Manhattan Eating Preferences

Adrian Tarniceriu

20.08.2019

1. Introduction

The most populous city in the United States, New York City is home to more than 8 million inhabitants (without including the metropolitan area). On top of this, it attracted 65.2 million tourists in 2018, which poses great challenges in terms of infrastructure, supplies, and alimentation. Being the most densely populated borough of New York, these issues are even more important in Manhattan.

This project focuses on the alimentation aspect of the Manhattan region. More specifically, it studies the eating place distribution based on neighborhoods, where the eating places are grouped in three categories: cafes, restaurants, and fast-food places.

This analysis could help visitors find a location more suitable for their preferences, but also provide entrepreneurs with an insight about possible business opportunities, such as opening a new restaurant (or shutting down an existing one).

2. Data

2.1 Data Source

The data used herein covers the Manhattan borough of New York City. This is sub-divided in 40 distinct neighborhoods. The location data, containing the list of neighborhoods and their coordinates (latitude, longitude), is obtained from [*https://cocl.us/new_york_dataset*](https://cocl.us/new_york_dataset).

	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

Fig 1. Example of location data

For each neighborhood, the venue data is obtained from Foursquare. In total, there were 1182 venues retrieved for the 40 neighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue ID	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	4bf58dd8d48988d1ca941735	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	4bf58dd8d48988d102941735	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	4bf58dd8d48988d147941735	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	4bf58dd8d48988d1e0931735	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	4bf58dd8d48988d148941735	Donut Shop
5	Marble Hill	40.876551	-73.91066	Blink Fitness Riverdale	40.877147	-73.905837	4bf58dd8d48988d176941735	Gym
6	Marble Hill	40.876551	-73.91066	TCR The Club of Riverdale	40.878628	-73.914568	4e39a891bd410d7aed40cbc2	Tennis Stadium
7	Marble Hill	40.876551	-73.91066	Land & Sea Restaurant	40.877885	-73.905873	4bf58dd8d48988d1ce941735	Seafood Restaurant
8	Marble Hill	40.876551	-73.91066	T.J. Maxx	40.877232	-73.905042	4bf58dd8d48988d1f6941735	Department Store
9	Marble Hill	40.876551	-73.91066	Starbucks	40.873755	-73.908613	4bf58dd8d48988d1e0931735	Coffee Shop

Fig 2. Example of venue data

2.2 Data Preprocessing

Because the venues can be of different types (eating places, sport, museums, etc.), and we are only interested in the eating places, we will filter out unrelated data. After this filtering, we will have 455 venues left.

Before continuing with the analysis, we do one more pre-processing step: here, we are only interested in cafes, restaurants, and fast-food places. Thus, each food-related Venue category will be assigned to one of these three categories as shown in the table below

Venue Category contains:	Assigned to:
'restaurant', 'bodega', or 'diner'	Restaurant
'cafe' or 'coffee'	Cafe
'joint', 'bagel', 'pizza', 'breakfast', 'burger', 'burrito', 'creperie', 'fast food', 'pastry', 'sandwich', 'snack', or 'taco'	FastFood

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue ID	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	4bf58dd8d48988d1ca941735	FastFood
1	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	4bf58dd8d48988d147941735	Restaurant
2	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	4bf58dd8d48988d1e0931735	Cafe
3	Marble Hill	40.876551	-73.91066	Land & Sea Restaurant	40.877885	-73.905873	4bf58dd8d48988d1ce941735	Restaurant
4	Marble Hill	40.876551	-73.91066	Starbucks	40.873755	-73.908613	4bf58dd8d48988d1e0931735	Cafe
5	Marble Hill	40.876551	-73.91066	Subway Sandwiches	40.874667	-73.909586	4bf58dd8d48988d1c5941735	FastFood
6	Marble Hill	40.876551	-73.91066	Boston Market	40.877430	-73.905412	4bf58dd8d48988d14e941735	Restaurant
7	Marble Hill	40.876551	-73.91066	SUBWAY	40.878493	-73.905385	4bf58dd8d48988d1c5941735	FastFood
8	Marble Hill	40.876551	-73.91066	Subway	40.877720	-73.905380	4bf58dd8d48988d1c5941735	FastFood
9	Marble Hill	40.876551	-73.91066	Terrace View Delicatessen	40.876476	-73.912746	4bf58dd8d48988d146941735	Restaurant

Fig 3. Example of clean pre-processed data

At this point, all needed data is grouped in one data-frame. Further processing will be described in the following sections (part 2 or week 5 of the final assignment).

3. Methodology

3.1 Exploratory Data Analysis

The goal of this project is to classify the neighborhoods based on the eating venue type. We will firstly explore the data by showing the number of Cafe, FastFood, and Restaurant places for each neighborhood. Afterwards, we will apply K-Means clustering to compare the exploratory results to a machine learning technique output.

In Fig. 4 we show the count of each eating place type per neighborhood. This data is further processed for Fig. 5 to show the most common venue type. Note that here we only display the results. More comments will follow in the "Discussion" section.

	Neighborhood	Cafe	FastFood	Restaurant	Venues
0	Battery Park City	2.0	4.0	1.0	7.0
1	Carnegie Hill	2.0	4.0	7.0	13.0
2	Central Harlem	0.0	3.0	11.0	14.0
3	Chelsea	1.0	1.0	11.0	13.0
4	Chinatown	0.0	3.0	10.0	13.0
5	Civic Center	1.0	1.0	11.0	13.0
6	Clinton	0.0	2.0	3.0	5.0
7	East Harlem	1.0	2.0	11.0	14.0
8	East Village	2.0	4.0	9.0	15.0
9	Financial District	2.0	2.0	5.0	9.0

Fig 4. Cafe/FastFood/Restaurant count per neighborhood

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Venues
0	Battery Park City	FastFood	Cafe	Restaurant	7.0
1	Carnegie Hill	Restaurant	FastFood	Cafe	13.0
2	Central Harlem	Restaurant	FastFood	Cafe	14.0
3	Chelsea	Restaurant	FastFood	Cafe	13.0
4	Chinatown	Restaurant	FastFood	Cafe	13.0
5	Civic Center	Restaurant	FastFood	Cafe	13.0
6	Clinton	Restaurant	FastFood	Cafe	5.0
7	East Harlem	Restaurant	FastFood	Cafe	14.0
8	East Village	Restaurant	FastFood	Cafe	15.0
9	Financial District	Restaurant	FastFood	Cafe	9.0
10	Flatiron	Restaurant	FastFood	Cafe	7.0
11	Gramercy	Restaurant	FastFood	Cafe	12.0
12	Greenwich Village	Restaurant	FastFood	Cafe	14.0
13	Hamilton Heights	Restaurant	Cafe	FastFood	14.0
14	Hudson Yards	Restaurant	Cafe	FastFood	9.0
15	Inwood	Restaurant	FastFood	Cafe	13.0
16	Lenox Hill	Restaurant	FastFood	Cafe	10.0
17	Little Italy	Restaurant	FastFood	Cafe	10.0
18	Lower East Side	Restaurant	Cafe	FastFood	13.0
19	Manhattan Valley	Restaurant	FastFood	Cafe	16.0

Fig 5. Most common venue per neighborhood

39 of the 40 neighborhoods have “Restaurant” as the most common venue, and one neighborhood has “FastFood” as the most common venue. The second most common venue is “FastFood” in 28 cases and “Cafe” in 12 cases.

3.2 Machine Learning Techniques

After the exploratory data analysis, we apply the K-Means algorithm to partition the neighborhoods based on the eating venue type. It is a significant strategy as a business can target these specific groups of customers and effectively allocate marketing resources. As there are three different venue types, we will use three different clusters.

The clustering results are shown in the following section.

4. Results

Fig. 6 shows the cluster assignment (0, 1, or 2) for the first 20 neighborhoods. The corresponding map-view is shown in Fig. 7. One may notice that the same-color points in Fig. 7 are not grouped together. This was expected, because the grouping has been done based on the venue types and not on the geographical location. 5 neighborhoods have been assigned to cluster 0, 19 neighborhoods to cluster 1, and 15 neighborhoods to cluster 2. One neighborhood does not have any returned venue and was removes from the analysis.

	index	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Venues
0	28	Manhattan	Battery Park City	40.711932	-74.016869	0.0	FastFood	Cafe	Restaurant	7.0
1	30	Manhattan	Carnegie Hill	40.782683	-73.953256	2.0	Restaurant	FastFood	Cafe	13.0
2	6	Manhattan	Central Harlem	40.815976	-73.943211	2.0	Restaurant	FastFood	Cafe	14.0
3	17	Manhattan	Chelsea	40.744035	-74.003116	1.0	Restaurant	FastFood	Cafe	13.0
4	1	Manhattan	Chinatown	40.715618	-73.994279	2.0	Restaurant	FastFood	Cafe	13.0
5	32	Manhattan	Civic Center	40.715229	-74.005415	1.0	Restaurant	FastFood	Cafe	13.0
6	14	Manhattan	Clinton	40.759101	-73.996119	0.0	Restaurant	FastFood	Cafe	5.0
7	7	Manhattan	East Harlem	40.792249	-73.944182	2.0	Restaurant	FastFood	Cafe	14.0
8	19	Manhattan	East Village	40.727847	-73.982226	2.0	Restaurant	FastFood	Cafe	15.0
9	29	Manhattan	Financial District	40.707107	-74.010665	0.0	Restaurant	FastFood	Cafe	9.0
10	38	Manhattan	Flatiron	40.739673	-73.990947	1.0	Restaurant	FastFood	Cafe	7.0
11	27	Manhattan	Gramercy	40.737210	-73.981376	0.0	Restaurant	FastFood	Cafe	12.0
12	18	Manhattan	Greenwich Village	40.726933	-73.999914	2.0	Restaurant	FastFood	Cafe	14.0
13	4	Manhattan	Hamilton Heights	40.823604	-73.949688	1.0	Restaurant	Cafe	FastFood	14.0
14	39	Manhattan	Hudson Yards	40.756658	-74.000111	1.0	Restaurant	Cafe	FastFood	9.0
15	3	Manhattan	Inwood	40.867684	-73.921210	1.0	Restaurant	FastFood	Cafe	13.0
16	10	Manhattan	Lenox Hill	40.768113	-73.958860	0.0	Restaurant	FastFood	Cafe	10.0
17	22	Manhattan	Little Italy	40.719324	-73.997305	0.0	Restaurant	FastFood	Cafe	10.0
18	20	Manhattan	Lower East Side	40.717807	-73.980890	1.0	Restaurant	Cafe	FastFood	13.0
19	25	Manhattan	Manhattan Valley	40.797307	-73.964286	2.0	Restaurant	FastFood	Cafe	16.0

Fig 6. Cluster labels for the Manhattan neighborhoods

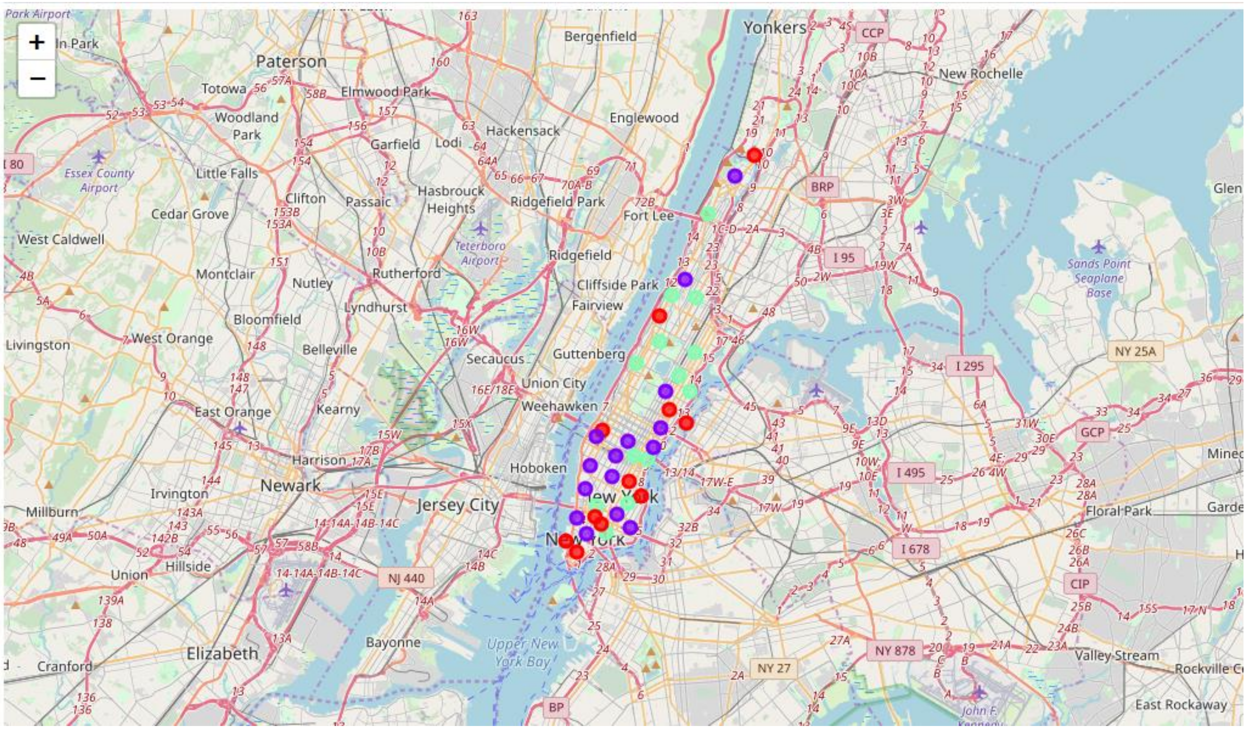

Fig 7. Neighborhood view based on the cluster label

5. Discussion

The goal of this project was to classify the neighborhoods based on the eating venue type. According to the exploratory data analysis, 39/40 neighborhoods have "Restaurant" as the most common venue type. Based on this, one might expect three classes ("Restaurant" as most common and "FastFood" as second common, "Restaurant" as most common and "Cafe" as second common, and other).

When inspecting Fig. 6, we see that "Restaurant" as most common and "FastFood" as second common neighborhoods have been assigned to all three clusters (e.g., Carnegie Hill, Chelsea, and Clinton, respectively). This is explained by the fact that the automatic K-Means classification also considered the total venue count, and the exact ratio between different venues, not only the popularity order, offering a better segmentation.

The difference in the naïve segmentation based on the most common venue type and the K-Means segmentation is a good example for the usability of machine learning techniques, which consider aspects that can be ignored by one's intuition. In this case, the examples are relatively basic, but the benefits will increase with more complex data.

6. Conclusion

This project presented the classification of Manhattan neighborhoods based on the most common eating venue (Cafe, FastFood, and Restaurant). It showed that the results using a naïve partitioning based on the most common venue type are not the same as when using machine learning techniques. The take-home message is that (especially when the data becomes more complex) simple intuitive methods are no longer enough. In these situations, machine learning is not only a recommendation, but a necessity.