MARCH 30, 2019

NEIGHBORHOOD SEGMENTATION AND VENUE OPENING SUGGESTIONS IN MANHATTAN

IBM DATA SCIENCE CAPSTONE

YELI WANG

Table of Contents

1. Introduction

Manhattan ranks the top of the five boroughs of the New York City, in both its population density, and its economic significance. The rich population gives a proper business sense for anyone who wish to open a restaurant in Manhattan. Meanwhile, an investor needs to consider several aspects prior to the decision of opening a restaurant, including picking the neighborhood, setting the price tier, and determining the cuisine type.

This report focuses on the analysis of local restaurants in Manhattan, and provides the insight for investors on opening restaurants by clarifying the following two questions: where and which type of restaurant should be opened.

2. Data

In order to support the analysis, the geographical information of Manhattan and the details of the venues were collected. The geographical data came from two data source, the JSON file from the IBM Data Science Course, and Wikipedia webpage where the Manhattan neighborhoods were listed.

The details of the venues were collected based on the neighborhoods from Foursquare API. At maximum 100 venues within 500 meters radius of that neighborhood under the food section were requested.

2.1 Data Collection

Foursquare API provides two types of requests. The regular API returns the general venue details. In this analysis, the regular API *explore* was used to collect the venues from a certain neighborhood with a user-specified radius and number of venues returned. One personal account can make 99,500 regular calls on a daily basis. While the rate limit is high, it only provides a certain level of information, which is insufficient for machine learning and segmentation in the later stage.

To obtain in-depth information on each venue, premium API calls have to be made. It gives the necessary information on all the categories of the venue, geographical location, detailed user ratings, price tier and opening date. One personal account can make 500 premium calls per day. In this study, after the venues were explored, the venue details were acquired through the premium call.

2.2 Data Cleaning

Foursquare API returned the results in JSON format. The output was ironed so that the necessary information could be extracted for analysis. The data exported from the regular API call is listed below.

- Venue ID
- Venue name
- Venue location
- Venue category

Note that the short version of the venue category was used, instead of the full category name. For example, “Pizzeria Sirenetta” is classified as French Restaurant, while the short name is French. By eliminating the word Restaurant, it is easy for data transformation and reference later.

The following data were collected from premium API call:

- Full address,
- Venue categories
- Tip counts
- Price tier
- Rating
- createdAt

It is not fully justified to classify one restaurant by one single category, because some Italian restaurants may also call themselves pizza shop. Consequently, the second category of the venue was also collected. If the venue has no second category, then the NaN was used instead.

The API response gives a dictionary object for price tier. Three keys are used to describe the price of the venue, namely a number label ranging from one to four for the price tier, a description of the label, and the currency. Generally one stands for cheap, and four for very expensive. The dictionary values were separated into three columns for easy access.

The createdAt object measures the seconds since epoch when the venue was created. Python built-in module datetime is used to decode the value into conventional date and time. Additionally, it was converted to the number of days since opening until now, so that all the time data was measured to the same reference.

2.3 Feature Selection

In this report, the Manhattan neighborhoods were arranged in different clusters. The prerequisite was the feature selection. Based on the discussion in data collection and data cleaning, the following features were selected.

- Venue categories
- Tip counts
- Price tier in numerical format
- Rating
- Days since opening.

The following table reports the correlation among each parameters. It could be concluded that none of these parameters are related to each other, except for rating vs. tips. The correlation coefficient reaches 0.49.

Figure 1 below shows the scatter plot of rating vs. tips. It can be inferred that the more tips one restaurant receives, the higher rating people tend to give, which is reasonable since people are more likely to provide feedbacks on venues that they feel satisfied. As a result, tip counts are removed from the feature set.

	RATING	TIPS	DAYS SINCE OPEN	PRICE TIER
RATING	1.000000	0.465219	0.000883	0.293448
TIPS	0.465219	1.000000	0.372426	0.254084
DAYS SINCE OPEN	0.000883	0.372426	1.000000	0.206203
PRICE TIER	0.293448	0.254084	0.206203	1.000000

Table 1: Correlation between features.

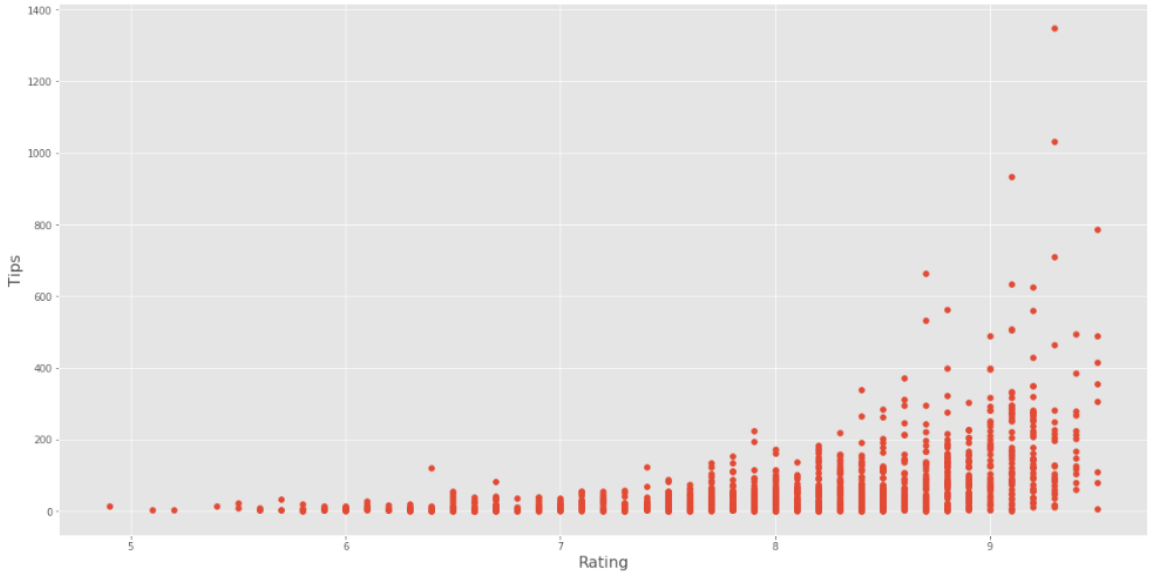

Figure 1: Rating vs. Tips

3. Methodology

Prior to clustering, the data was screened and the statistical results were calculated for better understanding on each parameter. Redundancy should be detected and removed for the sake of data quality. Moreover, the number of clusters applied for clustering analysis were tested so that the optimal one could be found.

3.1 Venue Exploration

In total 2,645 venues were collected through Foursquare API. The venue categories are first analyzed to see what the most popular cuisine is in Manhattan. The top 12 categories, which sum up to 1,385, accounts for 52 % of the total venues. The remaining venues categories are not considered in this analysis, as the numbers of restaurants are limited. By introducing excessive categories into the analysis, the result could be impacted. From Figure 2, the top three categories are Italian, Pizza and American.

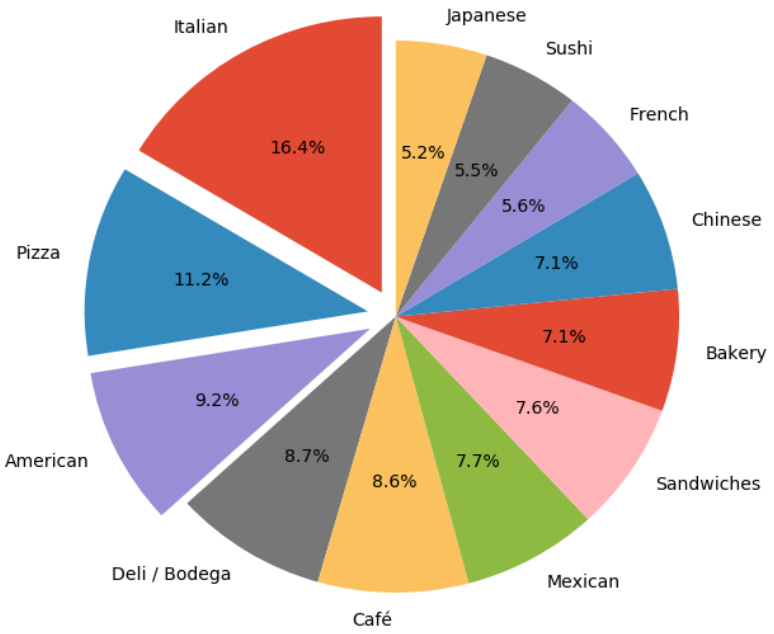

Figure 2: Pie Chart of Venue Categories.

After the data cleaning and feature selection, a new data matrix was built. Because the type of the restaurant categories was string, it was transformed to a dummy matrix, where each category becomes the column names. Only 1 and 0 wre assigned to each cell to represent what category one restaurant should fall under. Then the matrix was appended to the original data matrix.

After that, the whole data matrix was grouped by neighborhood, where the mean values of the remaining ungrouped columns were calculated. The purpose was to build a list of indicators of each neighborhood, so that the distance between each one could be calculated and used for clustering. It was common that some restaurant do not receive any tips or ratings from customers. They were dropped from the analysis to avoid dealing with NaN values.

By observation, Inwood and Washington Height are much alike in profile where Pizza, Mexican food are most common. Moreover, West Village and Flatiron show similarity in the overall ratings and price tier of the venues. Therefore, it was suspected that these neighborhoods would be classified under the same category.

In this study, two Machine Learning techniques are presented, namely k-Means clustering and hierarchical clustering. K-Means clustering partitions observations into several clusters. The observations within a cluster are very similar, while those across different clusters are much different. Hierarchical clustering builds a hierarchy of clusters. Agglomerative type is a bottom-up method, where clusters are merged together as one moves up the hierarchy. Divisive is a top-down approach and reversed version of agglomerative type.

3.2 Choice of Number of Clusters

One important parameter required for clustering analysis is the number of clusters k to form. Fewer clusters cannot differentiate each groups clearly, while more clusters might cause the centroids to converge to local minimal. Consequently, based on the k-Means method, the relationship between k versus the mean distance of the centers were studied. The elbow point were identified so that the optimal k could be picked.

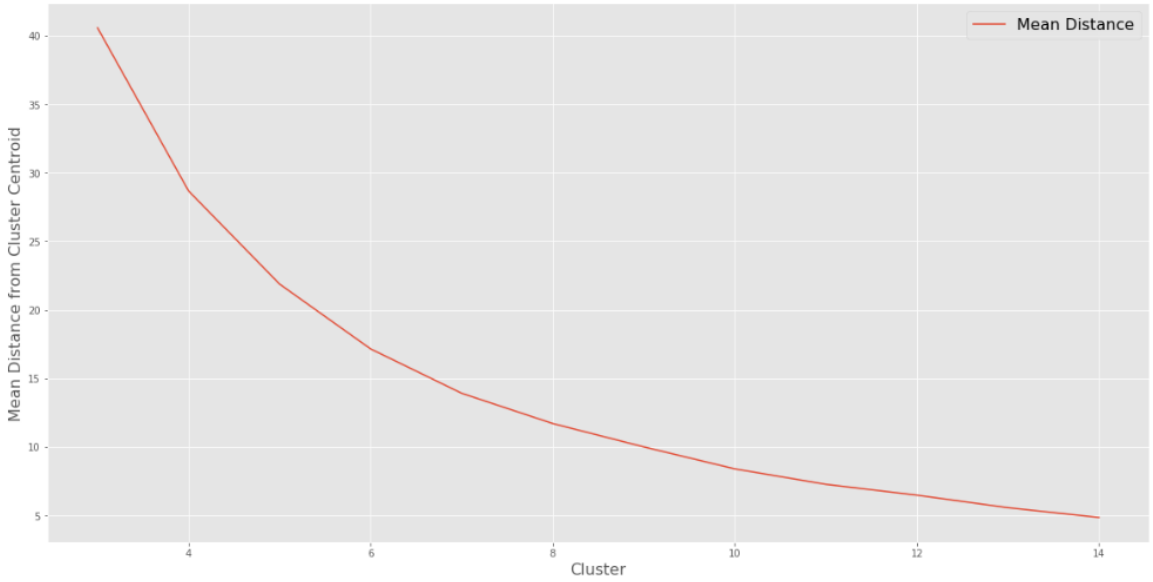

Figure 3: Cluster vs. Mean Distance from Cluster Centroid.

According to Figure 3, it can be seen that elbow point happened at k equals to 6. By increasing the clusters from 3 to 6, the mean distance drops sharply. The slope soothes after clusters are further increased. Therefore, this study sticks with six clusters for both clustering methods.

4. Results

In this section, the individual parameters were studied. The statistical results for each category were calculated, and the box plots were used for comparison. Clustering results were also presented for each clustering method.

4.1 Ratings

The statistical results on ratings are presented in Table 2. The overall rating is 8.0, with a standard deviation of 0.86. Among all the categories, Italian, French and Japanese received the highest reviews. By checking the box plots in Figure 4, it can be concluded that range of ratings of French or Italian is more concentrated than the rest of the categories, meaning that those restaurants generally offers high service to customers. Pizza or Sandwiches, on the other hand, show significant different from store to store.

	ITALIAN	PIZZA	AMERICAN	DELI/ BODEGA	CAFÉ	MEXICAN	SANDWICHES	BAKERY	CHINESE	FRENCH	SUSHI	JAPANESE
COUNT	223	137	126	80	110	101	103	92	87	78	75	71
MEAN	8.27	7.77	8.20	7.28	7.97	7.94	7.76	8.15	7.85	8.32	8.05	8.34
STD	0.67	0.95	0.78	0.90	0.86	0.75	0.96	0.76	1.04	0.64	0.80	0.76
MIN	6.1	4.9	5.5	5.5	5.1	5.6	5.7	5.9	5.2	5.8	5.9	6.1
25%	7.9	7.1	7.7	6.6	7.5	7.6	6.9	7.7	7.3	8.0	7.6	7.8
50%	8.3	7.9	8.3	7.3	8.1	8.0	7.9	8.3	7.9	8.4	8.2	8.5
75%	8.8	8.5	8.9	7.8	8.5	8.5	8.4	8.7	8.8	8.8	8.5	9.0
MAX	9.5	9.3	9.5	9.4	9.4	9.3	9.4	9.5	9.3	9.5	9.5	9.5

Table 2: Venue Rating Statistics.

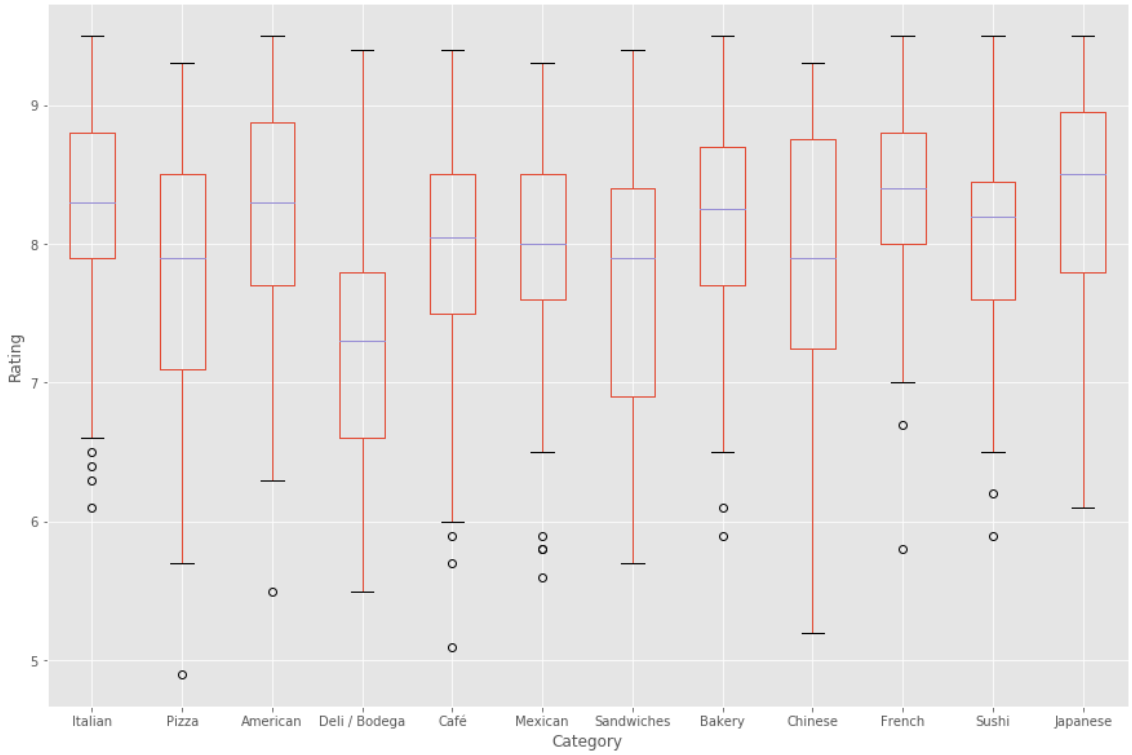

Figure 4: Venue Rating Box Plots.

4.2 Tier

The results on price tiers and associated box plots are shown below. Only two categories have a price tier close to three, which are Italian and French. This finding matches our impression that both cuisines represent a high-level of service and dishes. On the other hand, Pizza, Deli and Sandwiches are more affordable.

	ITALIAN	PIZZA	AMERICAN	DELI/ BODEGA	CAFÉ	MEXICAN	SANDWICHES	BAKERY	CHINESE	FRENCH	SUSHI	JAPANESE
COUNT	227	155	128	121	119	107	105	99	98	78	76	72
MEAN	2.62	1.20	2.39	1.17	1.34	1.74	1.35	1.34	1.43	2.72	2.21	2.44
STD	0.79	0.42	0.70	0.45	0.51	0.83	0.52	0.52	0.70	0.62	0.60	0.85
MIN	1	1	1	1	1	1	1	1	1	2	1	1
25%	2	1	2	1	1	1	1	1	1	2	2	2
50%	2	1	2	1	1	2	1	1	1	3	2	2
75%	3	1	3	1	2	2	2	2	2	3	2	3
MAX	4	3	4	3	3	4	3	3	4	4	4	4

Table 3: Venue Price Tier Statistics.

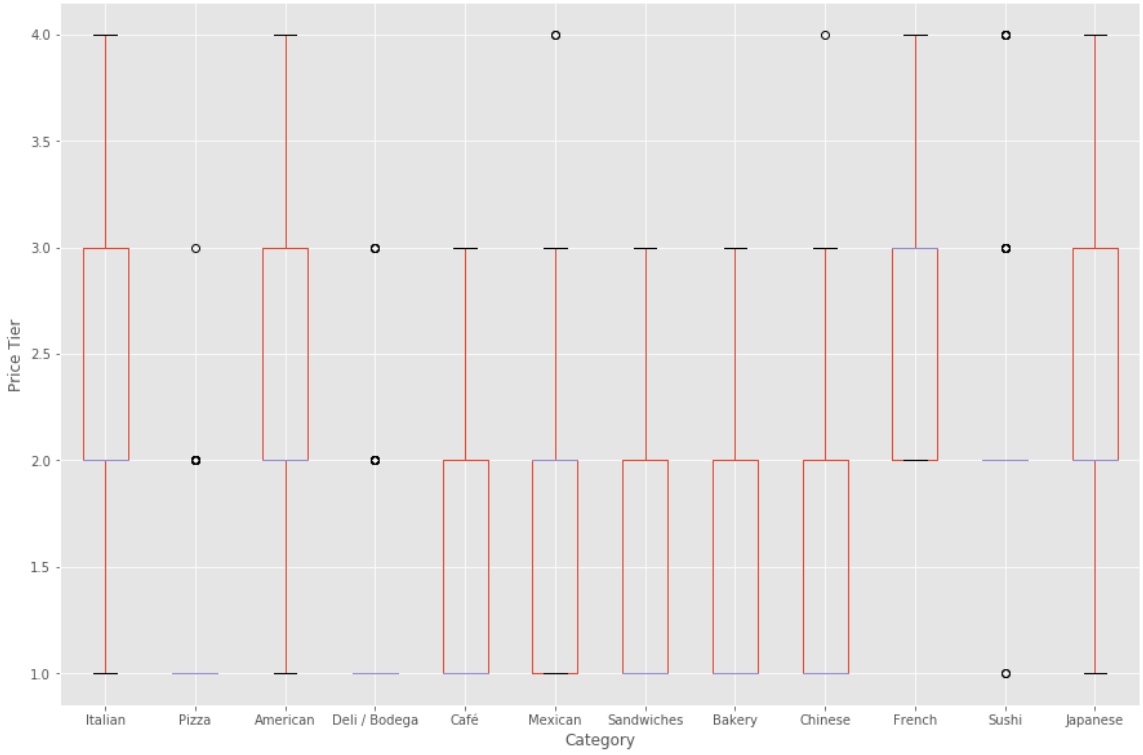

Figure 5: Venue Price Tier Box Plots.

4.3 Days

Under each category, there are restaurants that are either newly opened, or were established tens of years ago. It is hard to make reasonable justifications from it. Therefore, the data are accumulated across the time to see how many restaurants were opened in each year. According to Figure 7, it can be concluded that there is a sharp increase in the number from 2009 to 2011. Later the growth rate flattened and a stable rate was achieved. Café is one exception, where the overall growth rate was steady. The results between

2003 and 2009 should be filtered, probably because that Foursquare data did not trace back then.

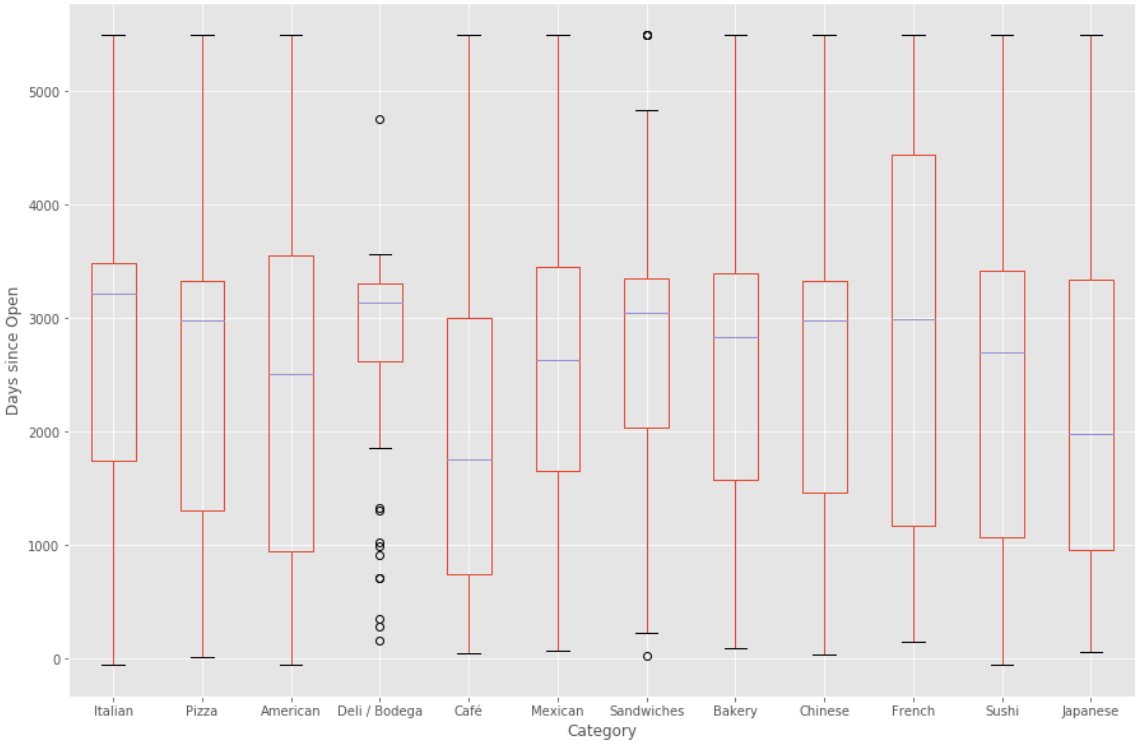

Figure 6: Venue Days Box Plots.

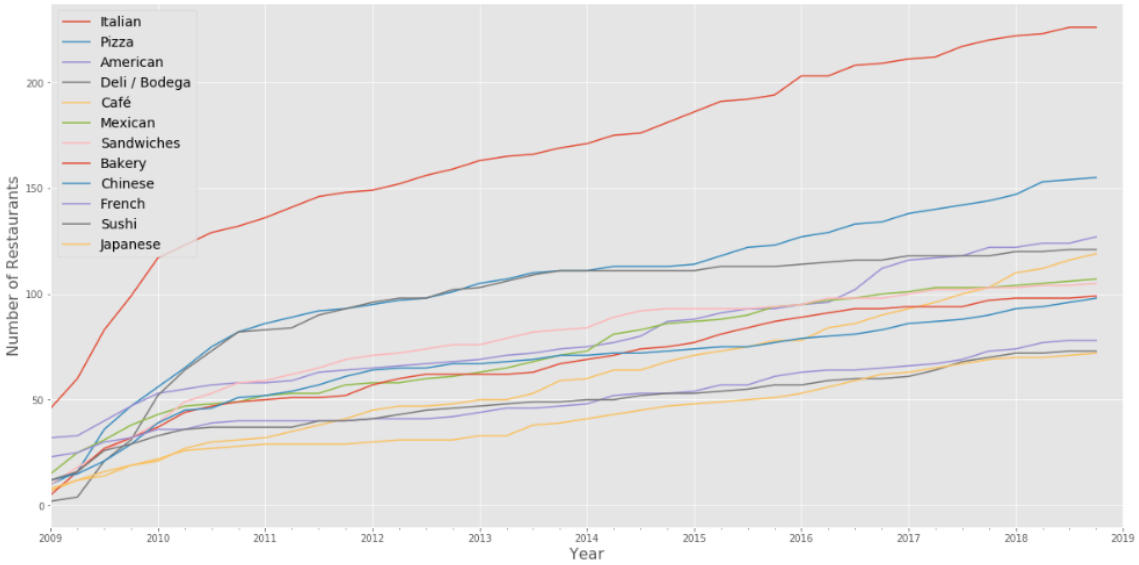

Figure 7: Venue Growth from 2009 to 2019.

4.4 Segmentation

4.1 K-Means Clustering

The results of the k-Means clustering is listed in Table 4. By grouping the neighborhoods under the same cluster, the mean value on venue rating, price tier, the number of tips received, and top 3 most common venues are summarized.

CLUSTER	RATING	PRICE TIER	TIPS	DAYS SINCE OPEN	1ST MOST COMMON VENUE	2ND MOST COMMON VENUE	3RD MOST COMMON VENUE
0	8.3	1.9	56.2	1820	American	Sandwiches	Italian
1	8.2	2.2	94.1	2867	Italian	American	Café
2	7.0	1.4	12.4	2719	Deli / Bodega	Sandwiches	Chinese
3	8.0	1.3	27.3	1788	Café	Italian	Chinese
4	7.6	1.3	36.9	2417	Pizza	Chinese	Bakery
5	8.0	1.9	53.2	2526	Italian	Pizza	Mexican

Table 4: Neighborhood Segmentation by k-Means Clustering.

It can be seen that cluster 2, 3, and 4 are very similar to each other. People could dine in those areas in a casual way or order deliveries, such as café, pizza and sandwiches places. Price tier is relatively low, and quality and service should be satisfactory, since the overall rating is between 7 and 8. For cluster 0, 1 and 5, the neighbors with high-end venues are grouped together, such as Italian, French or American. Both rating and price tier are high as expected. It can be inferred further that Cluster 4 has the most popular neighborhood, as the number of tips received is the highest than the others.

4.2 Agglomerative Hierarchical Clustering

CLUSTER	RATING	PRICE TIER	TIPS	DAYS SINCE OPEN	1ST MOST COMMON VENUE	2ND MOST COMMON VENUE	3RD MOST COMMON VENUE
0	7.2	1.3	16.6	2525	Deli / Bodega	Pizza	Sandwiches
1	6.6	2.0	7.0	3079	Sandwiches	Deli / Bodega	Sushi
2	8.1	2.0	68.1	2639	Italian	Pizza	American
3	8.5	1.4	114.8	2873	Chinese	Bakery	Sandwiches
4	8.5	2.0	69.6	1467	American	Japanese	Italian
5	8.0	1.3	27.3	1788	Café	Italian	Chinese

Table 5: Neighborhood Segmentation by Agglomerative Hierarchical Clustering.

Based on the results by agglomerative hierarchical clustering, we could observe that cluster 0, 3 and 5 are those with cheaper restaurants, and cluster 2 and 4 are ones that

are more expensive. Cluster 1 could be some outliers, where only a few neighbors are clustered together.

As both clustering method yield similar results, the discussion was based on those from k-Means.

5. Discussion

From the results presented below, it can be concluded that each neighborhood has its unique combinations of venues. Italian or French restaurants may be found more popular in certain areas than the others, from which it could be implied that opening a top-notch Italian restaurant at places where there are many café or pizza deliveries may not be a wise choice. Similarly, one will not want to set high price for sandwich shops, because the average price tier will be from cheap to moderate.

Figure 8 shows the map of Manhattan and each cluster is represented as a dot on it, with the yellow stands for cluster 0, and burgundy for cluster 5. The table below summarizes the places and the recommended venues for consideration. Readers should note that rental price, safety concerns, or other business factors are outside the scope of this study.

A detailed table of cluster results are included in Appendix.

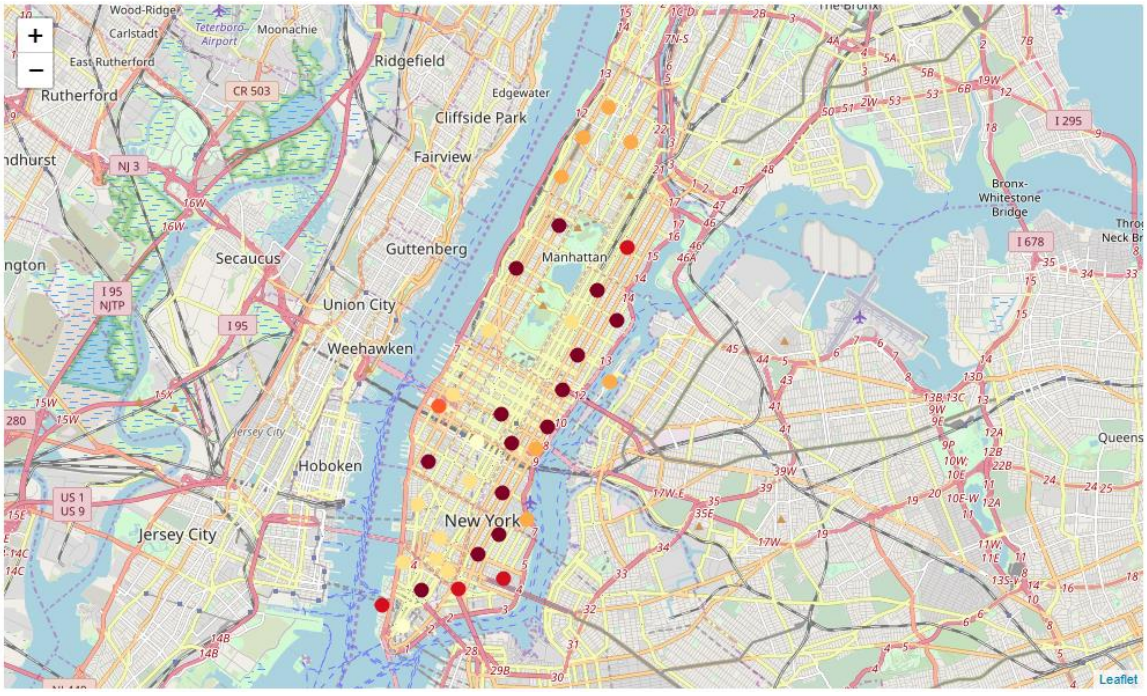

Figure 8: Neighborhood Segmentation Map.

CLUSTER	FAMOUS NEIGHBORHOOD	RATING	PRICE TIER	1ST MOST COMMON VENUE	2ND MOST COMMON VENUE	3RD MOST COMMON VENUE
0	Financial District, Midtown South	8.3	1.9	American	Sandwiches	Italian
1	Soho, Litter Italy, West Village, Flatiron	8.2	2.2	Italian	American	Café
2	Central Harlem, Morningside Heights	7.0	1.4	Deli / Bodega	Sandwiches	Chinese
3	Hudson Yards	8.0	1.3	Café	Italian	Chinese
4	East Harlem, Chinatown	7.6	1.3	Pizza	Chinese	Bakery
5	Upper West Side, Midtown, Chelsea	8.0	1.9	Italian	Pizza	Mexican

Table 6: Venue Recommendations.

5.1 Future Works

Foursquare provides access to the tips given to each venue. It is feasible to dump those reviews, and then research on them. Possibly, by mining texts, one could conclude what people actually cares about for each kind of restaurants, and then make recommendations to those owners for improvement.

6. Conclusion

In this study, the venues on the Manhattan were collected and grouped together based on their geographical locations to study the segmentation on neighborhoods. K-Means clustering and agglomerative hierarchical clustering with six clusters were applied, and similar clusters were formed. Recommendations were made on what type of venues were suitable to open in certain neighborhoods, including the category and price tier.

7. Appendix

NEIGHBORHOOD	RATING	PRICE TIER	TIPS	DAYS SINCE OPEN	AMERICAN	BAKERY	CAFÉ	CHINESE	DELI / BODEGA	FRENCH	ITALIAN	JAPANESE	MEXICAN	PIZZA	SANDWICHES	SUSHI	LABEL
BATTERY PARK CITY	7.7	1.4	24.8	1537	0.06	0.13	0.00	0.13	0.00	0.00	0.19	0.00	0.06	0.25	0.13	0.06	0
CARNEGIE HILL	7.7	1.8	30.0	2568	0.03	0.16	0.14	0.03	0.05	0.05	0.11	0.05	0.05	0.22	0.00	0.11	2
CENTRAL HARLEM	7.6	1.3	8.6	2576	0.10	0.05	0.05	0.20	0.20	0.10	0.00	0.00	0.00	0.20	0.10	0.00	0
CHELSEA	8.3	2.1	93.5	2407	0.08	0.17	0.10	0.02	0.02	0.08	0.17	0.08	0.08	0.08	0.06	0.04	2
CHINATOWN	8.5	1.4	114.8	2873	0.07	0.15	0.07	0.46	0.00	0.00	0.02	0.02	0.09	0.04	0.09	0.00	3
CIVIC CENTER	8.0	2.0	37.0	2201	0.10	0.12	0.07	0.00	0.05	0.10	0.17	0.02	0.07	0.07	0.15	0.07	2
CLINTON	7.7	1.9	49.5	2643	0.17	0.02	0.06	0.11	0.09	0.04	0.17	0.04	0.09	0.09	0.13	0.00	2
EAST HARLEM	7.6	1.1	9.7	2620	0.00	0.18	0.04	0.04	0.18	0.04	0.00	0.00	0.25	0.21	0.07	0.00	0
EAST VILLAGE	8.6	1.9	96.1	2695	0.05	0.00	0.08	0.08	0.08	0.10	0.13	0.10	0.15	0.18	0.00	0.05	2
FINANCIAL DISTRICT	8.1	1.8	42.8	2173	0.19	0.02	0.14	0.00	0.09	0.02	0.12	0.07	0.09	0.07	0.19	0.00	2
FLATIRON	8.6	2.3	119.7	2732	0.13	0.06	0.06	0.02	0.00	0.06	0.29	0.10	0.08	0.02	0.10	0.06	2
GRAMERCY	7.9	1.6	63.5	2661	0.13	0.03	0.07	0.03	0.13	0.00	0.17	0.03	0.13	0.13	0.10	0.03	2
GREENWICH VILLAGE	8.8	2.1	125.3	3033	0.09	0.04	0.07	0.07	0.00	0.07	0.34	0.04	0.05	0.09	0.05	0.09	2
HAMILTON HEIGHTS	7.2	1.2	9.6	2580	0.00	0.05	0.10	0.13	0.28	0.00	0.03	0.03	0.18	0.10	0.08	0.05	0
HUDSON YARDS	8.0	1.3	27.3	1788	0.00	0.00	0.50	0.17	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	5
INWOOD	7.4	1.2	9.6	2432	0.08	0.15	0.15	0.08	0.15	0.00	0.00	0.00	0.15	0.19	0.04	0.00	0
LENOX HILL	7.5	1.9	23.4	2463	0.03	0.07	0.10	0.07	0.10	0.03	0.21	0.03	0.05	0.11	0.03	0.16	2
LINCOLN SQUARE	7.7	2.0	55.4	2831	0.17	0.03	0.27	0.07	0.07	0.10	0.20	0.00	0.03	0.07	0.00	0.00	2
LITTLE ITALY	8.5	2.0	148.3	2748	0.05	0.07	0.12	0.10	0.02	0.07	0.33	0.07	0.02	0.07	0.05	0.02	2
LOWER EAST SIDE	7.5	1.3	53.8	2310	0.04	0.13	0.08	0.13	0.17	0.04	0.04	0.08	0.04	0.17	0.08	0.00	0
MANHATTAN VALLEY	8.0	1.4	31.3	2671	0.05	0.05	0.11	0.11	0.11	0.11	0.05	0.05	0.11	0.21	0.00	0.05	2
MANHATTANVILLE	7.0	1.4	16.2	2961	0.05	0.05	0.00	0.20	0.25	0.00	0.10	0.00	0.10	0.05	0.10	0.10	0
MARBLE HILL	7.1	1.1	3.0	2714	0.13	0.00	0.00	0.13	0.25	0.00	0.00	0.00	0.00	0.13	0.38	0.00	0
MIDTOWN	8.4	1.9	65.8	2546	0.16	0.12	0.04	0.04	0.06	0.08	0.04	0.10	0.06	0.08	0.16	0.08	2
MIDTOWN SOUTH	8.5	2.0	69.6	1467	0.21	0.12	0.12	0.03	0.00	0.03	0.15	0.15	0.00	0.06	0.09	0.03	4
MORNINGSIDE HEIGHTS	7.2	1.4	29.4	2473	0.13	0.00	0.22	0.09	0.22	0.00	0.04	0.00	0.04	0.17	0.09	0.00	0
MURRAY HILL	8.2	2.0	52.1	2412	0.13	0.04	0.08	0.06	0.02	0.06	0.08	0.10	0.06	0.13	0.13	0.10	2
NOHO	8.6	1.8	122.0	2646	0.05	0.05	0.05	0.00	0.04	0.09	0.22	0.07	0.13	0.15	0.05	0.09	2
ROOSEVELT ISLAND	6.1	1.3	7.3	2636	0.00	0.00	0.13	0.13	0.25	0.00	0.00	0.13	0.00	0.13	0.25	0.00	0
SOHO	8.5	2.1	116.6	3018	0.10	0.14	0.10	0.00	0.00	0.24	0.29	0.00	0.10	0.05	0.00	0.00	2
STUYVESANT TOWN	6.6	2.0	7.0	3079	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.50	0.00	1
SUTTON PLACE	7.8	2.3	22.4	2558	0.11	0.05	0.03	0.05	0.03	0.11	0.21	0.05	0.08	0.21	0.03	0.05	2
TRIBECA	7.9	2.0	64.1	2751	0.18	0.08	0.15	0.05	0.13	0.05	0.20	0.03	0.03	0.03	0.08	0.03	2
TUDOR CITY	6.7	1.4	18.5	2733	0.08	0.00	0.08	0.08	0.29	0.00	0.04	0.00	0.08	0.08	0.17	0.08	0
TURTLE BAY	7.7	2.0	36.7	2579	0.06	0.00	0.17	0.00	0.17	0.06	0.17	0.08	0.04	0.06	0.06	0.13	2
UPPER EAST SIDE	7.8	2.5	25.6	2859	0.13	0.06	0.02	0.02	0.13	0.08	0.33	0.02	0.04	0.08	0.02	0.06	2
UPPER WEST SIDE	8.1	1.8	52.0	2269	0.08	0.12	0.04	0.04	0.00	0.12	0.28	0.04	0.08	0.12	0.00	0.08	2
WASHINGTON HEIGHTS	7.0	1.2	8.8	2732	0.03	0.13	0.08	0.13	0.15	0.00	0.05	0.00	0.15	0.20	0.08	0.03	0
WEST VILLAGE	8.8	2.6	142.4	3186	0.17	0.02	0.04	0.04	0.00	0.10	0.33	0.10	0.06	0.08	0.04	0.04	2
YORKVILLE	7.7	1.7	18.3	2688	0.04	0.06	0.02	0.04	0.15	0.02	0.20	0.07	0.06	0.19	0.07	0.09	2