The Battle of Neighborhoods

Where to launch a food delivery startup?

By Eylul Tekin

April 26th 2020

1. Introduction

1.1. Background

There is a new startup company that plans to launch an application of food delivery service similar to Postmates, Doordash and Uber Eats. To pilot their application, the company will launch their application in one of the cities listed below:

1. New York City
2. San Francisco
3. Los Angeles
4. Chicago

The goal of this new startup is to decide which metropolitan city is the better to invest and launch their first market. The success of a market can be predicted by various factors, one important factor being the availability of options to deliver from. The current trends in the market is another important predictor. Because the target audience of the startup is younger users who are quite familiar with online apps as well as food delivery services, their food preferences are of higher priority. Based on recent research, consumers between the ages of 18 and 34 are especially interested in trying diverse cuisines. In addition, survey research conducted by Technomic indicates that more than half of Americans are more interested in trying different cuisines than they were one year ago. Thus, the objective is to identify the best metropolitan city to launch the pilot online food delivery service based on various and diverse cuisine options available within each city.

1.2. Problem

Business Problem: Which city has a more diverse restaurant scene that will boost the potential success of the initial launch?

To address the business problem at hand, I will obtain different groups (or clusters) of restaurants within each city. The number of optimal clusters can be utilized as an indicator of a diverse restaurant scenes. Furthermore, the content of each cluster as well as the number of members within each cluster can be used as other indicators of diversity. For instance, if the clusters within a city are too similar, then we can conclude that the culinary scene is not as diverse. For this problem, I will use descriptive modeling to describe existing culinary scene within each city. More specifically, I will use k-means cluster analyses to determine how many clusters of restaurants there are within each city and the content and diversity of each cluster. The frequencies of different food venues within a city will be used in k-means cluster analyses.

2. Data

Location data as well as food venue data is necessary to address the business problem. I will first acquire location data for each city. Neighborhoods will be the level of analyses. Thus, I need a list of neighborhoods in each city as well as their coordinates (i.e., latitudes and longitudes). If the neighborhood and location information is readily available in a dataset, I will use that information. If not, I will use the library geopy to obtain this information. Below I present the sources where I obtained the neighborhood and location data from:

2.1. Data sources

2.1.1. New York City
For New York City, the list of neighborhoods and their locations were readily available in a dataset from the link:

https://cocl.us/new_york_dataset

2.1.2. San Francisco
To get the list of neighborhoods in San Francisco, I used the following Wikipedia link. I used library, BeautifulSoup, to get the neighborhood list.

https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco

2.1.3. Los Angeles
To get the list of neighborhoods in San Francisco, I used the following Wikipedia link. As with San Francisco, I used library, BeautifulSoup, to get the neighborhood list.

https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles

2.1.4. Chicago
To get the list of neighborhoods in Chicago, I used the following Wikipedia link.

https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago

2.2. Data cleaning

Neighborhood lists were gathered for each city from resources listed below. Except New York City, the coordinates of the neighborhoods were missing. For these cities, geopy was used to obtain coordinates. If no coordinates were returned or the returned coordinates did not belong to the city of interest, the neighborhood was excluded.

2.2.1. New York City
In the readily available neighborhood data for New York City, there were 306 neighborhoods presented along with their coordinates.

Table 1. New York City neighborhood data

	Neighborhood	Latitude	Longitude
0	Wakefield	40.894705	-73.847201
1	Co-op City	40.874294	-73.829939
2	Eastchester	40.887556	-73.827806
3	Fieldston	40.895437	-73.905643
4	Riverdale	40.890834	-73.912585

2.2.2. San Francisco

After eliminating extra rows (e.g., 'External links') from data extracted from Wikipedia, the initial neighborhood list for San Francisco had 119 neighborhoods. The numbers at the beginning of each neighborhood were also deleted. After using geopy to obtain coordinates, 83 neighborhoods along with their coordinates remained in the dataset.

Table 2. Before and after of San Francisco neighborhood data

	Neighborhood
0	1 Alamo Square
1	2 Anza Vista
2	3 Ashbury Heights
3	4 Balboa Park
4	5 Balboa Terrace

	Neighborhood	Latitude	Longitude
0	Alamo Square	37.7764	-122.435
1	Anza Vista	37.7808	-122.443
2	Balboa Park	37.7214	-122.448
3	Bayview	37.7289	-122.392
4	Belden Place	37.7917	-122.404

2.2.3. Los Angeles

After eliminating extra numbers appeared in data extracted from Wikipedia, the initial neighborhood list for Los Angeles had 199 neighborhoods. After using geopy to obtain coordinates, 145 neighborhoods along with their coordinates remained in the dataset.

Table 3. Before and after of Los Angeles neighborhood data

	Neighborhood
252	Windsor Square
253	Winnetka
254	Woodland Hills
255	Yucca Corridor
256	[50]

	Neighborhood	Latitude	Longitude
0	Angelino Heights	34.0703	-118.255
1	Angeles Mesa	34.1358	-118.08
2	Arleta	34.2413	-118.432
3	Arlington Heights	34.0435	-118.321
4	Arts District	34.0412	-118.234

2.2.4. Chicago

The initial neighborhood list for Chicago had 246 neighborhoods. After using geopy to obtain coordinates, 225 neighborhoods along with their coordinates remained in the dataset.

Table 4. Chicago neighborhood data

	Neighborhood	Latitude	Longitude
0	Albany Park	41.9719	-87.7162
1	Altgeld Gardens	41.6549	-87.6004
2	Andersonville	41.9771	-87.6693
3	Archer Heights	41.8114	-87.7262
4	Armour Square	41.84	-87.6331

3. Methodology

3.1. Data preprocessing

After gathering neighborhood and coordinate data and cleaning it up, I first used the services of Foursquare API to acquire the types of restaurant that exist within each neighborhood of the four cities. Given the goal of the project, I limited my search on Foursquare API to 'food venues'. I created new datasets for each city by gathering top 100 food venues within 1000 radius of a neighborhood's coordinates. In some cases, the neighborhood was residential and no venues were returned.

3.1.1. New York City

There were 138 unique food venue categories in New York City. Amongst 306 neighborhoods, 302 of them returned at least one food venue through Foursquare API.

Table 5. New York City venue data

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
1	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Jimbo's	40.891740	-73.858226	Burger Joint
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
5	Wakefield	40.894705	-73.847201	SUBWAY	40.890468	-73.849152	Sandwich Place
6	Wakefield	40.894705	-73.847201	E&L Bakery	40.893564	-73.856997	Bakery
7	Wakefield	40.894705	-73.847201	Popeyes Louisiana Kitchen	40.898292	-73.854719	Fried Chicken Joint
8	Wakefield	40.894705	-73.847201	Subway	40.897792	-73.855219	Sandwich Place
9	Wakefield	40.894705	-73.847201	Domino's Pizza	40.898443	-73.854851	Pizza Place

3.1.2. San Francisco

There were 118 unique food venue categories in San Francisco. Amongst 83 neighborhoods, 64 of them returned at least one food venue through Foursquare API. This is quite interesting, suggesting there are many residential neighborhoods in San Francisco.

Table 6. San Francisco venue data

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alamo Square	37.77636	-122.434689	The Mill	37.776425	-122.437970	Bakery
1	Alamo Square	37.77636	-122.434689	Nopa	37.774888	-122.437532	New American Restaurant
2	Alamo Square	37.77636	-122.434689	4505 Burgers & BBQ	37.776125	-122.438142	BBQ Joint
3	Alamo Square	37.77636	-122.434689	Souvla	37.774577	-122.437809	Souvlaki Shop
4	Alamo Square	37.77636	-122.434689	Bar Crudo	37.775707	-122.438019	Seafood Restaurant

3.1.3. Los Angeles

There were 104 unique food venue categories in Los Angeles. Amongst 145 neighborhoods, 142 of them returned at least one food venue through Foursquare API.

Table 7. Los Angeles venue data

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Angelino Heights	34.070289	-118.254796	Guisados	34.070262	-118.250437	Taco Place
1	Angelino Heights	34.070289	-118.254796	Tsubaki	34.072938	-118.251298	Japanese Restaurant
2	Angelino Heights	34.070289	-118.254796	Konbi	34.075383	-118.253893	Japanese Restaurant
3	Angelino Heights	34.070289	-118.254796	Leo's Tacos	34.067743	-118.260974	Taco Place
4	Angelino Heights	34.070289	-118.254796	Ostrich Farm	34.076272	-118.255919	American Restaurant

3.1.4. Chicago

There were 110 unique food venue categories in New York City. Amongst 225 neighborhoods, 224 of them returned at least one food venue through Foursquare API.

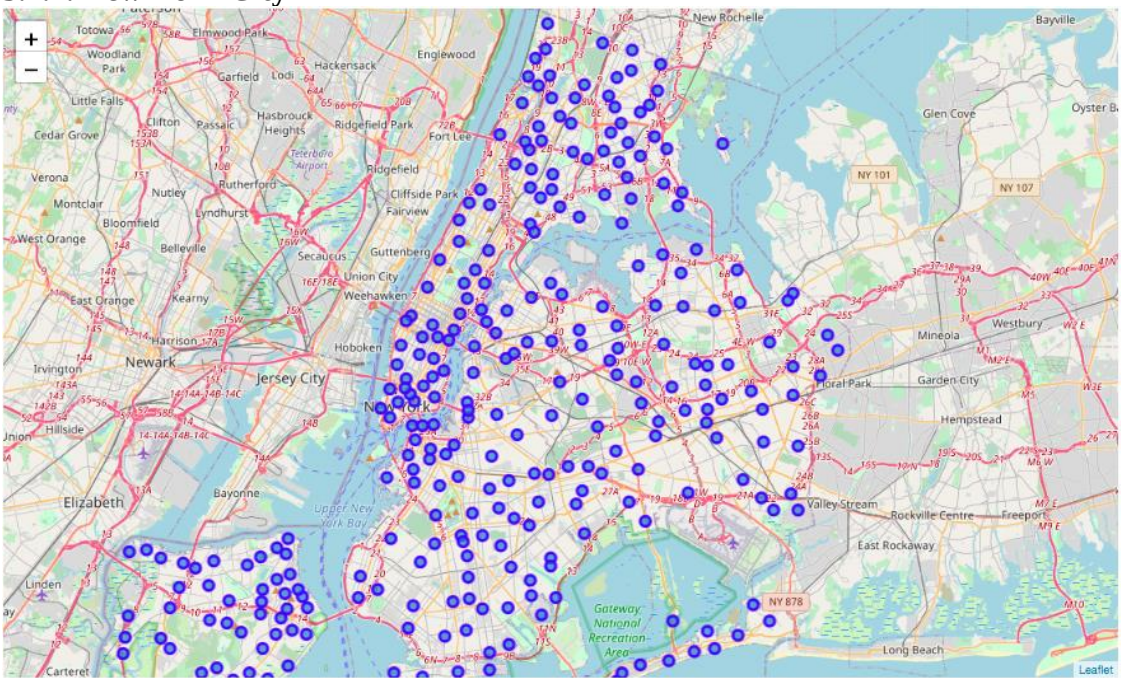
Table 8. Chicago venue data

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Albany Park	41.971937	-87.716174	Tre Kronor	41.975842	-87.711037	Scandinavian Restaurant
1	Albany Park	41.971937	-87.716174	Great Sea Chinese Restaurant	41.968496	-87.710678	Chinese Restaurant
2	Albany Park	41.971937	-87.716174	Merla's Kitchen	41.976063	-87.713559	Restaurant
3	Albany Park	41.971937	-87.716174	2 Asian Brothers	41.975832	-87.709655	Vietnamese Restaurant
4	Albany Park	41.971937	-87.716174	Peking Mandarin Resturant	41.968292	-87.715783	Chinese Restaurant

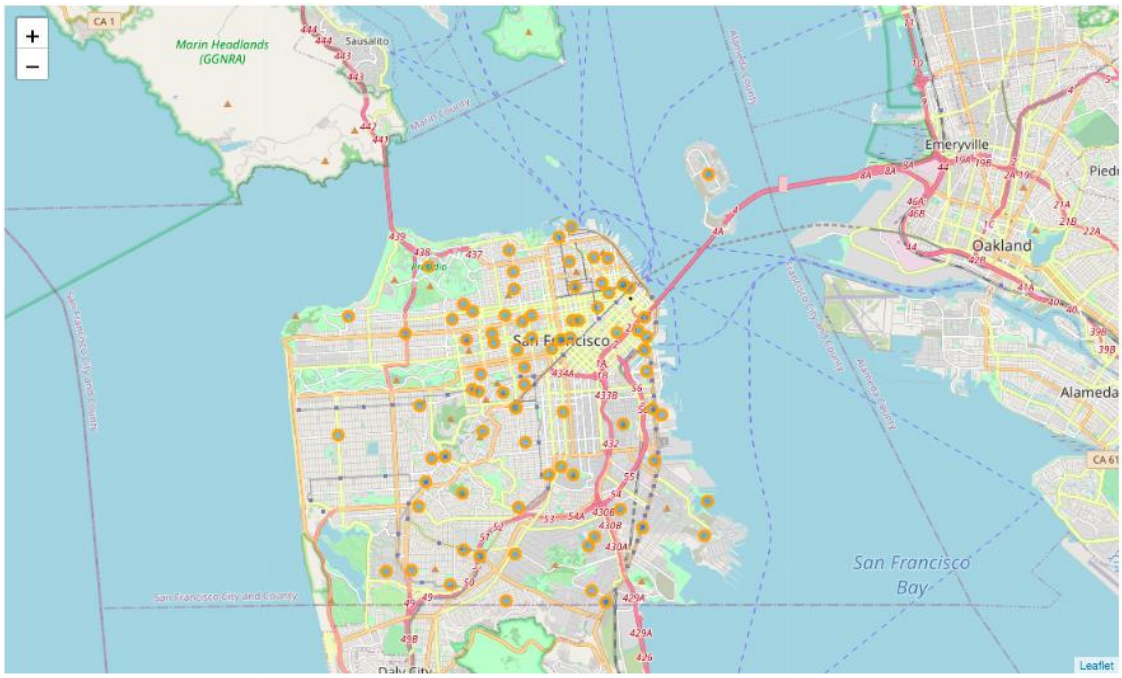
3.2. Data visualization

Before conducting the cluster analyses for each city, neighborhoods were mapped on the city using the library folium. This gave a preliminary idea of city density. Looking at the maps below, neighborhoods in New York City and Chicago have high density, whereas neighborhoods in Los Angeles and San Francisco have lower density.

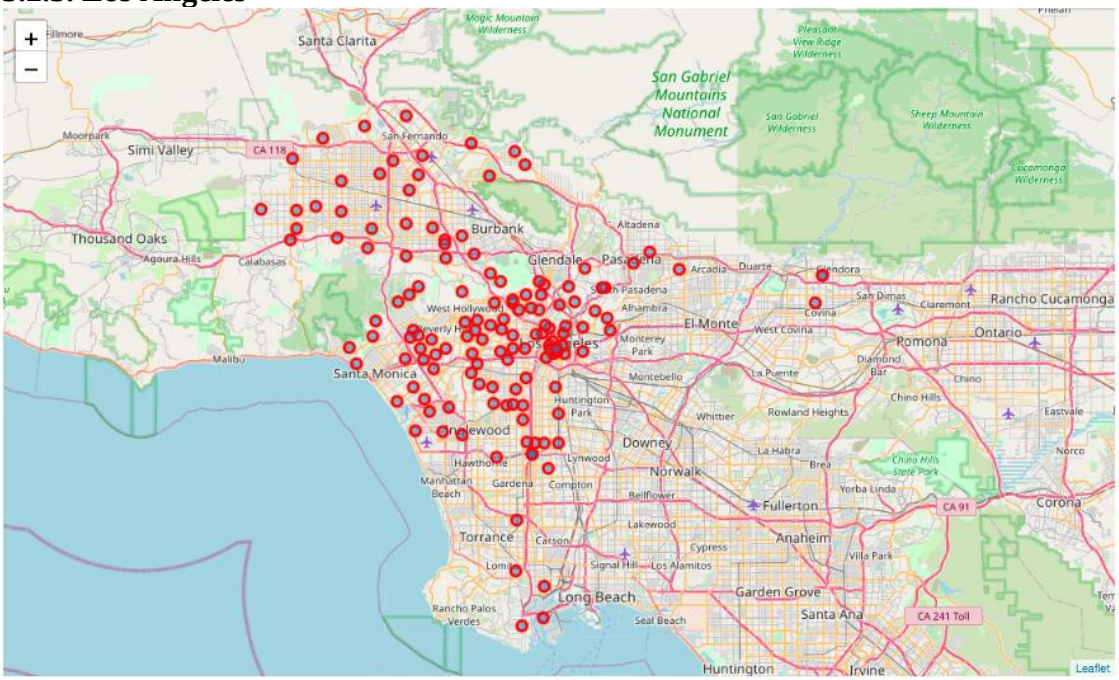
3.2.1. New York City



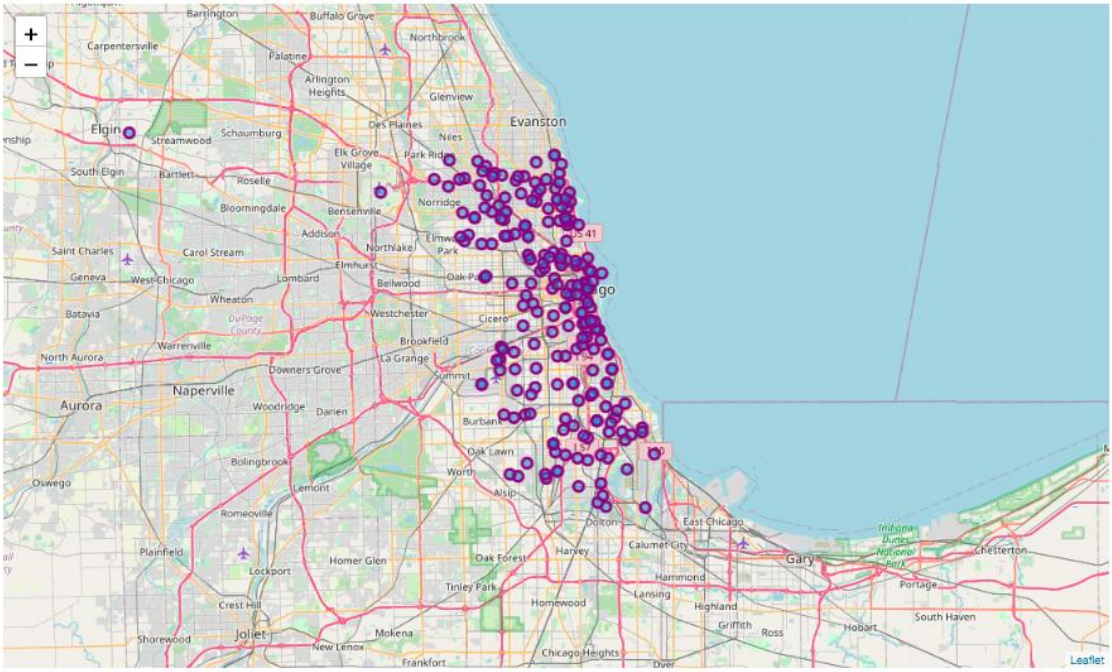
3.2.2. San Francisco



3.2.3. Los Angeles



3.2.4. Chicago



3.3. Data transformation

One hot encoding was used to dummy code the datasets based on food venue category. Then by grouping venue categories by neighborhoods, I calculated frequencies for each venue category within each neighborhood. By using frequencies, I aimed to identify cuisine clusters within each city and examine their frequency and diversity. An example of table from Chicago is provided below. According to this table, 5% of the food venues in Albany Park is categorized as Asian restaurants, and another 5% of them are categorized as bakery.

Table 9. The first two rows of Chicago venue frequency data

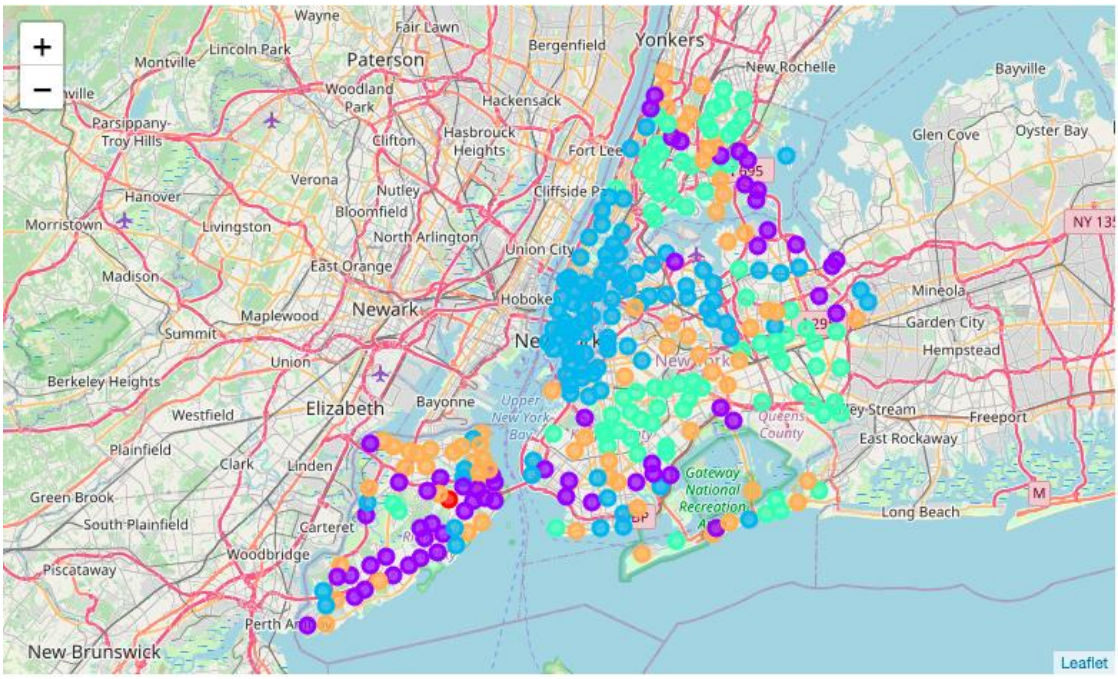
Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery
Albany Park	0.0	0.0	0.0	0.0	0.0	0.047619	0.015873	0.0	0.047619
Altgeld Gardens	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000

3.4. Cluster analyses

I employed the unsupervised k-means machine learning algorithm to find how the neighborhoods within each city were grouped based on food venue frequency. Because k-means cluster analysis returns the specified number of clusters regardless of k's optimality, I first used the elbow method to determine the optimal number of clusters for each city separately. The elbow method indicates the number of clusters that lead to highest information gain from the dataset. After that number, the explained variation can still increase by adding new clusters but 1) the increase is not as high, and 2) increasing the number of clusters makes it harder to describe the dataset. After finding the optimal number for each city, I conducted k-means cluster analysis using it.

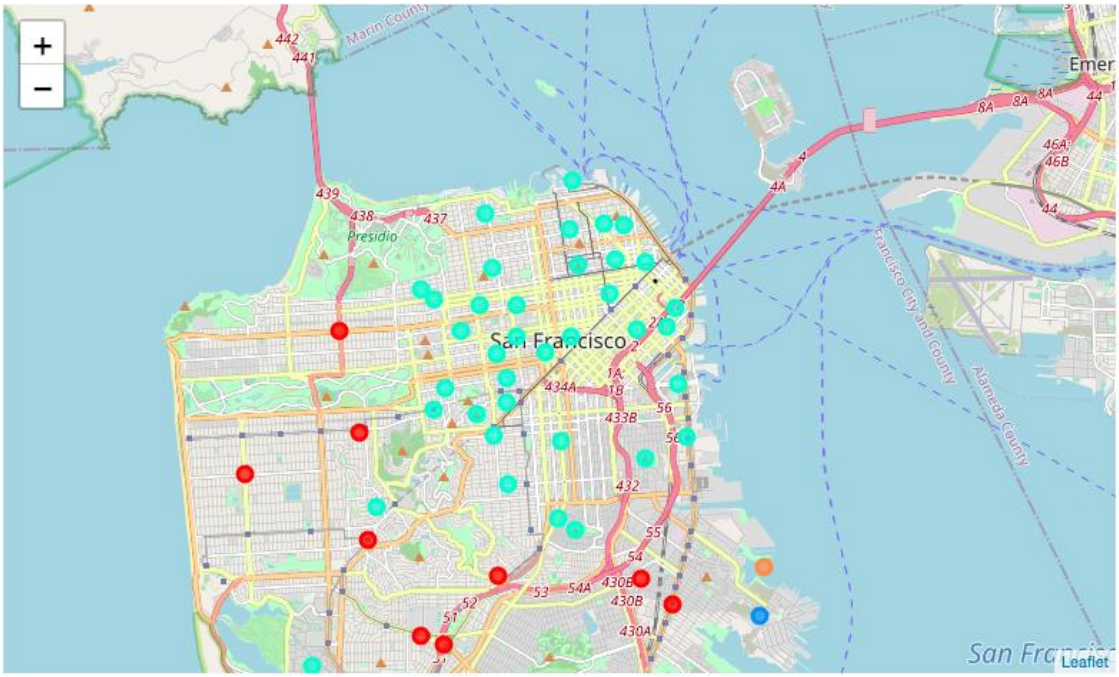
3.4.1. New York City

According to the elbow method, the optimal number of clusters was 5 in New York City.



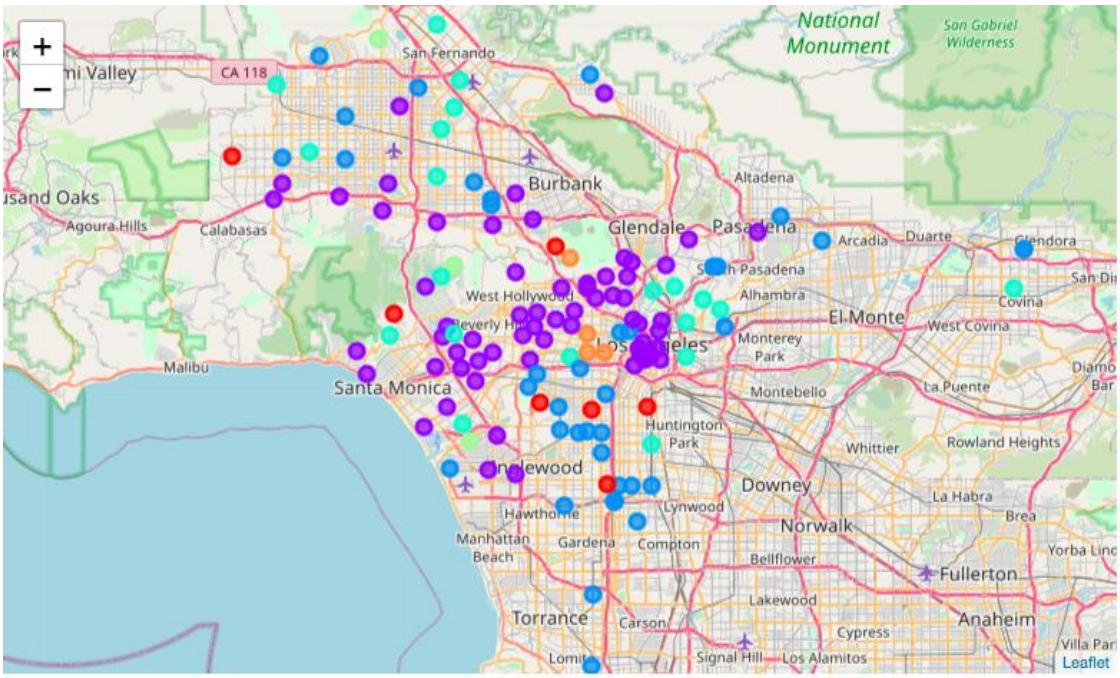
3.4.2. San Francisco

According to the elbow method, the optimal number of clusters was 6 in San Francisco. However, the k-means returned some NaN clusters and no members for Cluster 1.



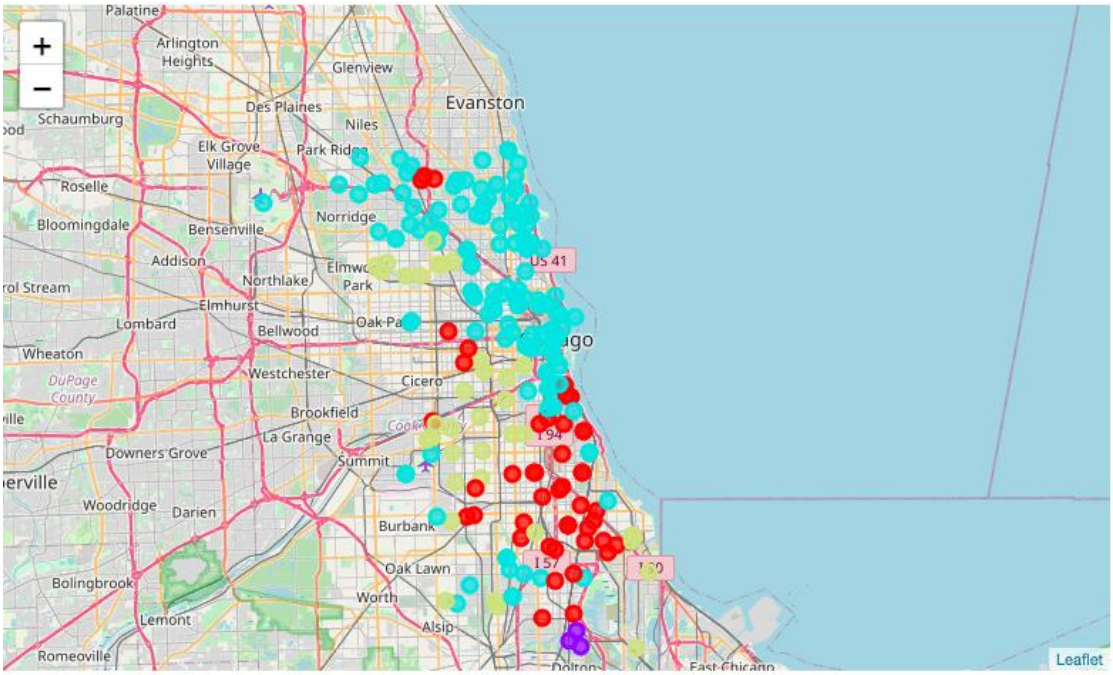
3.4.3. Los Angeles

According to the elbow method, the optimal number of clusters was 6 in Los Angeles.



3.4.4. Chicago

According to the elbow method, the optimal number of clusters was 4 in Chicago.



3.5. Cluster examination

After cluster analyses, cluster labels were obtained for each neighborhood. To examine the content of each cluster within a city, I created a dataset that showed the frequencies of each food venue category for each cluster. Below is the example of such dataset for New York City. Using such datasets for each city, I plotted top 15 most frequent venue categories for each cluster.

Table 10. New York City cluster frequency data

Cluster Labels	0	1	2	3	4
Afghan Restaurant	0.0	0.000000	0.000526	0.000000	0.000786
African Restaurant	0.0	0.000345	0.001010	0.001533	0.000000
American Restaurant	0.0	0.023077	0.043204	0.016217	0.047149
Arepa Restaurant	0.0	0.000213	0.001883	0.000122	0.000777
Argentinian Restaurant	0.0	0.000000	0.002561	0.000000	0.000000

3.5.1. New York City

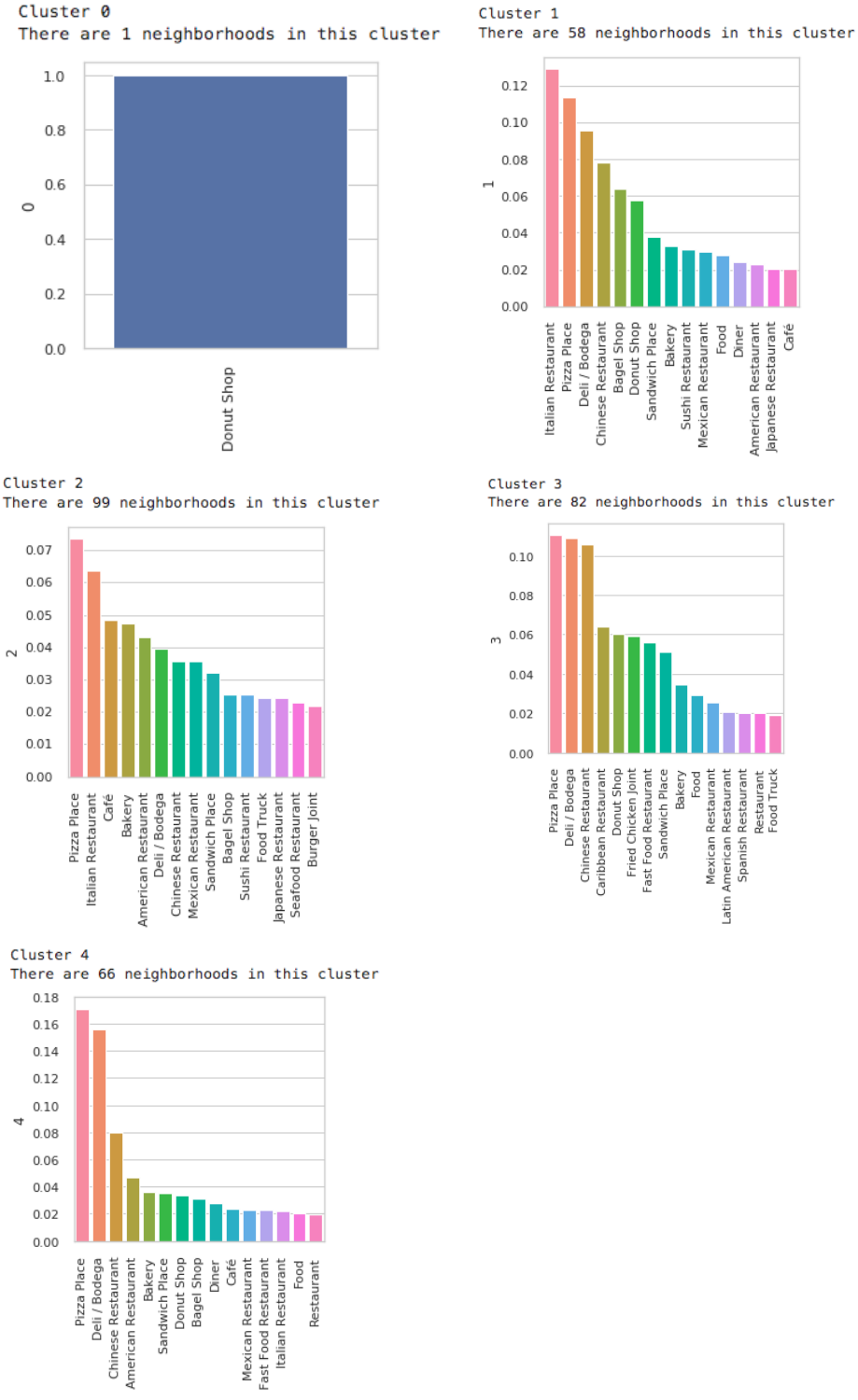


Figure 1. Cluster examination for New York City

3.5.2. San Francisco

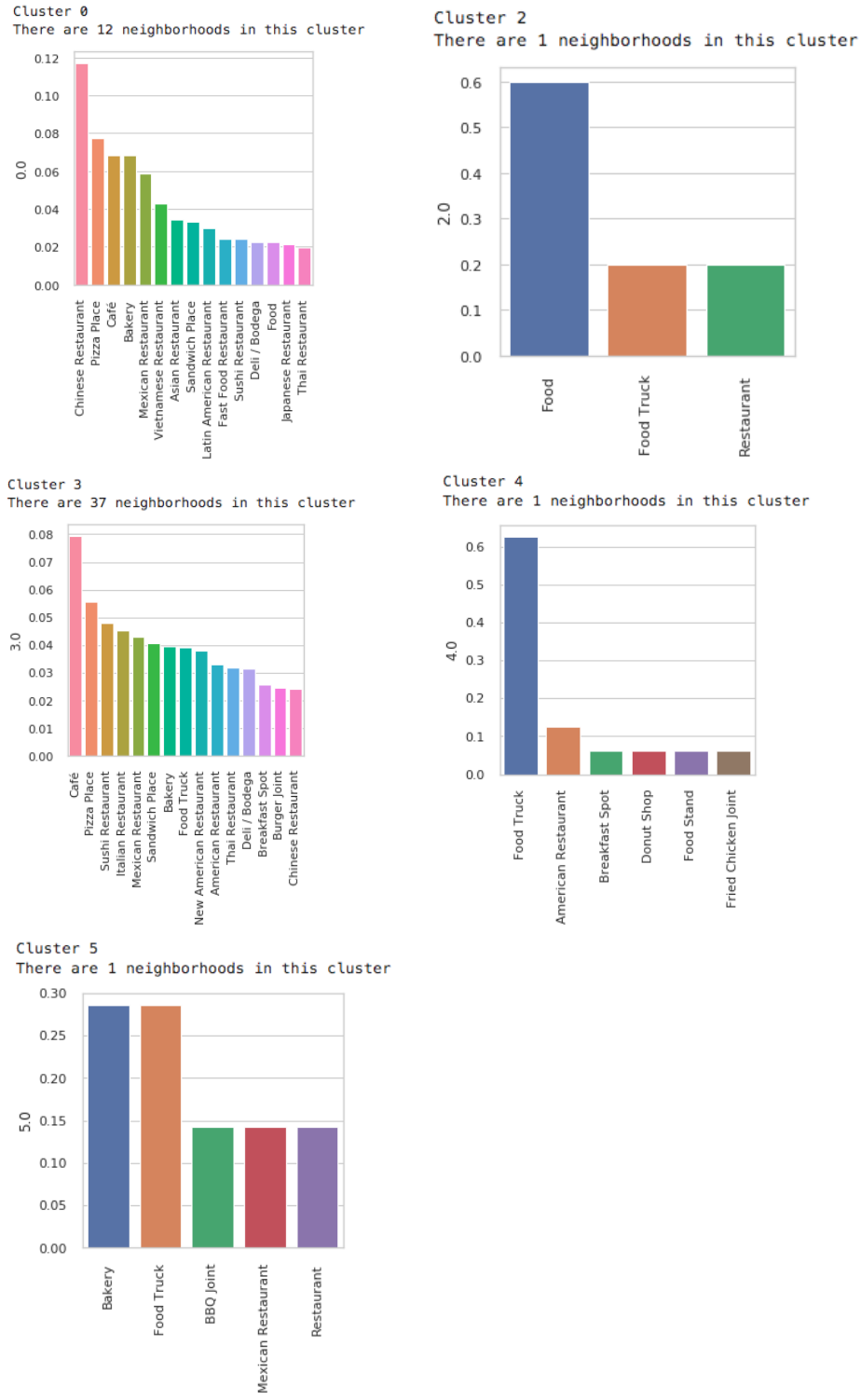


Figure 2. Cluster examination for San Francisco

3.5.3. Los Angeles

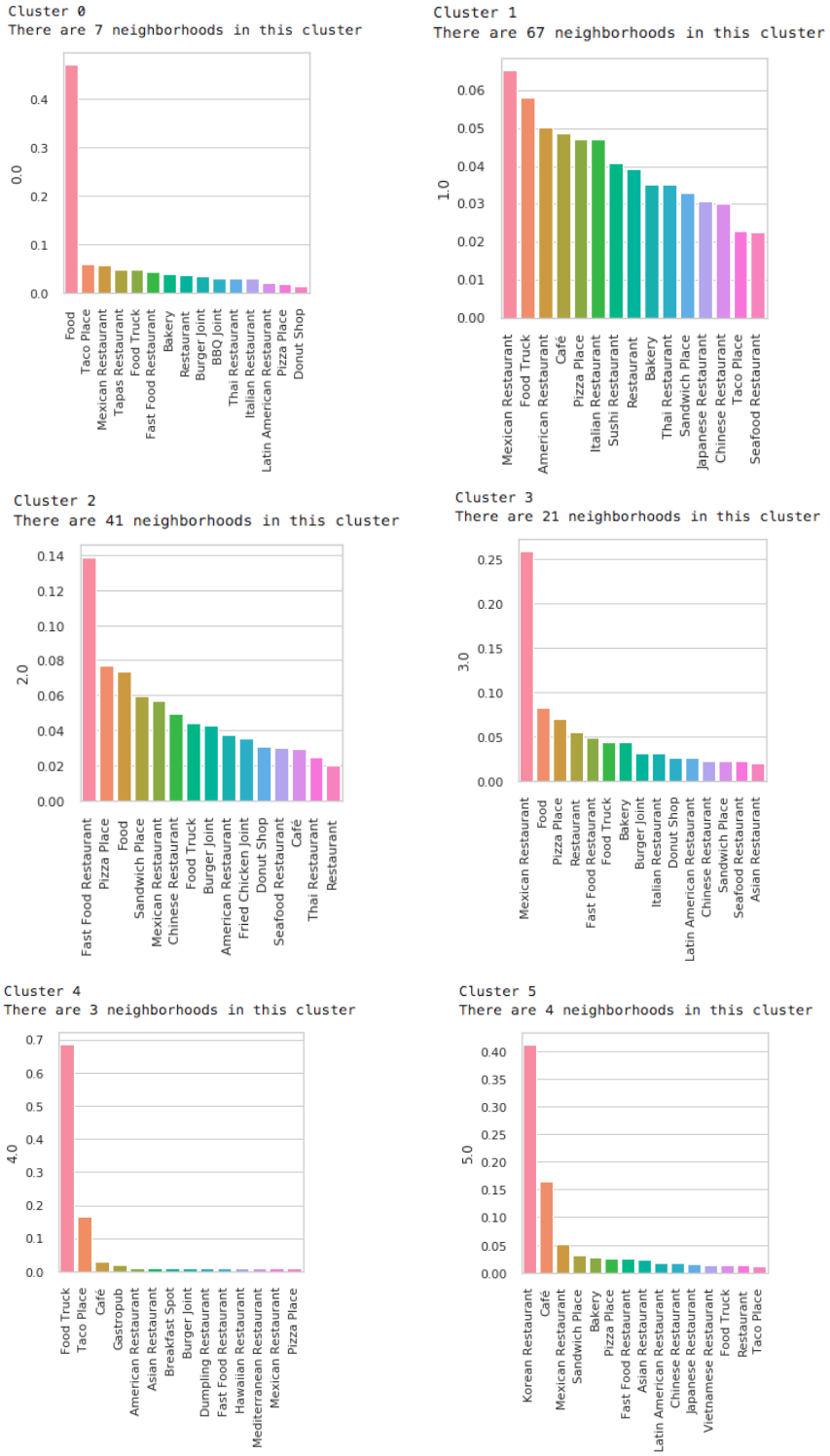


Figure 3. Cluster examination for Los Angeles

3.5.4. Chicago

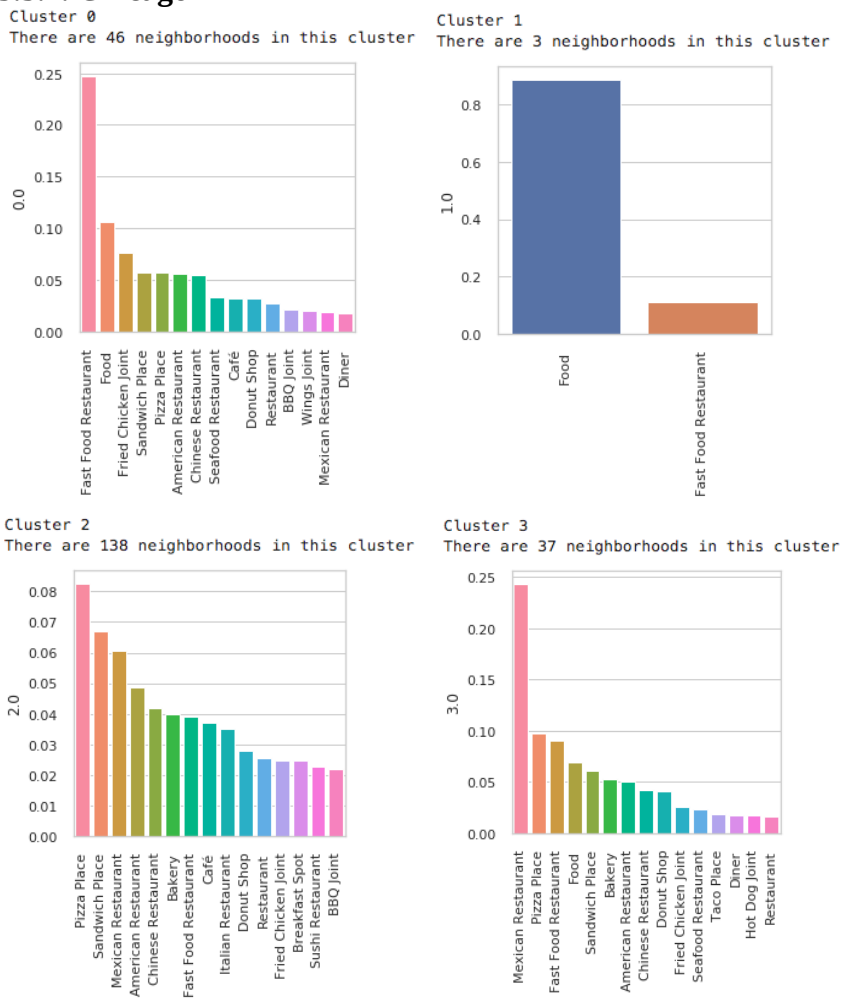


Figure 4. Cluster examination for Chicago

4. Results

In this project, I conducted four k-means cluster analyses for four cities to describe restaurant clusters within these cities. Below I report what I have found for each city.

4.1. New York City

There are five clusters in New York City; however, Cluster 0 is not informative. There is only one neighborhood in this cluster with a single donut shop. Examining remaining four clusters, Italian restaurants, pizza places and delis dominate all four of them. Cluster 1 has 58 neighborhoods, and relatively high numbers of Chinese restaurants as well as bagel and donut shops. Cluster 2 has 99 neighborhoods and is the most diverse cluster. Besides Italian

restaurants, pizza places and delis, it offers cafes, bakeries, and decent amount of American, Chinese and Mexican restaurants as well as sandwich places. Cluster 3 has 82 neighborhoods and offers Caribbean cuisine as well as fried chicken and fast food options. Lastly, Cluster 4 has 66 neighborhoods, and is not much diverse. The restaurant scene is mostly dominated by pizza places and delis, followed by Chinese restaurants.

4.2. San Francisco

Although I conducted a k-means analysis with six clusters for San Francisco, the algorithm found only five clusters in San Francisco and did not cluster any neighborhoods within Cluster 1. Furthermore, there were multiple neighborhoods that were not identified as parts of any clusters. Examining existing clusters, we see that Cluster 0 has 12 neighborhoods and is relatively diverse. Although there are a lot of Chinese restaurants, other options as pizza, Mexican and Vietnamese restaurants are available. Cluster 3 is the most diverse cluster in San Francisco. There are similar amounts of pizza, sushi, sandwich options as well as Italian and Mexican restaurants. Clusters 2 and 4 only have a single neighborhood and are not informative.

4.3. Los Angeles

There are six clusters in Los Angeles; however, Clusters 0 and 4 are not informative because the most frequent categories are ‘food’ or ‘food truck’. Cluster 1 has 67 neighborhoods in it, and is quite diverse with Mexican, American, Italian, Japanese, Thai, Chinese, pizza and sushi restaurants. Cluster 2 has 41 neighborhoods that most frequently have Mexican and Chinese restaurants as well as fast food, and pizza and sandwich places. Cluster 3 includes 21 neighborhoods and is dominated by Mexican places without much diversity. Lastly, Cluster 5 has 4 neighborhoods and is dominated by Korean restaurants.

4.4. Chicago

There are four clusters in Chicago; however, Cluster 1 is not informative. Examining other clusters, Cluster 0 has 46 neighborhoods which are dominated by fast food restaurants followed by fried chicken, pizza and sandwich places as well as Chinese and American restaurants. Cluster 2 is the most crowded cluster with 136 neighborhoods and offers pizza and sandwich places as well as Chinese, Mexican and American restaurants. Lastly, Cluster 3 is dominated by Mexican restaurants followed by sandwich and fast food places.

5. Discussion

Our business problem was to determine the city with more diverse cuisine to launch a new food delivery startup. A direct way to determine diversity of the restaurant scene is to look at the number of different clusters. By doing this, we can assume that the more clusters there are, the more diversity there is. Using this method, we would conclude that LA is the city with more diverse cuisine because it has 6 clusters. However, only using cluster number without examining the content might lead to wrong conclusions. For instance, some clusters do not give much information about cuisine diversity. Therefore, after conducting the cluster analyses, I also

examined the content of each cluster to understand the available cuisines as well as the number of neighborhoods in each cluster.

Based on this metric, I recommend Los Angeles for the launch of the startup because Los Angeles has the most differences amongst clusters. It also has a very diverse cluster that includes 66 neighborhoods. This suggests that many restaurant options from various categories are available within that cluster.

6. Conclusion

In this project, I used cluster analysis to describe frequencies of different food venue categories within four metropolitan cities and made a location recommendation. Cluster analysis allows us to describe our data in a more meaningful way and similar business problems can be solved by using the approach I demonstrated here. To name a few examples, these business questions might be where to open a certain type of restaurant, or how similar and dissimilar certain neighborhoods or cities are.