“A cup of coffee, please!”

— A Data-Oriented Guide to Coffee Establishment in Manhattan

Steven Du, Capstone Project Report, Nov 2019

Part I: Introduction

Napoleon Bonaparte once commented: ‘I would rather suffer with coffee than be senseless.’ We, the modern generation, perhaps are in a good position to understand his feeling. The unique stress of modern life often forces us to rely on the power of coffee to stay awake, to work and ‘suffer’ at the same time. Nonetheless, who could resist the temptation of a cup of hot, newly brewed coffee along with pastries after waking up from dream?

Needless to say, coffee is an indispensable part of modern life, especially for people who reside in metropolises. Given coffee’s staple drink status, coffee establishments (places that primarily sell coffee or coffee-related drink) naturally play a key role in urban residents’ life. A visit to a familiar coffee establishment in the morning can be an essential part of people’s morning routine. Students or freelancers frequently choose coffee establishments as the places for working and studying. Coffee establishments are also perfect meeting place between friends or co-workers.

Choosing a Good Coffee Establishment in Manhattan

Choosing a good coffee establishment, while may not be as serious or time-consuming as choosing a fancy restaurant, could lead to considerable happiness (or misery if not lucky) to one’s life. A little extra knowledge on where and how to look for a good coffee establishment can thus be quite valuable for every residents and visitors in a large city.

Among the cities in America, New York’s love for coffee is quite remarkable. According to a report published by WalletHub[1], the New York city boasts more coffee shops per capita than any other cities in the United State, and is ranked as the second best coffee city in America, only after Seattle.

---

[1] McCann, Adam, “Best Coffee Cities in America, Sep 24, 2019. Available at: https://wallethub.com/edu/best-cities-for-coffee-lovers/23739/

This study aims to compose a data-orient guide to the coffee establishments in Manhattan for people who current resides in New York and people who plans to visit this city in near future. Rather than trying to pick up the 'best' coffee establishment as one might do for restaurants, this guide aims to provide a detailed summary on the quality as well as popularity of coffee establishments in Manhattan Island and offer coffee lovers a few tips on how to choose from the numerous coffee establishments in their everyday life.

Major Tasks of this Study

To achieve our goal, this study will utilize the API of Foursquare to retrieve a complete (or nearly complete) list of coffee establishments along with their locations in Manhattan area. By making use of the premium endpoint, 'details', this study will also summarize various features of the coffee establishments, including the overall rating, the number of rating received, the price tier, number of likes, number of tips and photos. By analyzing these data, this study will aim to answer the following questions:

- Are all the coffee establishments similar to each other in term of quality? Does the price matter in determining the quality of coffee establishments? Is there any major difference between the large coffee chains such as Starbucks and the local coffee shops?
- What is the relationship between quality and popularity? Does a popular coffee establishment also serve better coffee?
- Can we locate certain area(s) in Manhattan where the quality and popularity of coffee establishments are significantly better or worse than other areas in Manhattan?

Part II: Data

Data Collection

To obtain the data needed for further analysis, I firstly utilized the 'search' endpoint in the API of Foursquare to retrieve a list of coffee establishments in Manhattan. To obtain a complete list of coffee establishments, I conducted multiple searches requests in which each search was centered around one neighborhood in Manhattan[2]. Afterward,

[2] The coordinates of all neighborhoods in New York area are obtained from the geographical data of New York studied in the Week-3 module of this class.

the ‘detail’ endpoint was used to collect key statistics, including metrics on the popularity and quality, of each coffee establishments. In the end, I assigned each venue to a Neighborhood Tabulation Area (NTA, which includes one or multiple neighborhoods) based on the geographical data[3] of the boundaries of NTAs in New York. The detailed procedure of the data collection is summarized in Figure 1.

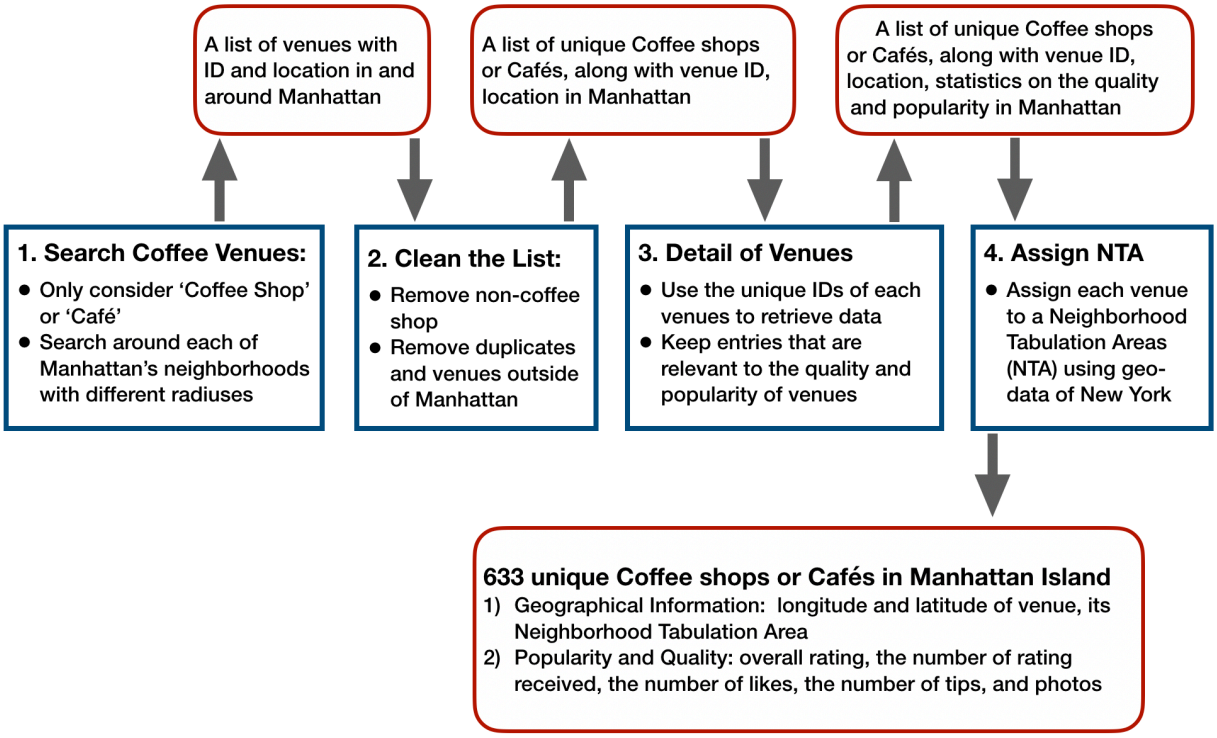

Figure 1: Procedure of Data Collection

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 phrases 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[4] maintained and updated by Foursquare.

---

[3] Available at https://geo.nyu.edu/catalog/nyu-2451-34561

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

Judging from this list, it is plausible that Foursquare may 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 apply our common sense, we might argue that a ‘Coffee Shop’ often focuses on coffee-related drinks along with a small selection of snacks, while a ‘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 quality of the venues.

Admittedly, ‘Coffee Shop’ or ‘Café’ are not the only places one can purchase coffee. Bakery, bagel shop, small restaurant or even convenience store can all be reasonable choice. 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 quality or popularity 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 may actually introduce unnecessary complexity into this study. 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 would significantly increase the number of venues I need to examine using 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.

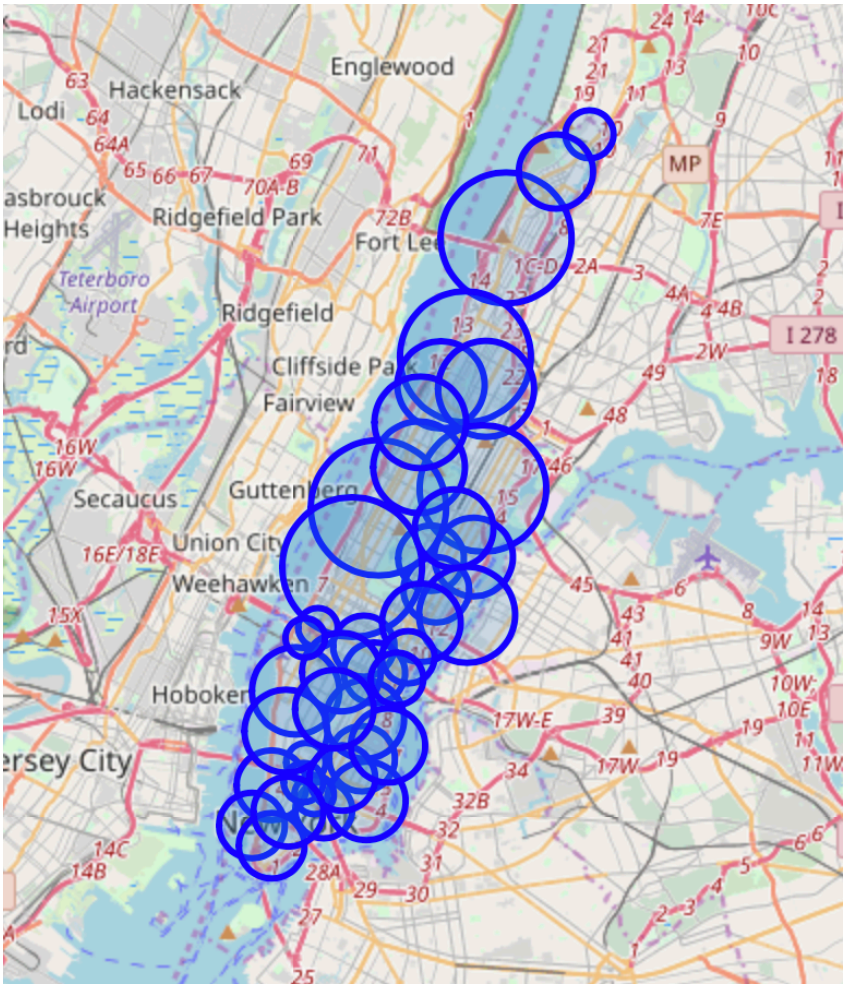

Figure 2: Centers and Radius of Search Call in Manhattan

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.

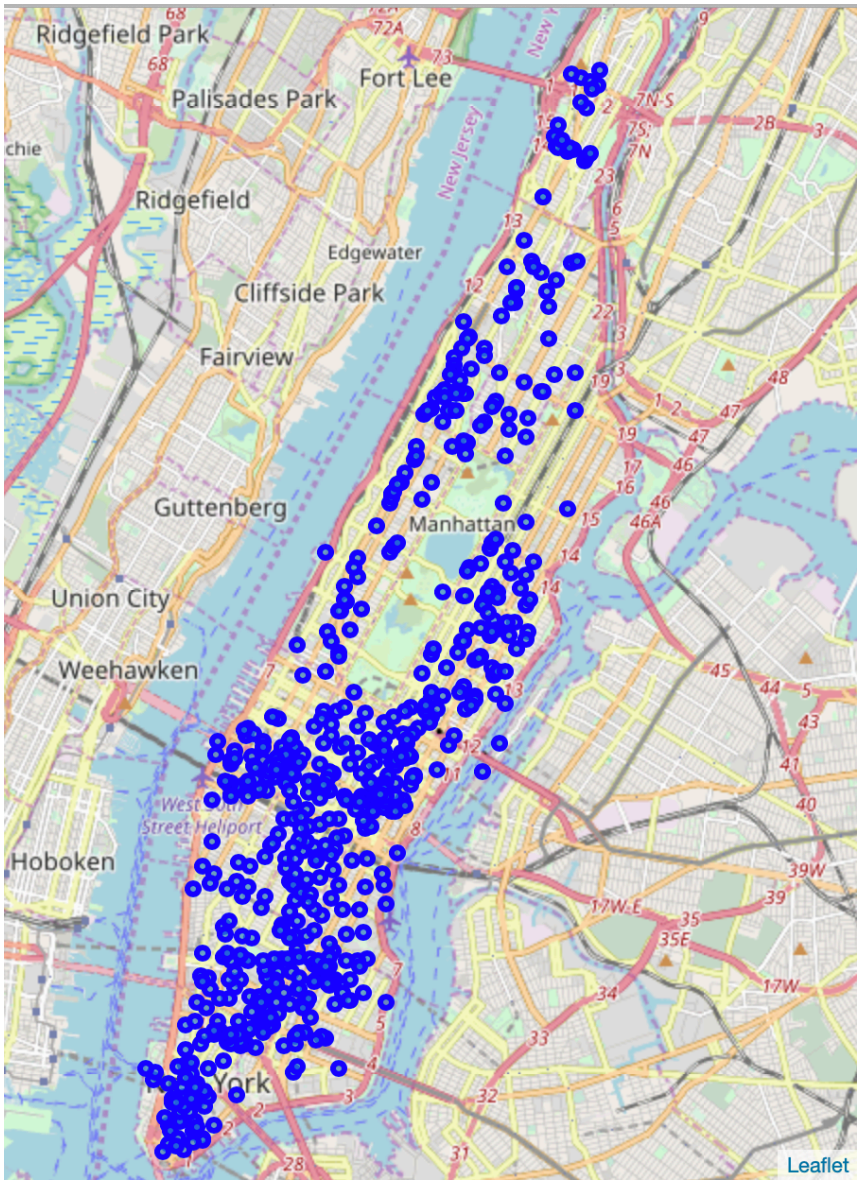

Figure 3: Coffee Shops and Cafés in Manhattan

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 quality and popularity of these coffee establishments, the following features were recorded along side with the geographical information:

Feature Name	Meaning of Feature
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).

4) Neighborhood Tabulation Areas

The neighborhood data we used to retrieve venues does not contain the geographical boundaries of neighborhoods. Moreover, some neighborhoods, especially the ones located in the lower Manhattan area are extremely close to each others. Consequently, for any given venue, it is hard to determine which neighborhood it locates. Fortunately, New York city also divides the city into Neighborhood Tabulation Areas (NTAs) and each NTA contains one or several Neighborhoods. In Manhattan Island, there are 40 neighborhoods and 29 NTAs. I download the geo-data of the Neighborhood Tabulation Areas of Manhattan at: https://geo.nyu.edu/catalog/nyu-2451-34561 and use the *within* method from *Shapely* package to assign each venue to the NTA it belongs to.

5) Create Additional Features

Create Logarithm of Counts:
For each venues, four different types of counts (*Ratings, Likes, Tips and Photos*) are collected. However, these counts data tend to contain some extremely values. For instance, while many venues may receive around 100-200 likes, a few venues can boast over 1000 likes. Thus, the natural logarithm transformation is applied to obtain the logarithm of the counts. To handle the situation where the count might be 0, 0.5 is added to the count before logarithm transformation is performed.

Price Tier of the Venue:
While there are four different tiers (1-4, from the least to most expensive) for the restaurant-type venues in Foursquare, the majority of the coffee establishments are of price tier 1 or 2. In fact, only 4 out of 633 coffee establishments has price tier 3. For this reason, I create an indicator variable *Price* to represent the price of the coffee establishments: 0 for low price (price tier 1) and 1 for high price (price tier 2 or 3).

Starbucks:
I wish to explore the different between large chain coffee shops such as Starbucks and individual coffee establishments. For this purpose, an indicator variable Starbucks is created where 0 is used to represent Non-Starbucks establishment and 1 is used to represent Starbucks.

Café:
An indicator variable *Café* is also created, in which 0 is used to represent that this establishment is categorized as 'Coffee Shop' and 1 is used to represent that this establishment is categorized as 'Café'.

Note that all the Starbucks are categorized as 'Coffee Shop'. Thus, this list of coffee establishments essentially includes three categories: 'Coffee Shop that is not Starbucks', 'Starbucks' and 'Café'.

A Brief Description of Data

In the following table I present a brief summary of the dataset I will use in the following analysis. Additional figures can also be find in the exploratory analysis part of the Method section.

Feature Name	Description
Venue ID	Unique ID of the venue. 633 unique coffee establishment are collected in the list.
Venue Name	Nam of the Venue
Venue Latitude	Latitude of the Venue
Venue Longitude	Longitude of the Venue
Neighborhood Tabulation Areas	Name of the neighborhood tabulation area (29 in total for Manhattan Island).

Feature Name	Description
Café	Whether the venue is categorized as Café (1) or Coffee Shop (0). 196 out of 633 coffee establishments are categorized as Café.
Starbucks	Whether the venue is a Starbuck Shop (1) or not (0). 157 out of 633 coffee establishments are categorized as Starbucks.
Price	Whether the price is high (1) or low (0). 108 out of 633 coffee establishments are of the high price.
Rating	Rating of the establishment. 82 out of 633 coffee establishment did not receive any rating at all, which will be dropped in the following up analysis as missing data. For the rest, the rating ranges between 5 and 9.3. The 25% quantile, median and 75% quantile of the rating are 6.0, 7.7 and 8.45 respectively.
Likes Count **Ratings Count** **Tip Count** **Photos Count**	Counts of the likes, ratings, tips and photos received by the venue. For a common coffee establishment, the values of these counts can range between dozens to two or three hundreds. However, some coffee establishments may receive thousands of likes or ratings
Likes (Log) **Ratings (Log)** **Tips (Log)** **Photos (Log)**	Logarithm transformation of the above counts (add 0.5 to avoid logarithm of 0).

Part III: Methodology and Analysis

Exploratory Analysis

Firstly we will take a look of our data sett with the help of a few graphs to gain some insight on the relationship between features

1) Popularity: Four Counts (and Logarithm), Which One Matters?

All the four counts (counts of likes, ratings, tips and photos) can be regarded as indicators of the popularity of the coffee establishment. But are they related?

Figure 4 displays the pairwise plots of the four counts (upper panel) along with the pairwise plots of the logarithm of the four counts. It can be seen that all these four counts are strongly positively related to each other. That is to say, if we want to predict the rating (quality) based on the counts (popularity), to avoid multicollinearity, we only need to consider one of the counts into our model. For this reason, we will focus on the counts of the **number of likes** and use it as a measure of popularity.

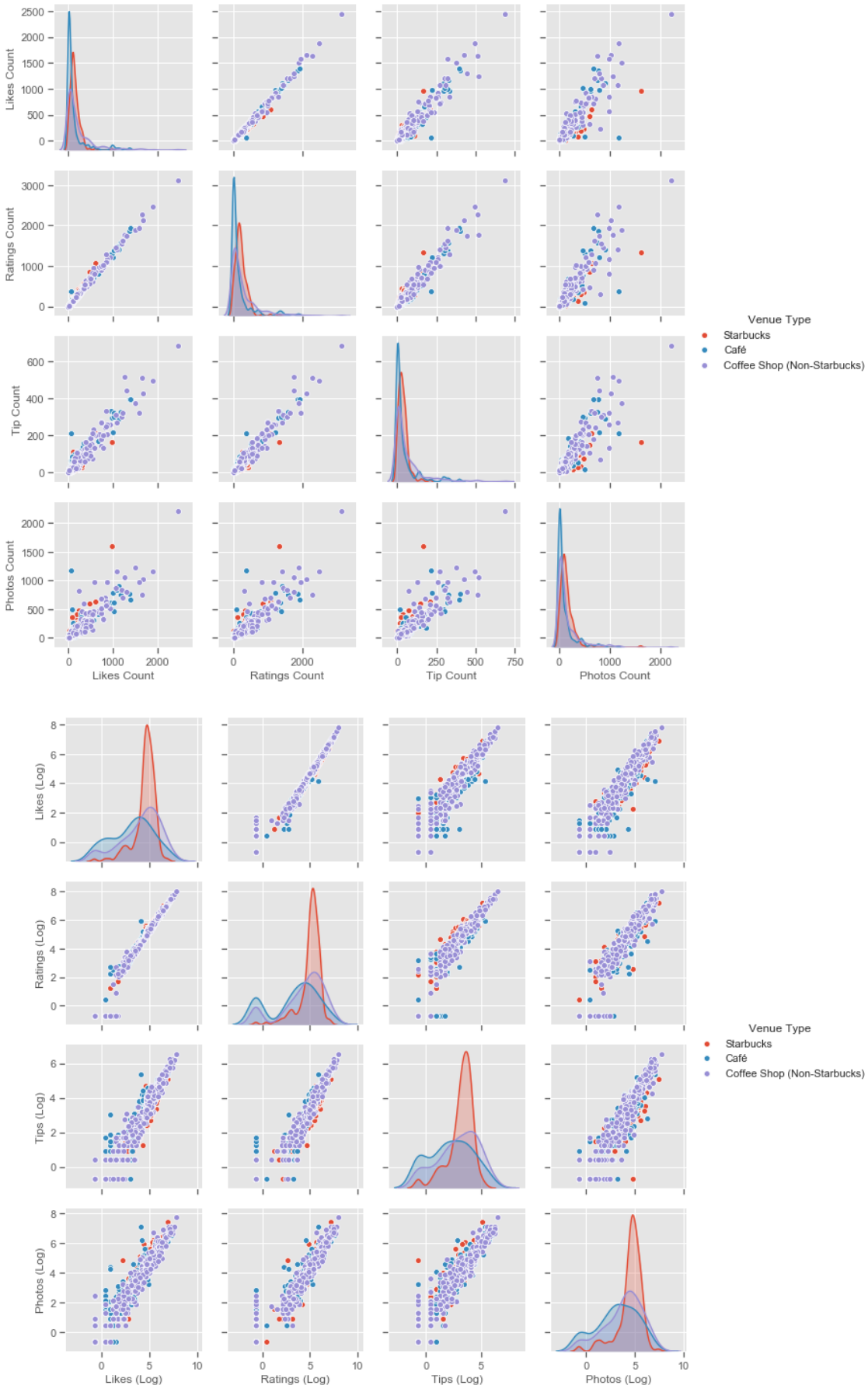

Figure 4: Counts - Measures of Popularity - for Different Types of Venues.
(Upper: Counts; Lower: Logarithm of Counts).

Moreover, if we compare the plot of counts and the plot of the logarithm of the count, it is clearly that, while the distributions of the counts are highly skewed with quite a few extremely large values, the distributions of the logarithm of the counts do not suffer the same issue. Also, based on the distributions of the logarithm of the counts, a considerable proportion of the coffee establishments receive 0 or only a few counts. My close inspection shown that these establishments generally also do not have a rating value. Since the primary goal of this study is to understand what determines the quality of coffee shop, we will drop the venues that do not receive any rating as missing data without rating.

Lastly, to investigate whether the types of venues (Starbucks, Café, non-Starbucks Coffee Shop) are related to the popularity of the venues, we also draw the box plots of the logarithm of Likes Count for each type of the venues in Figure 5.

As can be seen from this figure, while non-Starbucks Coffee Shops tend to receive higher number of likes on average compared to the Cafés, the spreads are similar. This suggests that the distinction between non-Starbucks Coffee Shops and Cafés might be quite limited in term of popularity. However, the distribution of the logarithm of the likes of Starbucks is much concentrated at a relatively hight value.

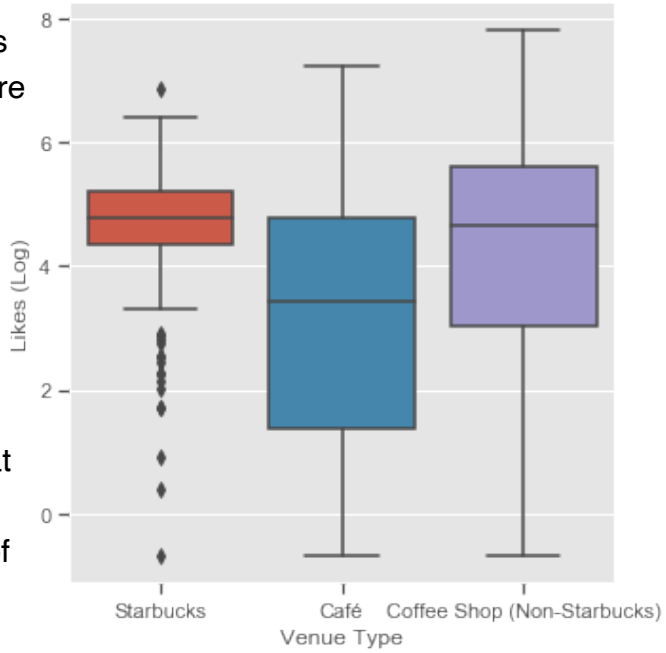

Figure 5: Logarithm of Likes for Different Types of Venues

2) Quality vs Popularity

To investigate the relationship between quality (measured by rating) and popularity (measured by the number of likes), the scatter plot between the logarithm of the number of likes and the rating is displayed in Figure 6. Different colors are used to distinguish different types of venues.

Based on this Figure, it can be seen that both features are generally positively related. That is, a coffee establishment that is more popular (more likes) also tend to have better quality (higher rating). And as to the difference between the types of venues, it appears

that, while there is no noticeable difference between Non-Starbucks Coffee Ships and Cafés, for the same level of rating, Starbucks tend to receive more likes compared to the other coffee establishments.

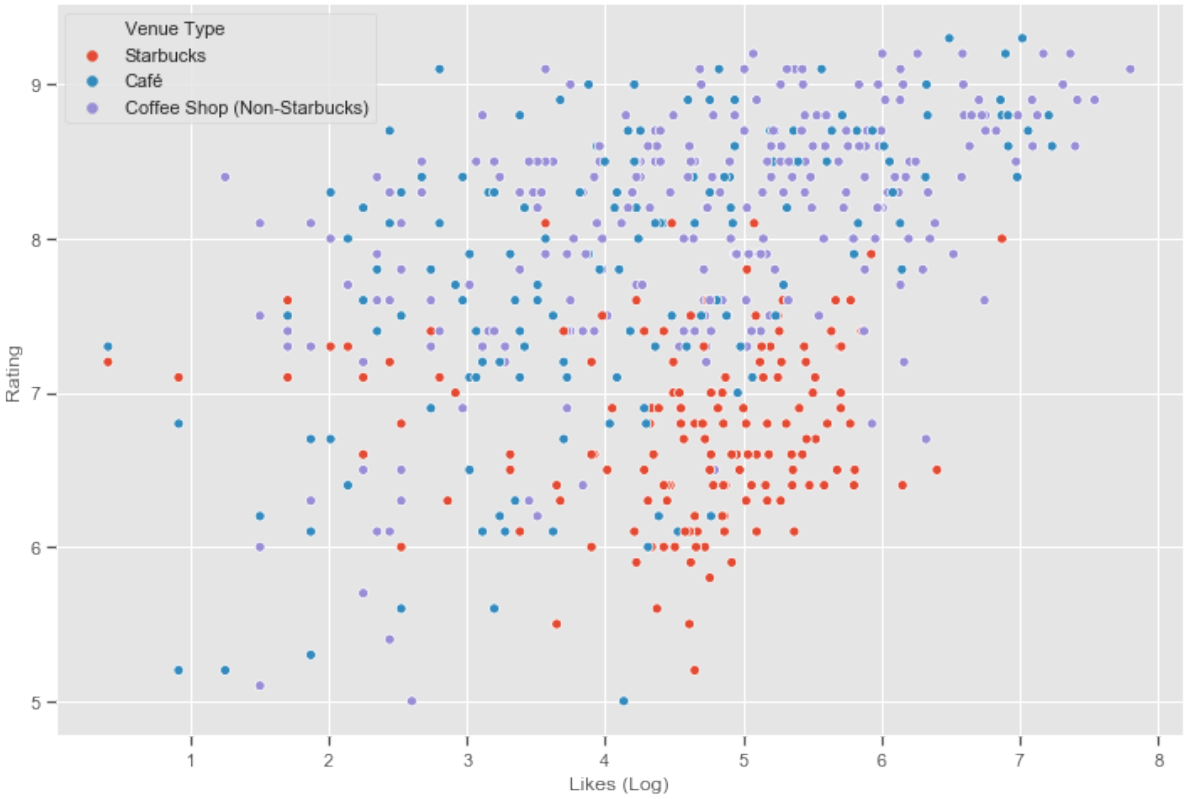

Figure 6: Popularity vs Quality

From the box plots (Figure 7) of the ratings for three types of venues, we can also see clearly that, while the distributions of rating between non-Starbucks Coffee Shops and Cafés are similar (with non-Starbucks Coffee Shops tend to receive higher rating on average), the average rating of Starbucks is considerably lower.

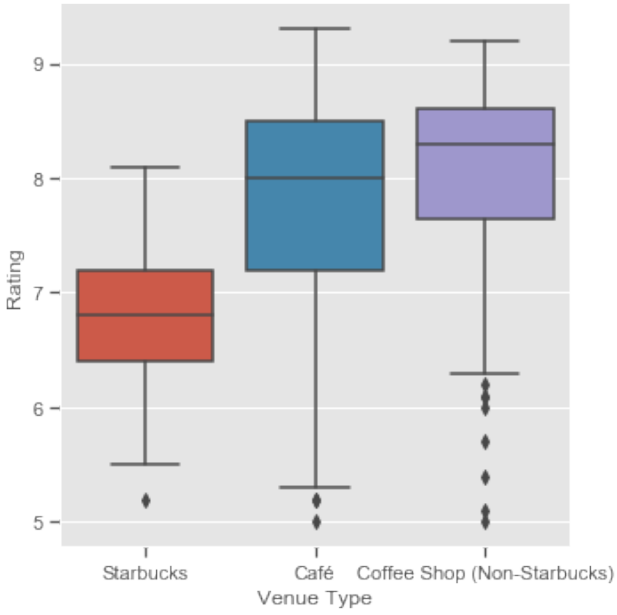

Figure 7: Rating for Different Types of Venues

3) Effect of Price

Coffee establishments are generally cheap. In this analysis, we categorize coffee establishments with price tier 1 (83%) as low-priced, and coffee establishments with price tier 2 (16%) and 3 (4 out of 633) as high-priced. The box plots of rating and the logarithm of likes count for both low and high-priced coffee establishments are summarized in Figure 8.

Based on these figures, it can be conclude that on average coffee establishments with higher price tend to offer better quality coffee and also attract larger popularity. Nonetheless, the variation among low-priced coffee establishments is also larger. Consequently, there are plenty of good choices among low-priced 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!

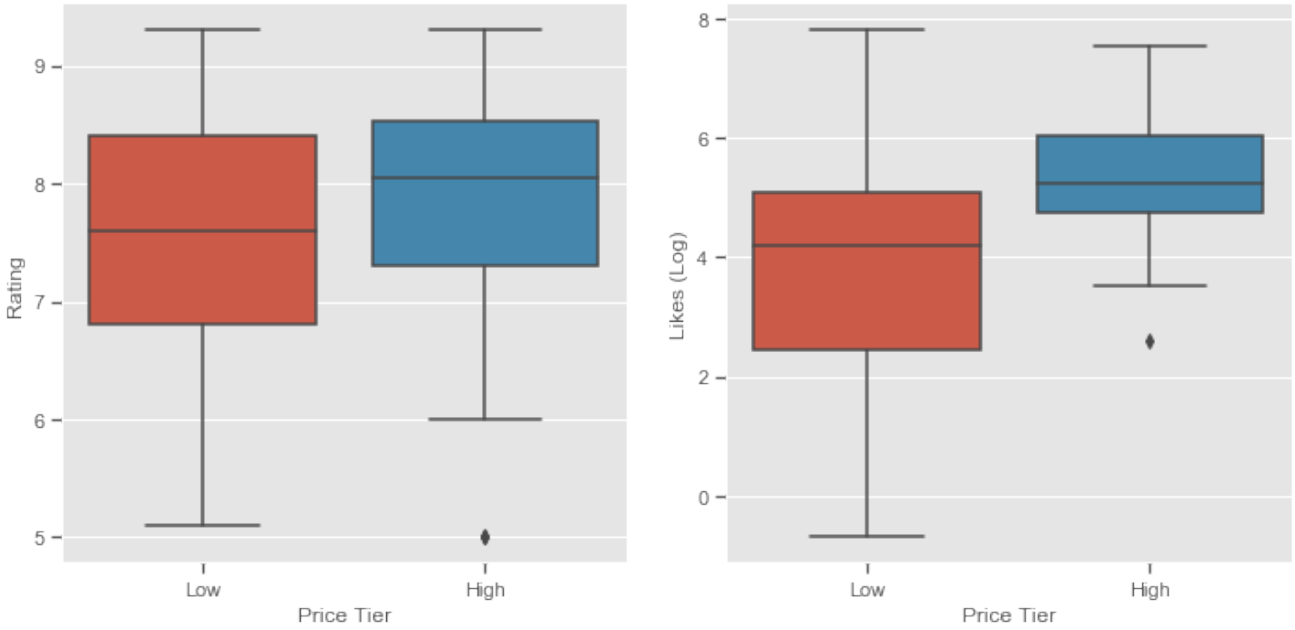

Figure 8: Effect of Price

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 Neighborhood Tabulation Areas.

The maps in Figure 9 suggested that, there are considerable differences in term of the quality and popularity of coffee establishments in different NTAs. Generally speaking, coffee establishments located in the southern part of Manhattan tend to have better quality and also attract more popularity. However, different neighborhoods do have different focus. For instance, the NTA with the highest average number of likes is the West Village area, but the NTA with the highest average rating is the Lower East Side. Moreover, although the northern part of Manhattan tends to host coffee establishments with lower ratings and numbers of likes, Hamilton Heights is a notable exception. Here you can find coffee establishments with very high rating comparable to the Lower East Side but with relatively fewer likes. In summary, depends on whether you want to visit a high-quality or a popular coffee shops, you may need to visit different areas of Manhattan.

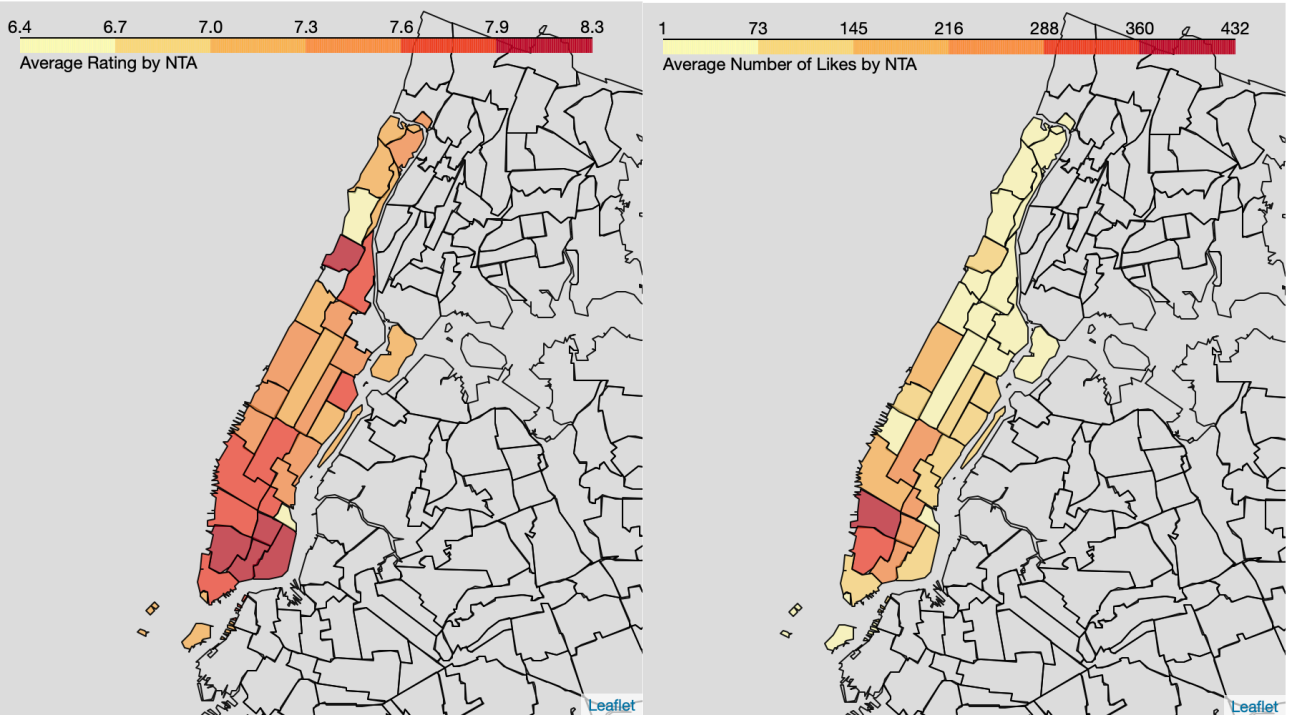

Figure 9: Average Ratings (Left) and Average Number of Likes (Right) of Coffee Establishments in Different Neighborhood Tabulation Areas (NTAs) of Manhattan.

In summary, from our exploratory analysis, the following statements appear to be valid:

- The quality and popularity of coffee establishments are generally positively related.
- Compare to the other coffee establishments, Starbucks tend to attract a reasonable level of popularity but receive lower rating.

- High-priced coffee establishments tend to receive higher rating and more likes but low-priced coffee establishments provide more variability.
- Coffee establishments in different neighborhood tend to have different characteristics.

Below we will conduct more detailed analysis to verify our finding.

Multiple Regression Model for the Rating of Coffee Establishments

As demonstrated in Figure 6, the rating of coffee establishments seems to be positively related to the logarithm of the number of likes. In addition, it seems that such relationship can differ considerably between different types of coffee establishments (especially for Starbucks). Thus, if we wish to explore the relationship between rating and logarithm of the number of likes using regression model, it is necessary to take into consideration of the interactions between the types of coffee and the logarithm of the number of likes. Moreover, we also wish to explore the role of price for determine the rating of coffee establishments. Based on these consideration, we will first try to fit the following multiple regression Model:

$$\begin{aligned}\text{Rating} &= \beta_0 + \beta_1 \text{Log-Likes} \\ &\quad + \beta_2 \text{Price} + \beta_3 \text{Café} + \beta_4 \text{Starbucks} \\ &\quad + \beta_5 \text{Log-Likes} \times \text{Price} + \beta_6 \text{Log-Likes} \times \text{Café} + \beta_7 \text{Log-Likes} \times \text{Starbucks} \\ &\quad + \epsilon\end{aligned}$$

Since we choose Rating as the dependent variable, we need to remove all the venues without ratings, and only focus on the venues that receiving ratings. In order to determine the significance of the fitted parameters, we will use the package *statmodels*, to fit the linear regression model and we have the following result:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.7834	0.147	46.012	0.000	6.494	7.073
Log_Likes	0.2860	0.031	9.333	0.000	0.226	0.346
Price	−1.3433	0.404	−3.322	0.001	−2.138	−0.549
Café	−0.4733	0.222	−2.128	0.034	−0.910	−0.036
Starbucks	−0.0283	0.287	−0.099	0.921	−0.592	0.535
Log_Likes:Price	0.2058	0.076	2.700	0.007	0.056	0.356
Log_Likes:Café	0.0837	0.049	1.699	0.090	−0.013	0.181
Log_Likes:Starbucks	−0.2763	0.060	−4.576	0.000	−0.395	−0.158

In the model displayed above, the coefficients for the indicator variable Starbucks and the interaction terms between the logarithm of likes and the indicator variable Café have

relatively high p-value, which mean that we can check simpler model without these two predictors. For this problem, we decide to make the comparison between models based on the values of BIC, and this model has BIC equals 1156.

After a few trials, we settle at the following model with BIC equals 1143. The R-squared of this model equals 0.52 and the p-values of all the coefficients are smaller than 0.01.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.6018	0.105	62.917	0.000	6.396	6.808
Log_Likes	0.3164	0.024	13.331	0.000	0.270	0.363
Price	-1.4357	0.401	-3.577	0.000	-2.224	-0.647
Log_Likes:Price	0.2235	0.075	2.982	0.003	0.076	0.371
Log_Likes:Starbucks	-0.2747	0.014	-20.008	0.000	-0.302	-0.248

According to this model, we have the following regression equation equation:

$$\begin{aligned}\text{Rating} &= 6.6018 + 0.3164\ \text{Log-Likes} - 1.4357\ \text{Price} \\ &\quad + 0.2235\ \text{Log-Likes} \times \text{Price} - 0.2747\ \text{Log-Likes} \times \text{Starbucks} \\ &\quad + \epsilon\end{aligned}$$

The following statements can be drawn based on this equation:

1. With all else being equal, whether the coffee establishments are categorized as ‘Coffee Shop’ or ‘Café’ has no impact on the rating of the coffee establishments.
2. The rating is positively correlated with the logarithm of the likes for all types of coffee establishments. However, the exact effect varies low-priced and high-priced coffee establishments, and between Starbucks and Non-Starbucks.
3. If the number of likes increases, the impact on the average rating is stronger for the high-priced coffee establishment compared to the low-priced coffee establishment.
4. If the number of likes increases, the impact on the average rating is much weaker for Starbucks compared to the other coffee establishments.
5. A high-priced coffee establishment only receive higher average rating than an otherwise-equivalent low-priced coffee establishment when the number of likes is greater than about 616 (a quite high number) . That is, for most coffee establishments, charing higher price can actually decrease the rating, assuming the number of likes remain the same.
6. A Starbucks tends to receive lower rating compared to an otherwise-equivalent non-Starbucks coffee establishments.

In the following table, we summarize the effect on the average rating when the number of likes is doubled based on the estimated regression model.

	Effect on Average Rating When the Number of Likes is Doubled
Non-Starbuck, Low-priced	Increase by 0.2193
Non-Starbuck, High-priced	Increase by 0.3742
Starbuck, Low-priced	Increase by 0.0289
Starbuck, High-priced	Increase by 0.1838

Clustering Analysis for Coffee Establishments in Manhattan

In this section, we will explore how the coffee establishments differ between NTAs in Manhattan. Rather than focusing on individual coffee establishments, the main aim of this section is to explore the general characteristics of coffee establishments in various NTAs, such as which NTAs tend to host coffee shops with high ratings.

For this purpose, the K-Means method will be applied twice to reduce the impact of coffee establishments with extreme high ratings or number of likes. First of all, K-Means algorithm will be used to cluster the coffee establishments around Manhattan into several major archetypes. We can then compute the proportions of archetypes in each NTA in Manhattan. The vectors of proportions are then used as input to cluster NTAs

1) Clustering Coffee Establishments

Our discussion in the previous section indicates that the distinction between ‘Coffee Shop’ and ‘Café is insignificant as long as other factors are taken into consideration. Thus, we will only use the rating, the logarithm number of likes, the price and whether the coffee establishment is Starbucks as the input features for the K-Means algorithm. All the features are standardized before the

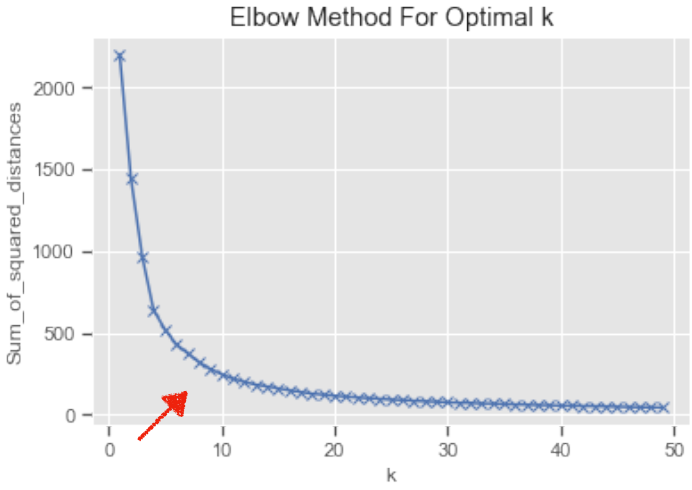

Figure 10: Elbow Plot for Clustering Coffee Establishments

K-mean algorithm is applied. And the elbow method is used to select the most suitable number of clusters (see Figure 10).

Based on the elbow plot, I choose to cluster the coffee establishments into 6 major archetypes. This clustering scheme not only leads to a reasonably low sum of squared distance within clusters, but also enables relatively simple interpretation of each archetypes. In the table below we summarize the characteristics of all the six archetypes based on the average values of features within each archetype:

	Rating	Likes (Log)	Price	Starbucks
Low Rating, Low Likes, Low Price	6.411538	2.516939	0.019231	0.038462
Starbucks, Low Rating, Mild Likes, High Price	6.531818	4.812814	1.000000	0.909091
Starbucks, Low Rating, Mild Likes, Low Price	6.769173	4.621938	0.000000	1.000000
High Rating, Low Likes, Low Price	8.002797	3.542682	0.000000	0.000000
High Rating, High Likes, High Price	8.197647	5.571075	1.000000	0.000000
High Rating, High Likes, Low Price	8.507759	5.702888	0.000000	0.000000

Two notable facts can be observed from this clustering scheme. First, while quality (high rating) is generally associated with popularity (high number of likes), a number of coffee establishments in Manhattan still manage to provide high quality coffee despite of the lacking of popularity. Such archetype of coffee shops also belongs to the low price tier and can thus be great choice for people who values quality more than popularity. Second, the low rating coffee establishments are mainly composed by Starbucks and other cheap coffee establishments that lack both quality and popularity.

2) Clustering NTAs

After we categorize coffee establishments in to 6 archetypes, we can compute the proportions of archetypes in each NTA and use these proportion vectors as the inputs for clustering NTAs. We again rely on Elbow method to determine the optimal number of clusters (Figure 11).

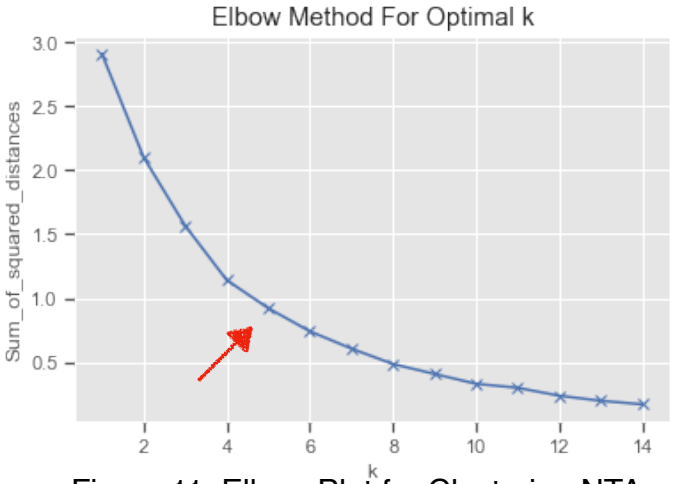

Figure 11: Elbow Plot for Clustering NTAs

The number of clusters is chosen as 5, which leads to three major clusters of NTAs and two single-NTA clusters. The key traits of each clusters, along with the average proportions of archetypes are summarized in the table below. We also show the clustering of NTAs over a map of Manhattan Island (Figure 12).

	High Rating, High Likes, High Price	High Rating, High Likes, Low Price	High Rating, Low Likes, Low Price	Low Rating, Low Likes, Low Price	Starbucks, Low Rating, Mild Likes, High Price	Starbucks, Low Rating, Mild Likes, Low Price
Mostly High Rating	0.164518	0.333294	0.298979	0.069196	0.023695	0.110318
High Rating and Some Starbucks	0.150398	0.116396	0.289231	0.066658	0.051074	0.326243
Mixture Ratings with Low Likes	0.062004	0.094907	0.378638	0.291832	0.023810	0.148810
Mainly Low Rating with some High Rating	0.111111	0.111111	0.000000	0.666667	0.000000	0.111111
Low Price Starbucks	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Based on these information, we may summarize the main traits of these clusters as follow:

- The first major cluster ‘Most High Rating’ includes most of regions in the southern end of Manhattan (with famous neighborhoods such as east and west villages, SoHo and Chinatown), along with Hamilton Height in the north. These neighborhoods mainly host high rating coffee establishment with a relatively small potion of low rating ones and can be the top choice for people who love high quality coffee.
- The second major cluster ‘High Raring and Some Starbuck’ is mainly composed by areas in middle Manhattan such as Midtown and Lincoln Square, along with a few neighborhoods in southern Manhattan and the Marble Hill and Inwood areas on the northern end. These neighborhoods do host a significant number of high rating coffee establishments. However, compared to areas in the first cluster, shop in these areas tend to be less popular and the proportion of Starbucks is also significantly higher. This area can be the idea place for enjoying coffee at a more leisure paces.
- The third major cluster, ‘Mixture Rating with Low Likes’ mainly contains regions located to the north of central park. Coffee shops in these region are generally cheaper and less popular. Nonetheless, a considerable proportion of coffee shops do offer high quality coffee and the chance for a visitor to find enjoyable coffee experience is still quite decent.
- The fourth and fifth clusters only contains one NTA. The fourth cluster is the Washington Heights South neighborhood located in the northern part of Manhattan

and mainly hosts coffee shops with low rating (along with a small proportion of high rating shops). The fifth cluster is the NTA composed by the Stuyvesant Town and Cooper Village neighborhoods. And the only coffee shops you can find in this NTA is low-priced Starbucks.

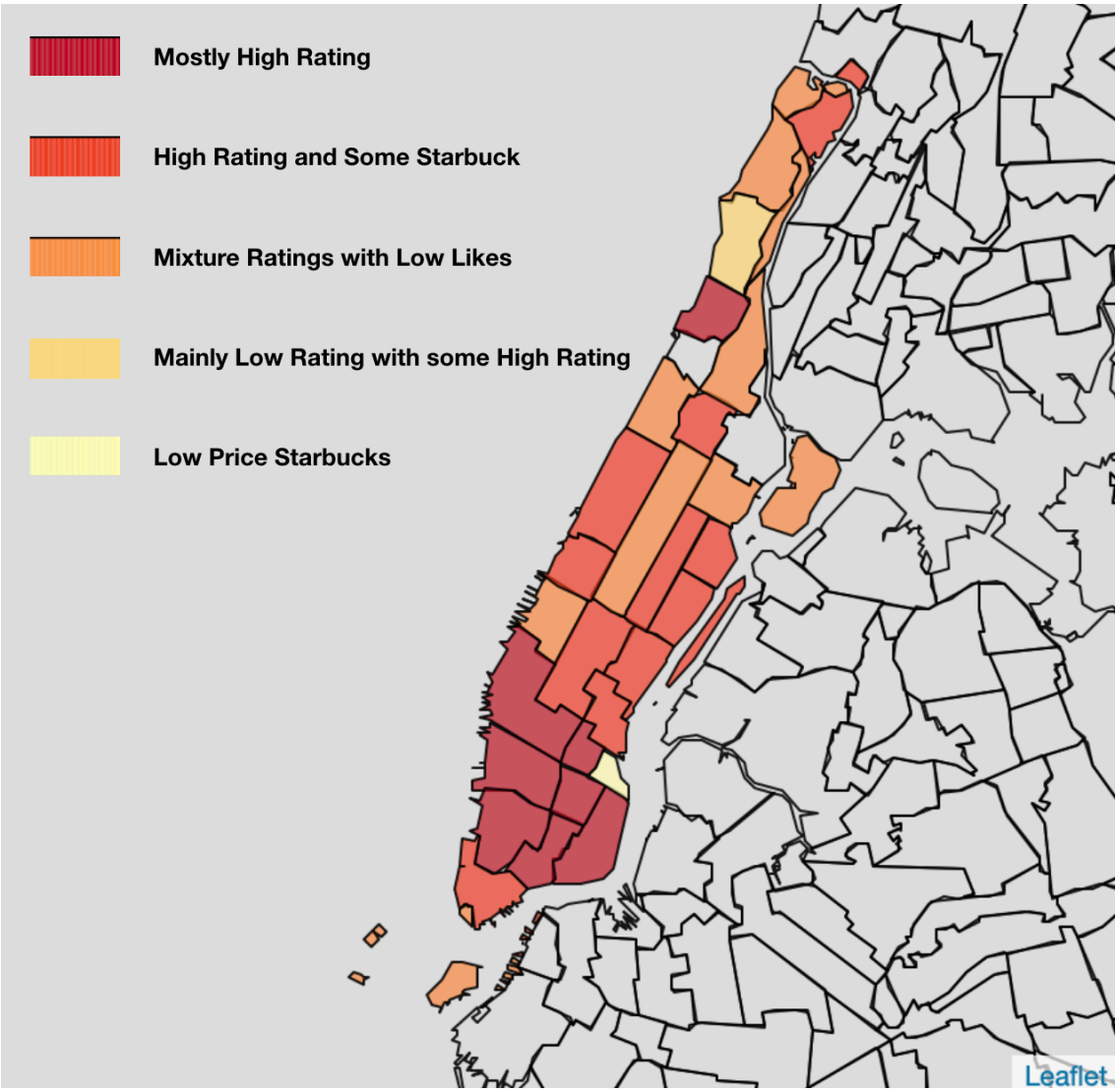

Figure 11: An Area-Based Map for Coffee Establishments in Manhattan

Table in the following page also contains a detailed list of the clustering of NTAs.

Neighborhood Type	Neighborhood Tabulation Area
Mostly High Rating	West Village
	SoHo-TriBeCa-Civic Center-Little Italy
	Lower East Side
	Hudson Yards-Chelsea-Flatiron-Union Square
	Hamilton Heights
	Gramercy
	East Village
	Chinatown
High Rating and Some Starbucks	Yorkville
	Central Harlem South
	Battery Park City-Lower Manhattan
	Upper West Side
	Upper East Side-Carnegie Hill
	Turtle Bay-East Midtown
	Murray Hill-Kips Bay
	Midtown-Midtown South
	Marble Hill-Inwood
	Lincoln Square
	Lenox Hill-Roosevelt Island
Mixture Ratings with Low Likes	Clinton
	East Harlem South
	Morningside Heights
	Washington Heights North
	park-cemetery-etc-Manhattan
	Central Harlem North-Polo Grounds
Mainly Low Rating with some High Rating	Washington Heights South
Low Price Starbucks	Stuyvesant Town-Cooper Village

Part IV: Result

Based on our analysis above, we can draw the following conclusions on the coffee establishments in Manhattan areas:

1. The number of likes, number of ratings, number of tips and number of photos of coffee establishment are all strongly linear correlated to each others. Practically speaking, one of these counts would sufficient for describing the popularity of the coffee establishment.
2. Generally speaking, the rating of a coffee establishment is positively related to the number of likes it received. The more likes a coffee establishment receive, the higher the rating. The exact relationship, however, depends on the type and price of coffee establishments.
3. With the same number of likes, a low-priced coffee establishment tends to receive higher rating compared to a high-priced coffee establishment.
4. With the same number of likes, a Starbucks tends to receive lower rating compared to a non-Starbucks coffee establishment.
5. Whether a coffee establishment is categorized as ‘Coffee Shop’ or ‘Café’ has no significant impact on the rating of coffee establishment if the number of likes and the price of coffee of coffee establishment are taken into account.
6. Coffee establishments in Manhattan can be roughly categorized into 6 archetypes. Three of these archetypes feature coffee establishments with high rating but varying degree of popularity and price. Two of these archetypes represent low-priced and high-priced Starbucks. And the last one archetype features low rating, low-priced coffee establishments with low number of likes.
7. Rather than being indifference to the archetypes, different neighborhoods in Manhattan tend to ‘prefer’ different archetypes.
8. The proportion of high rating archetypes is considerably higher in most neighborhoods of southern Manhattan, and Hamilton Height in the north.
9. Coffee establishments in Lower Manhattan, most regions in middle Manhattan, along with Marble Hill and Inwood on the northern tip of Manhattan are divided between high-rating non-Starbucks coffee shops and Starbucks.
10. Most neighborhoods in the northern part of Manhattan are characterized by mixed-rating coffee establishments with low popularity.

Part V: Discussion

One of the noteworthy observation of this study is the unique characteristics of Starbucks. If we compare the rating and number of likes of Starbucks and other coffee establishment, it is clearly that Starbucks attract a large number of likes (as well as tips and photos) despite of their relatively low rating. There are two possible explanation to this phenomenon. First, Starbucks are underrated by users; Second, Starbucks employed special strategy to receive more than average numbers of likes, tips or photos. While second explanation seems to be more likely, additional data is needed to draw the final conclusion.

Moreover, the inhomogeneous distribution of coffee establishments over the different neighborhood in Manhattan also suggests that the performance of Starbucks might be different in different market. In the regions belong to the second NTA cluster 'High Rating with Some Starbucks', the proportion of Starbucks is quite higher (over 30%). Considering the fact that the major archetype of the non-Starbucks coffee shops in these regions is the archetype with high rating but low number of likes, it might be possible that Starbucks hold certain advantage over shops in such archetype. In contrast, the proportion of Starbucks is usually within the range of 10% to 15% in the regions with significant numbers of either high rating and high number of likes archetype, or the low rating and low number of likes archetype. The interpretation of such effect is again beyond the scope of this study.

As our goal is to focus on the general properties of coffee establishment, logarithm transformation is applied to the counts to reduce the impact of coffee shops with extremely high number of likes. Still, as such coffee shops often draw significant attention from public, it can be very interesting to study why certain coffee shop could achieve such level of popularity. One potential staring point of such inquiry is to focus on the coffee establishments with rating greater than 8 and then explores the properties of such coffee shops.

While our study does reveal very interesting relationship between the rating and the number of likes for coffee establishments in Manhattan, as well as the inhomogeneous distribution of coffee establishments over the different neighborhood in Manhattan, this study does have its limitation. And we will discuss these limitation in the rest of this section.

First of all, we use the rating of coffee establishment as a measure of quality and use the number of likes as a measure of popularity. To what extend these two measurements could truly represent the quality and popularity of corresponding coffee establishment is debatable. For instance, our common experience tells us that, we are more likely to review a restaurant we may only visit once every few months and are less likely to review a coffee shop we visit everyday. In this sense, the rating and the number of likes can be biased measures of quality and popularity. There is also the possibility that, certain shop owner may hire people to post fake reviews, which could further reduce the reliability of these two measures.

Secondly, our study is limited to a few features of the coffee shops. A more extensive study may also explore the tips and the photos posted by the users. For instance, one might want to check how many tips are about the quality of coffee, how many tips are about the quality of deserts and how may tips are about the environment of the shops. Such detailed information could allow us to have a much better understanding on why people might like or hate the corresponding coffee establishment. Unfortunately, with a personal Foursquare account, the access to these features is very limited and such inquiry is beyond the capability of this project.

Thirdly, our study on the distribution of coffee shops in Manhattan relies on a reasonable geographical division of neighborhoods. While NTAs are used to divide Manhattan into a number of regions in this study, whether such division is optimal for our purpose is unclear. It is interesting to see if it is possible to form relatively regular geographical division of Manhattan by clustering the coffee shop directly. Doing so would require a clustering approaches that can take both the numerical features and the spatial information of subjects into consideration.

Last but not the least, as our study is an observational study by nature, the relationships between variables or between the coffee shops and neighborhoods we discovered from our study can not be regarded as causal relations. In particular, even though our regression analysis indicates that the rating is positively related with the number of likes, we can not simply conclude that a shop owner could certainly increase shop's rating if he or she manages to double the number of likes received by the shop.

Part VI: Conclusion

To summarize, this study applies the multiple linear regression and K-means clustering algorithm to study the coffee shops in the Manhattan area. Our analysis indicate that the rating and the number of likes are generally positively related, though the degree of relationship depends on the type and the price of coffee establishment under consideration. We also reveal considerable difference in term of the quality and popularity of coffee establishment between neighborhoods in Manhattan area.

By providing a clear, data-supported explanation on the relationship between rating and the number of likes for coffee establishments in Manhattan area, as well as the distinct characteristics of coffee establishments in different neighborhood, this study can serve as a guide for choosing suitable coffee establishment in the neighborhoods of Manhattan area.

Generally speaking, for both local residents or visitors to Manhattan area, our study would recommend non-Starbucks coffee shop over Starbucks, and low-priced coffee shop over high-priced one. While we acknowledge the fact that the quality is usually correlated with the popularity, we do want to point out that, Manhattan, especially the neighborhoods located in the northern par of Manhattan, do host a large number of coffee establishments that offer high quality service but at the same time attracts less popularity. In addition, while there is a reasonable chance to find high quality coffee shops in most neighborhoods of Manhattan, high quality coffee shops are more common in the southern part of Manhattan and in the Hamilton Height neighborhoods.

Thanks for reviewing this report and I would greatly appreciate any comments you might provide.