Neighborhood Segmentation And Venue Opening Suggestions in Manhattan

IBM DATA SCIENCE CAPSTONE
YELI WANG

Agenda

- ❑Why Neighborhood Segmentation
- ❑Data Collection and Cleaning
- ❑Venue Exploration
- ❑Neighborhood Segmentation
  - ❑K-Means Clustering
  - ❑Agglomerative Hierarchical Clustering
- ❑Discussion
- ❑Appendix

Why Neighborhood Segmentation

Manhattan ranks the top of the five boroughs of the New York City, in both its population density, and its economic significance.

Before open an restaurant, investor needs to consider:

- Neighborhood selection
- Price target
- Cuisine type

By segmentation, one can answer the questions above, and make justified decision.

Data Collection and Cleaning

Data Collection

- Manhattan neighborhood from JSON and Wikipedia, 40 neighborhoods
- Venue details from Foursquare API

Data Clearning

- Regular API ***explore***: Venue ID, name, location, category
- Premium API ***details***: Full address, tip counts, price tier, rating, createdAt

Feature Selection

	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

Venue Exploration

Parameters while exploring venues

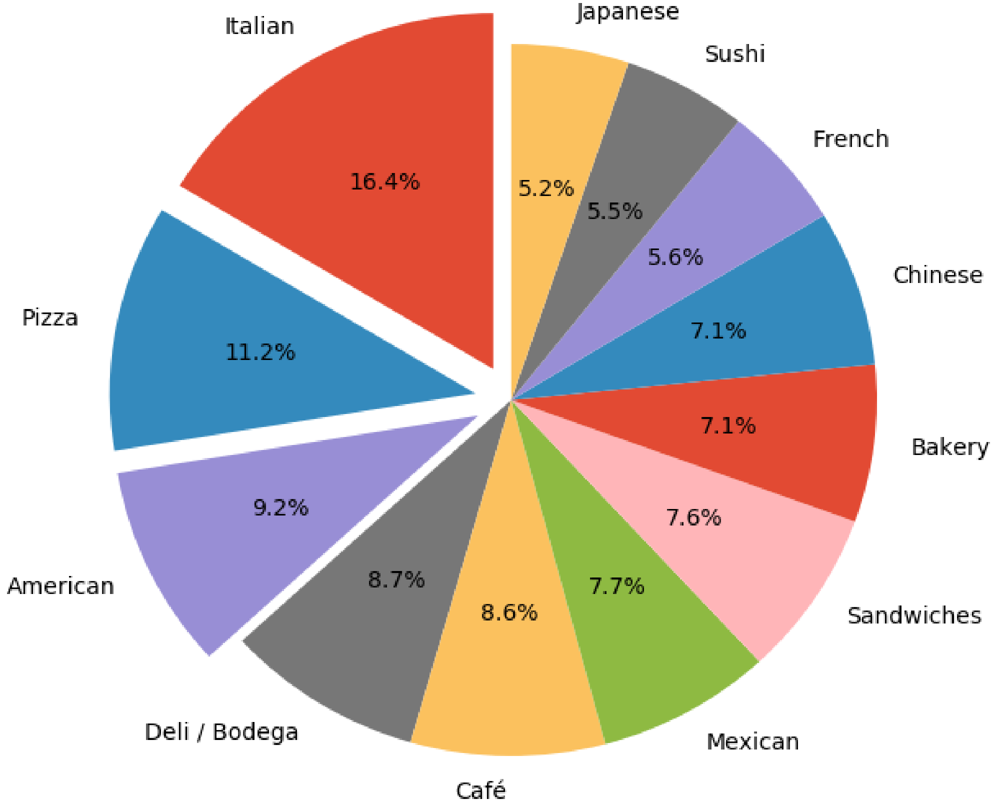
- 500m radius within the neighbor
- At maximum 100 venues returned

2,645 venues in total collected

Top 12 venues picked, 1,385 venues in total

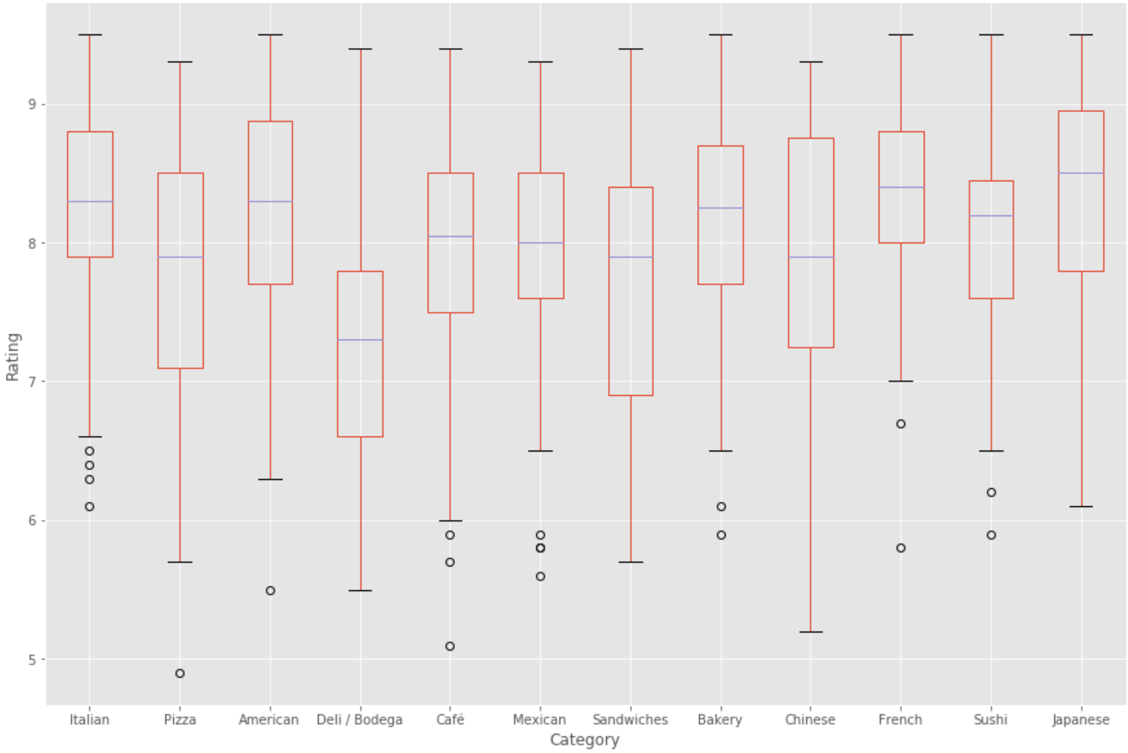
Top 3 venues

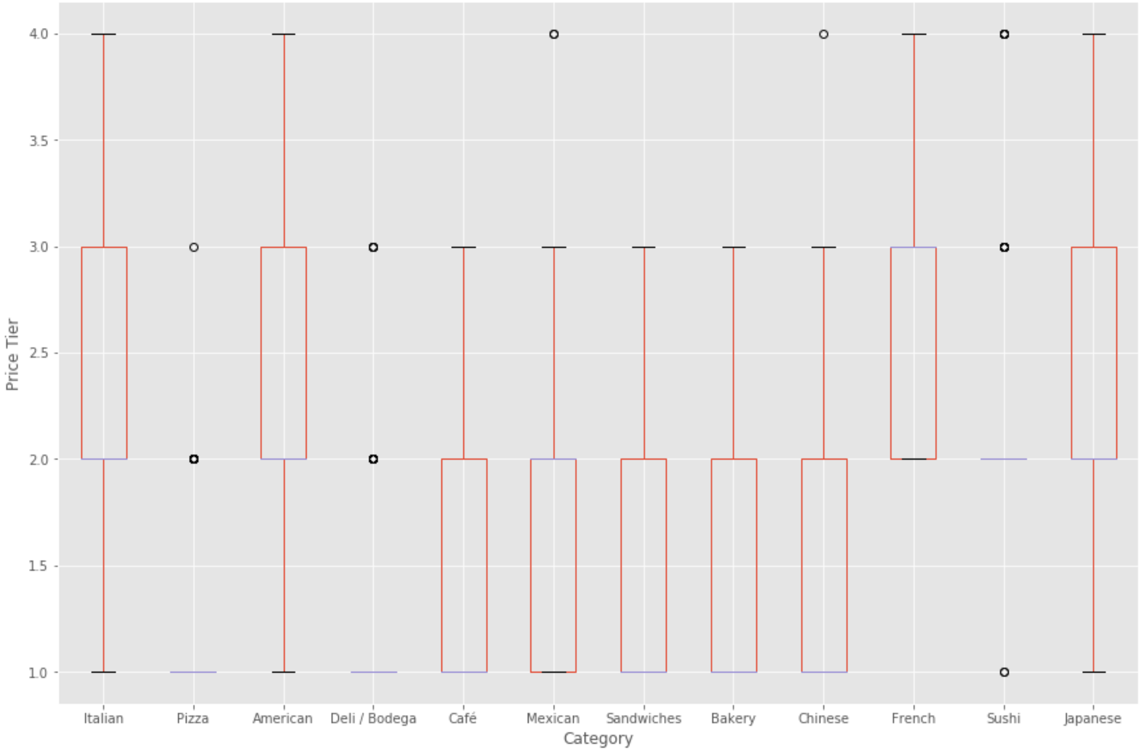
- Italian
- Pizza
- American



Venue Exploration

Italian, French restaurants receive higher and more concentrated reviews



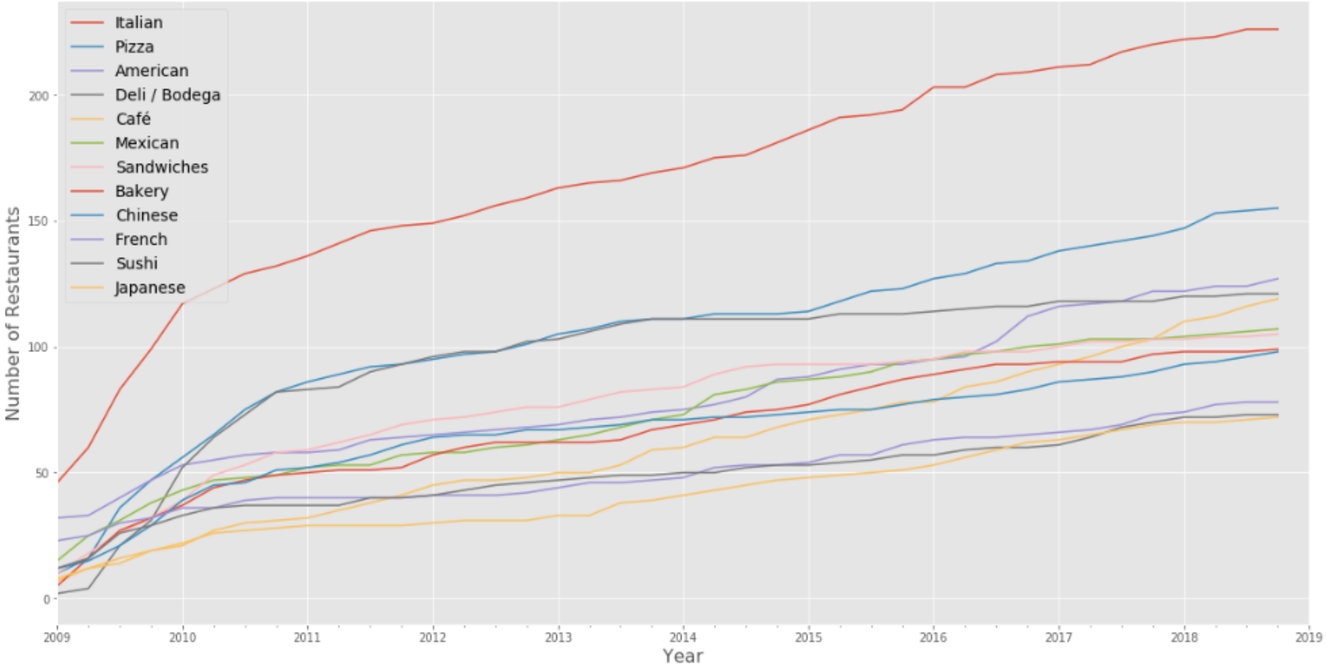


Venue Exploration

Restaurants expanded sharply from 2009 to 2011, and then growth rate stabled after 2011.

Exception: Café grew steadily for the past 10 years.

Data from 2003 to 2009 are excluded, probably because Foursquare data did not trace back then.



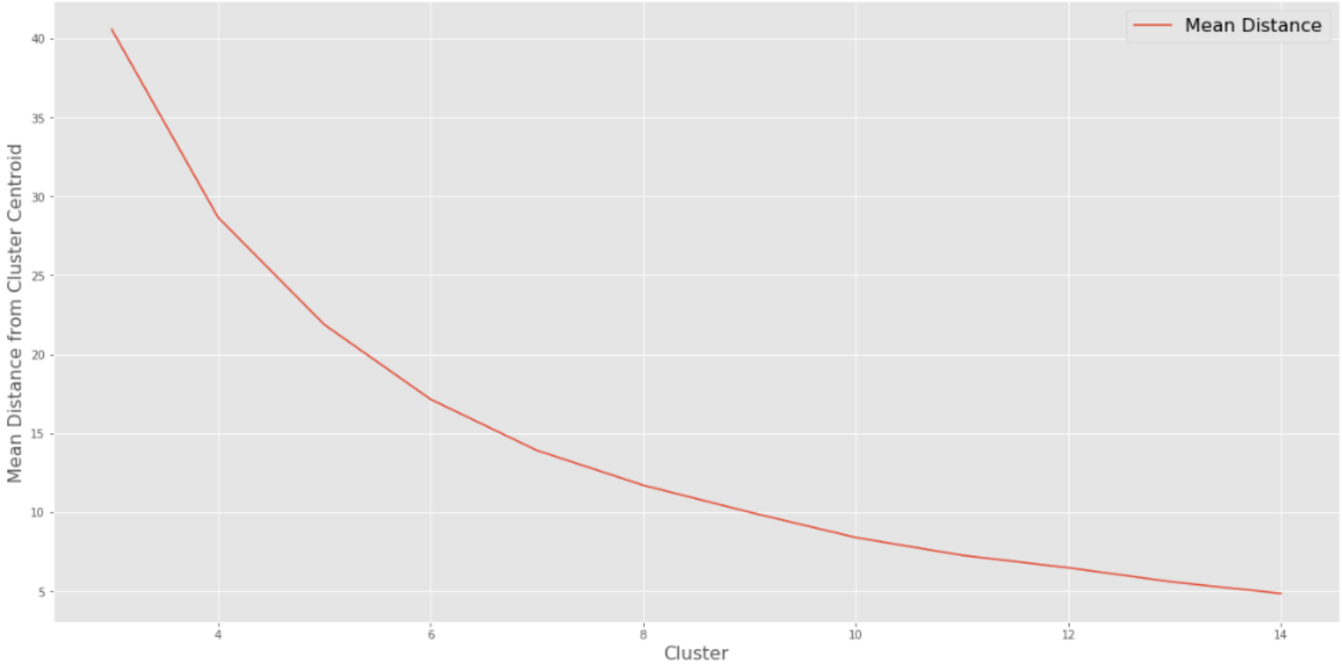
Neighborhood Segmentation

Choice of number of clusters

- Fewer clusters cannot differentiate each groups
- More clusters cause the centroids to converge to local minimal.

Mean distance of each data point to cluster centroid was measured

Elbow point at ***k = 6***



K-Means Clustering

Cluster 2, 3, 4

- Casual dining, deliveries
- Affordable price and average ratings

Cluster 0, 1, 5

- Luxurious venues
- Usually expensive, but top services

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

Agglomerative Hierarchical Clustering

Cluster 0, 3, 5

- Deli, Café, Pizza, and Chinese are most popular
- Price tier is generally 1

Cluster 1 - Outlier

Cluster 2, 4

- Italian, American and Japanese are most common
- Price tier is 2, and ratings are above the average

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

Discussion

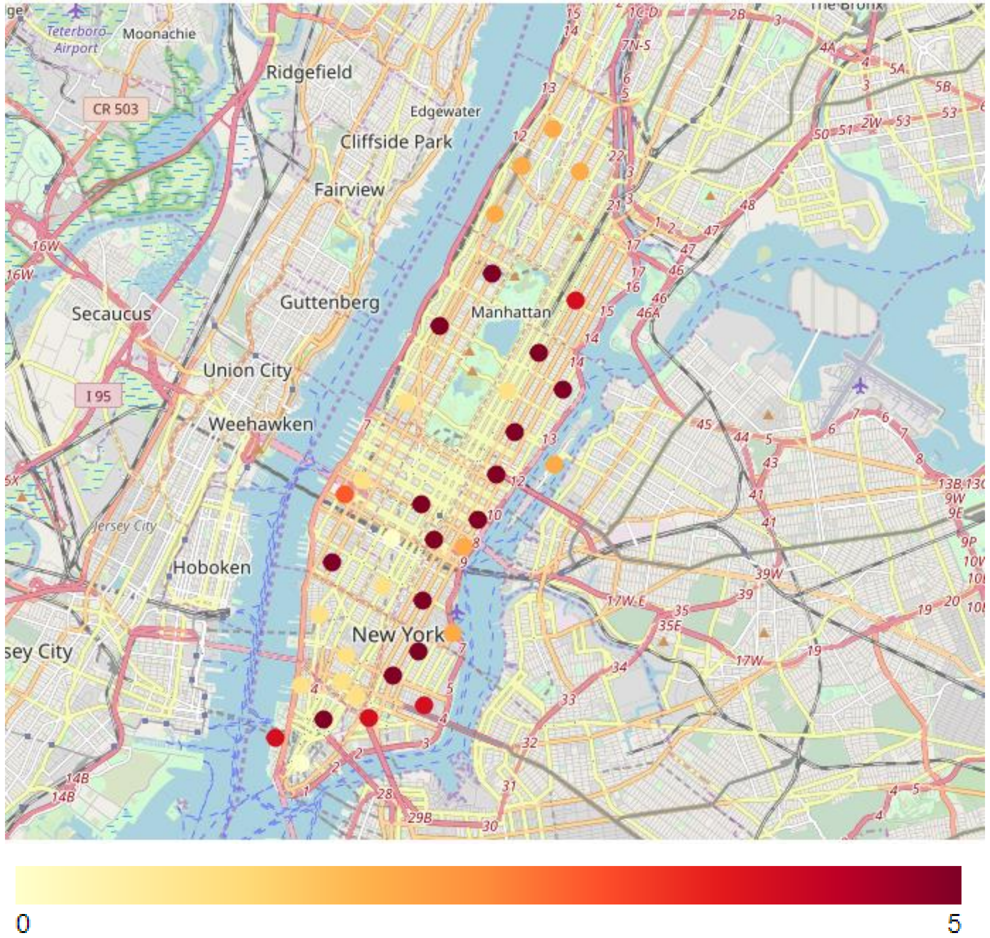
Each neighborhood has its unique combination of venues.

Restaurants should be opened where they are popular among the local residents, and fit the culture of that neighbor. And price should be set based on that area.

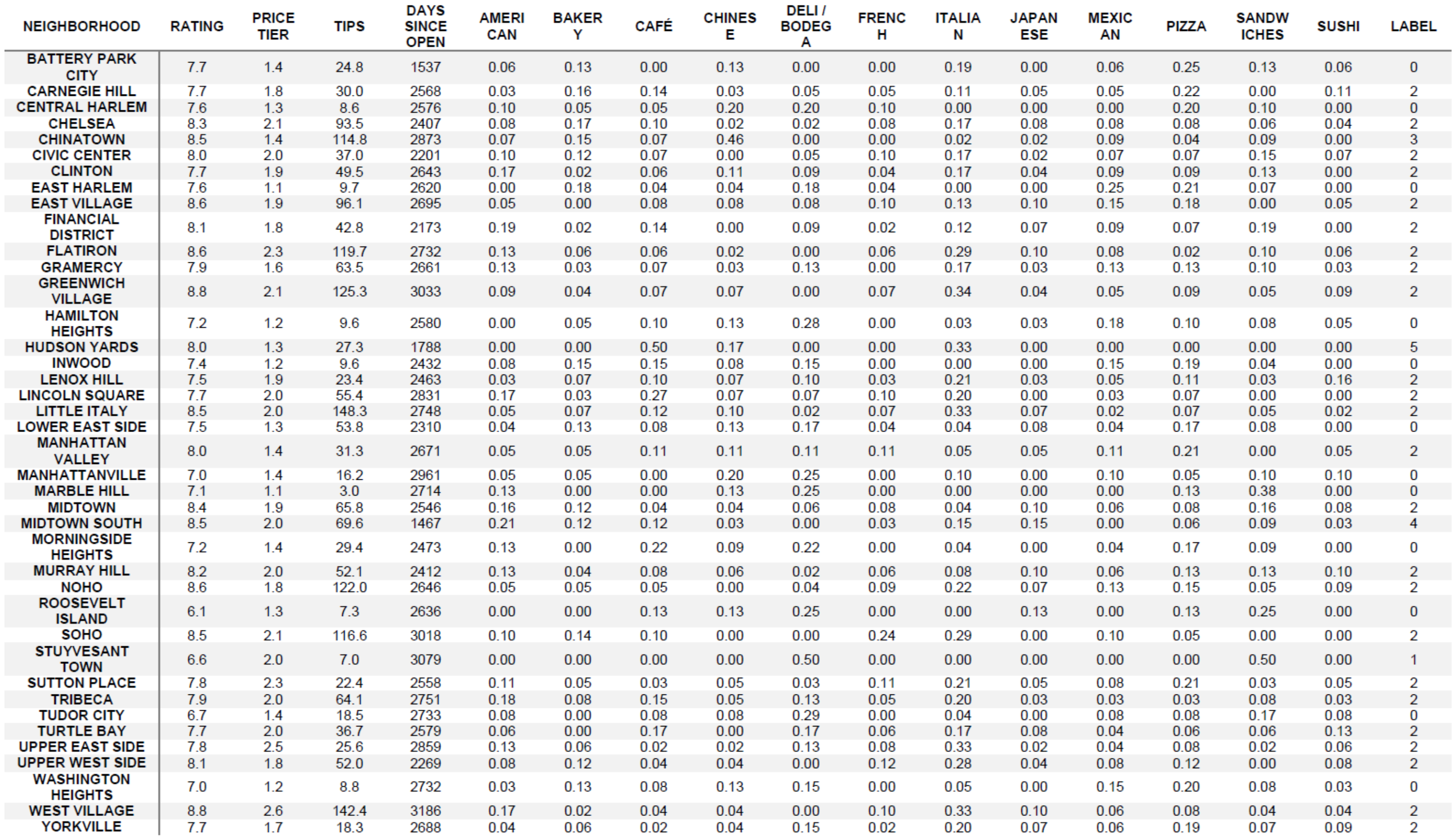
Refer to Appendix for detailed neighborhood segmentation.

Future works:

- Text mining on venue tips
- Research on what customer mostly care about



Appendix: List of Neighbor Clusters



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