IBM Data Science Capstone Project
– Battle of Neighborhoods –

Final Report

Getting the right locations for a new fine cuisine Restaurant in Chicago

Krisztián Fintor

21th of February, 2021

1. Introduction

This project is about using data science toolset on a real-life problem and demonstrating the creation of value by applying the learned skills. This report presents this capstone project.

2. Problem definition and background

A European restaurant chain owner wants to open a new restaurant overseas notably in Chicago in the USA.

The investor operates the increasingly popular fine cuisine restaurants, specializing in a special vegan menu offers.

He chose this big city based on the basis of preliminary market research of the Midwest region, but he wants to learn more about the conditions within the city so that he can choose a suitable place to open his restaurant.

The investor prefers those areas that are frequently visited by tourists and local residents, and they're gastronomically popular, but there aren't many fine cuisine restaurants.

3. Audience, stakeholders

Despite the fact that the current project is about a single investor, the problem and the proposed solution can be well applied to meet the needs of companies and investor groups.

4. Data and data source

The following data is required for the successful implementation of the project:

- List of community areas of Chicago. Data source: https://en.wikipedia.org/wiki/Community_areas_in_Chicago

- Location data (geo-coordinates) of community areas of Chicago. Data source: geo-coordinates of community areas will be obtained by using Nominatim Geocoding service from Geopy library in the notebook.

- Top venues of community areas. Data source: this data will be obtained from Foursquare through an API.

5. Methodology, Explanation of data usage

5.1. Outline schedule

1. The first steps in achieving the project objectives will be to obtaining, cleaning, sorting and exploring data.

2. After that exploring the top venues in community areas by Foursquare will be the next step.

3. The K-means unsupervised machine learning technique will be used for creating clusters of community areas. In order to choose a relevant number of clusters silhouette scoring will be used.

5.2. Preparation and exploration of obtained data

After the names of the community areas of Chicago (hereinafter referred to as 'districts') have collected and added to a data frame, checking and correcting of names of the districts will be executed. The unnecessary rows are also will be deleted from the data frame.

Collecting of geographical coordinates (latitude, longitude) of districts will be the next step based on district names using a geocoding process. Postal codes are used in those cases where coordinates cannot be able to get by using name of the district.

In order to avoid multiple districts with the same geographic coordinates, a checking for duplicates was carried out on the data series. In case of multiple districts with the same coordinates the names of the districts have been merged into one coordinate.

At the end of the preparatory operations, the number of districts decreased from the original 77 to 72 (Fig. 1).

	Name	Latitude	Longitude
0	Rogers Park	42.009574	-87.675550
1	West Ridge	41.879788	-87.633113
2	Uptown	41.969450	-87.660513
3	Lincoln Square	40.148032	-89.363308
4	North Center, North Park	41.858657	-87.612199
...	...	...	...
67	Washington Heights	41.705596	-87.655931
68	Mount Greenwood	41.691818	-87.699001
69	Morgan Park	41.885592	-87.651928
70	O'Hare	41.977921	-87.903141
71	Edgewater	41.999149	-87.657370

72 rows × 3 columns

Figure 1.: *Prepared data frame of district locations*

5.3. Visualization of locations

In order to visualize positions of the districts we used geographical coordinates of Chicago and the folium library (Fig. 2.).

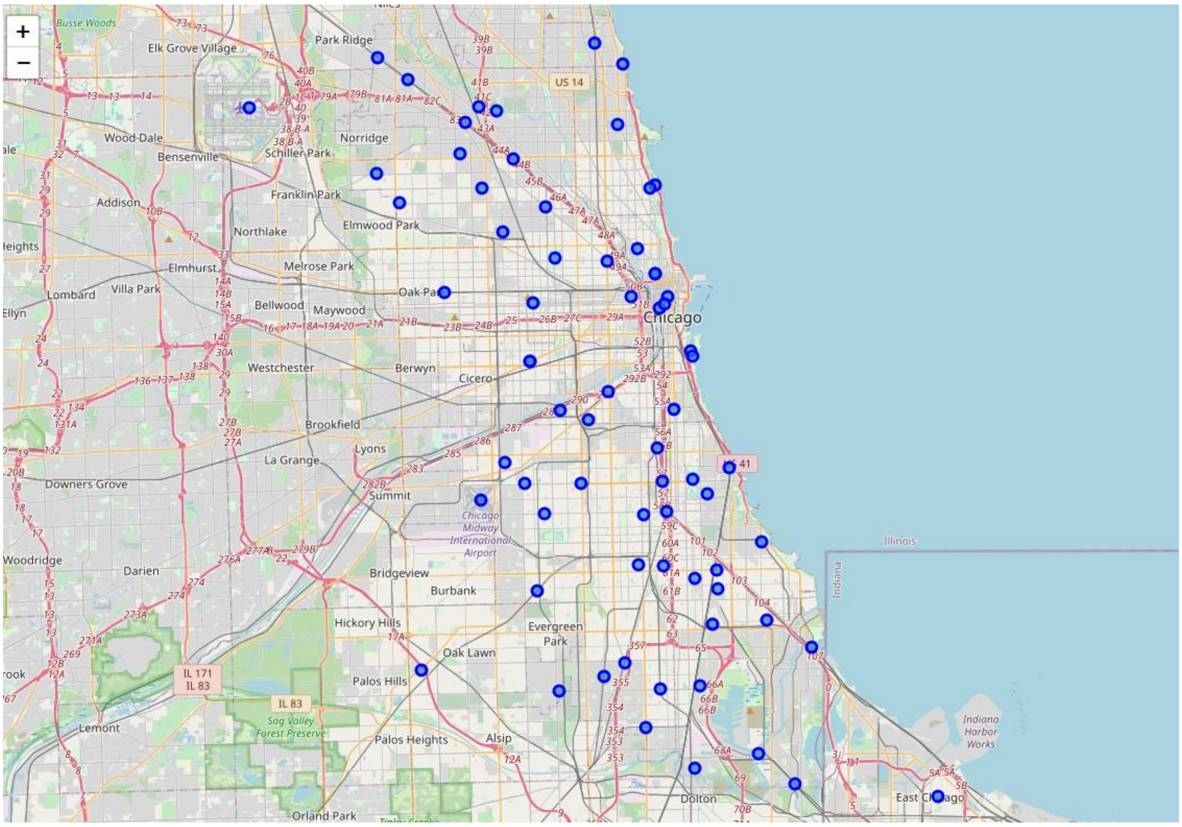

Figure 2.: *Location of districts on Chicago map*

5.4. Top venues are in the neighborhood of districts locations.

In this step the top venues will be collected from the neighborhood of each districts. The venues will be collected by using Foursquare API. Data from Foursquare are received in json format, and after rearranging the data we have up to 100 venues in the neighborhood of each districts. Collecting of venues have been carried within the 1500 m radius of the coordinates of the districts. The collected and arranged data contains some basic information about the collected venues including its name, location coordinates and category (Fig. 3.).

	District Name	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park	42.009574	-87.67555	El Famous Burrito	42.010421	-87.674204	Mexican Restaurant
1	Rogers Park	42.009574	-87.67555	Taqueria & Restaurant Cd. Hidalgo	42.011634	-87.674484	Mexican Restaurant
2	Rogers Park	42.009574	-87.67555	Bark Place	42.010080	-87.675223	Pet Store
3	Rogers Park	42.009574	-87.67555	Morse Fresh Market	42.008087	-87.667041	Grocery Store
4	Rogers Park	42.009574	-87.67555	Mind Crusher Tattoo	42.003801	-87.672525	Tattoo Parlor

Figure 3.: *Top venues of districts of Chicago*

By this procedure we got 342 unique venue category in Chicago. Then we check and put in a data frame how many pieces of each venue category are found in each district by using the one hot encoding procedure. Grouping rows by districts and by taking the mean of frequency of occurrence of each category will be the next one.

We quickly can check the top 5 most common venue categories in each district by the mean of frequency of occurrence. By using the mean frequency of occurrence of each venue categories a new data frame that contains the top 10 most common venues in each district can be created (Fig. 4.).

	District Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park	Pizza Place	Park	Middle Eastern Restaurant	Hookah Bar	Sandwich Place	Ice Cream Shop	Supermarket	Mobile Phone Shop	Donut Shop	Coffee Shop
1	Archer Heights	Mexican Restaurant	Pizza Place	Discount Store	Sandwich Place	Bank	Fast Food Restaurant	Video Store	Bar	Grocery Store	Bakery
2	Armour Square	Coffee Shop	Sandwich Place	Bookstore	Park	Café	History Museum	Grocery Store	Sushi Restaurant	Thai Restaurant	Pet Store
3	Ashburn	Park	Sandwich Place	Pizza Place	Mexican Restaurant	Fast Food Restaurant	Fried Chicken Joint	Seafood Restaurant	Pharmacy	BBQ Joint	Furniture / Home Store
4	Auburn Gresham	Seafood Restaurant	Discount Store	Fried Chicken Joint	Fast Food Restaurant	Park	Pharmacy	Grocery Store	Bus Station	Hot Dog Joint	Dive Bar

Figure 4.: *Top 10 most common venues in districts of Chicago (first vive rows)*

5.5. Filtering of venue categories

We don't need all of the 342 venue categories. Only those venue categories are important that can be decisive in selecting the right venue for a new restaurant. As I have already described in the problem definition chapter, I prefer locations that are in vogue from a gastronomic point of view, where only a few fine cuisine restaurants are there.

Hence, venues of restaurants are extremely important part of the data. Another important aspect is accessibility, so transport hubs (e. g. Bus Stations, Train Stations, Airport area) where many people arriving to the city are also favorable. Those districts where many hotels can be found are also important because many tourists can be found in the vicinity of hotels. Screening on the basis of the above-mentioned criteria has narrowed the list of specific venues to be taken into account to 60 venue categories.

5.6. Clustering

After filtering we have a dataset that are appropriate for clustering. For clustering we will use the K-Means clustering which is an unsupervised machine learning algorithm. In order to avoid the trial and error approach, the silhouette score was used to helping determine the right number of clusters (Fig. 5.).

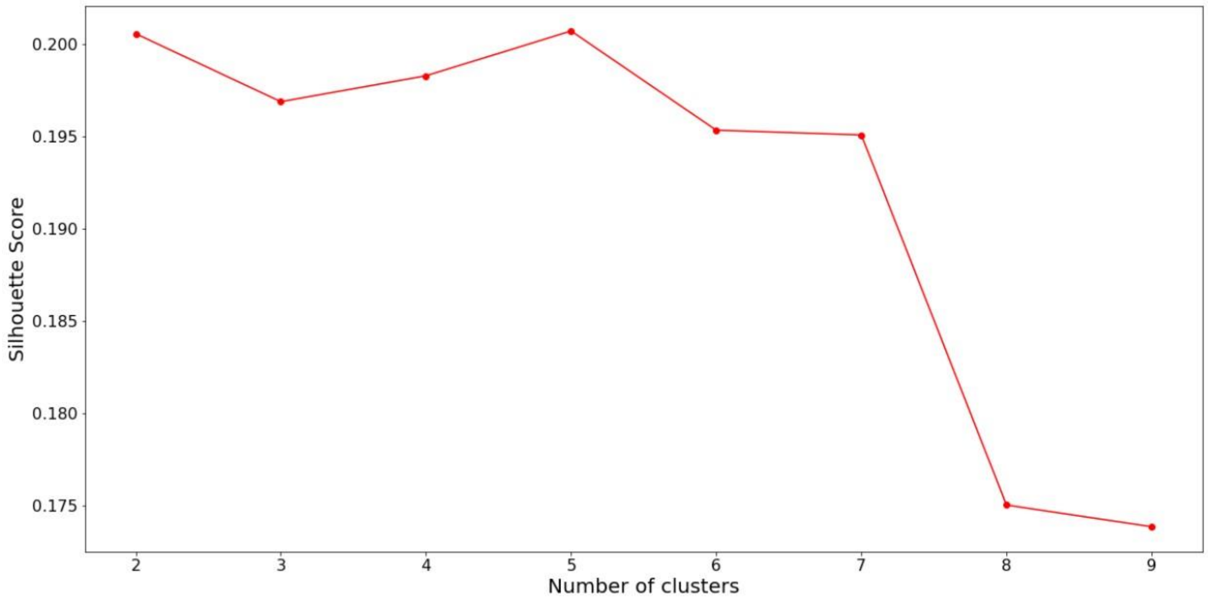

Figure 5.: *Silhouette score plot in order to determine number of clusters*

From the graph above we can read out the optimal number of clusters (where the Silhouette score is the highest) which is 5 in our case. The next step is to run the K-means clustering and get the following table as a result (Fig. 6):

Name	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Rogers Park	42.009574	-87.675550	0	Park	Beach	Pizza Place	Sandwich Place	Café	Fast Food Restaurant	African Restaurant	Mexican Restaurant	Coffee Shop	Supermarket
West Ridge	41.879788	-87.633113	2	Hotel	Theater	Steakhouse	Snack Place	Coffee Shop	New American Restaurant	Italian Restaurant	Gym	Park	Cuban Restaurant
Uptown	41.969450	-87.660513	0	Coffee Shop	Grocery Store	Vietnamese Restaurant	Mexican Restaurant	Breakfast Spot	Chinese Restaurant	Pizza Place	Sushi Restaurant	Vegetarian / Vegan Restaurant	Pet Store
Lincoln Square	40.148032	-89.363308	0	Pharmacy	Convenience Store	Pizza Place	Coffee Shop	Sandwich Place	Bar	Construction & Landscaping	Gym / Fitness Center	Discount Store	Donut Shop
North Center, North Park	41.858657	-87.612199	0	Aquarium	History Museum	Park	Planetarium	Pizza Place	Historic Site	Grocery Store	Burger Joint	Beach	Coffee Shop

Figure 6.: *Results of the K-Means clustering*

And now we can show the clusters we just created on the map (Fig. 7):

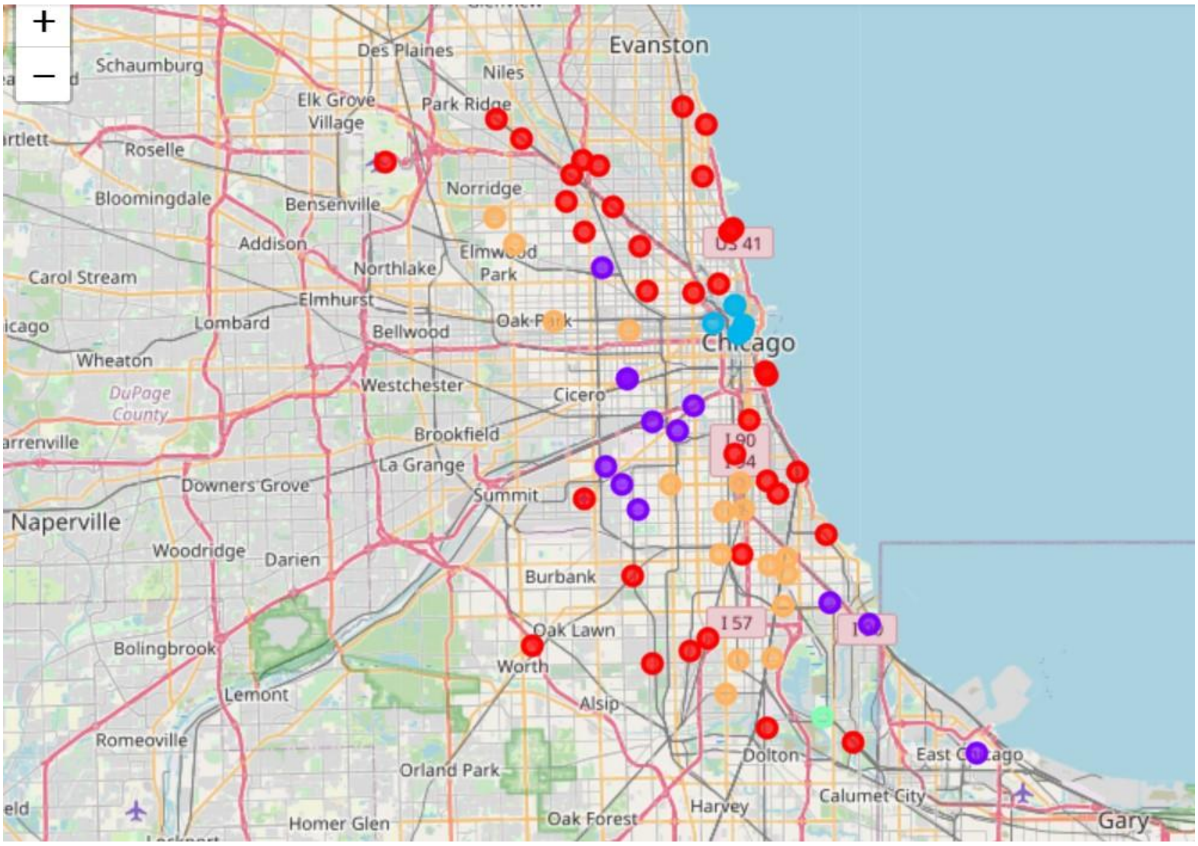

Figure 7.: *Visualization of clusters on Chicago map*

5.7. Some limitations of applied technics

Despite the fact that the algorithm used has successfully separated clusters, we need to consider some important limiting factors:

- We have collected venues only in 1500 m radius territory of district locations. In addition, the number of venues that could be collected was limited to 100 per query. Hence we could not take into account every venue category in a district.
- Because of the identical coordinates of some districts locations, we had to combine a few districts that could also distort the results.
- Arbitrary filtering of venue categories may also cause differences in results.

6. Results and Discussion

6.1. Interpretation clusters

Cluster 0

This is the largest cluster, including about half of Chicago's districts. It includes suburban and inner city districts also. This cluster shows the greatest diversity in different venue categories. What can be said in general is that the districts of this cluster are contain many cafés and bars, while here is the highest diversity in restaurant categories. In many districts there are a large number of tourist attraction institutions (Aquariums, Museums, Planetariums etc.) (Fig. 8).

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Rogers Park	Park	Beach	Pizza Place	Sandwich Place	Café	Fast Food Restaurant	African Restaurant	Mexican Restaurant	Coffee Shop	Supermarket
2	Uptown	Coffee Shop	Grocery Store	Vietnamese Restaurant	Mexican Restaurant	Breakfast Spot	Chinese Restaurant	Pizza Place	Sushi Restaurant	Vegetarian / Vegan Restaurant	Pet Store
3	Lincoln Square	Pharmacy	Convenience Store	Pizza Place	Coffee Shop	Sandwich Place	Bar	Construction & Landscaping	Gym / Fitness Center	Discount Store	Donut Shop
4	North Center, North Park	Aquarium	History Museum	Park	Planetarium	Pizza Place	Historic Site	Grocery Store	Burger Joint	Beach	Coffee Shop
6	Lincoln Park	Coffee Shop	Mexican Restaurant	Vegetarian / Vegan Restaurant	Spa	Sushi Restaurant	Bakery	Grocery Store	Gym	Gay Bar	Pub
8	Edison Park	Italian Restaurant	Salon / Barbershop	Sandwich Place	Breakfast Spot	Coffee Shop	Pizza Place	Bank	Mexican Restaurant	Bakery	American Restaurant
9	Norwood Park	Park	American Restaurant	Bar	Donut Shop	Coffee Shop	Italian Restaurant	Fast Food Restaurant	Polish Restaurant	Sandwich Place	Pizza Place
10	Jefferson Park	Bar	Pizza Place	Park	Convenience Store	Bakery	Chinese Restaurant	Ice Cream Shop	Pharmacy	Coffee Shop	Grocery Store
11	Forest Glen	Grocery Store	Sandwich Place	Chinese Restaurant	Bar	Park	Pizza Place	Train Station	Filipino Restaurant	Pharmacy	Donut Shop
12	Albany Park	Pizza Place	Park	Middle Eastern Restaurant	Hookah Bar	Sandwich Place	Ice Cream Shop	Supermarket	Mobile Phone Shop	Donut Shop	Coffee Shop
13	Portage Park	Bar	Pizza Place	Pharmacy	Video Store	Park	Coffee Shop	Sandwich Place	Italian Restaurant	Discount Store	Donut Shop
14	Irving Park	Sandwich Place	Discount Store	Mexican Restaurant	Italian Restaurant	Pizza Place	American Restaurant	Chinese Restaurant	Latin American Restaurant	Coffee Shop	Bar

***Figure 8**.: First part of the Cluster 0 data*

Cluster 1 ("The Empire of Mexican Gastronomy in Chicago")

The districts in this cluster could be called "The Empire of Mexican Gastronomy in Chicago". Since, the first and second most common venue category in all districts of the cluster is Mexican Restaurant.

Apart from a few parks, banks or pharmacies, predominantly discount commercial units are the most common locations in these districts. Although the districts are gastronomically popular, they offer the same cuisine. In addition, due to the very few sights, these districts are unlikely to be popular tourist destinations (Fig. 9).

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
18	Hermosa	Mexican Restaurant	Sandwich Place	Grocery Store	Discount Store	Fast Food Restaurant	Pharmacy	Park	Fried Chicken Joint	Donut Shop	Cuban Restaurant
26	North Lawndale	Mexican Restaurant	Pharmacy	Bank	Pizza Place	Seafood Restaurant	Food	Mobile Phone Shop	Liquor Store	Fast Food Restaurant	Nightclub
27	South Lawndale	Mexican Restaurant	Pharmacy	Bank	Pizza Place	Seafood Restaurant	Food	Mobile Phone Shop	Liquor Store	Fast Food Restaurant	Nightclub
33	Grand Boulevard	Sandwich Place	Discount Store	Park	Pharmacy	Sporting Goods Shop	Mexican Restaurant	Donut Shop	Gas Station	American Restaurant	Seafood Restaurant
43	Calumet Heights	Mexican Restaurant	Park	Fast Food Restaurant	Discount Store	Bank	Pharmacy	Sandwich Place	Shoe Store	Grocery Store	Currency Exchange
47	East Side	Harbor / Marina	Mexican Restaurant	Bar	Pizza Place	Park	Supermarket	Fast Food Restaurant	Seafood Restaurant	Shipping Store	Light Rail Station
52	Archer Heights	Mexican Restaurant	Pizza Place	Discount Store	Sandwich Place	Bank	Fast Food Restaurant	Video Store	Bar	Grocery Store	Bakery
53	Brighton Park	Mexican Restaurant	Fast Food Restaurant	Park	Sandwich Place	Donut Shop	Taco Place	Hot Dog Joint	Video Store	Ice Cream Shop	Supermarket
54	McKinley Park	Fast Food Restaurant	Mexican Restaurant	Park	Sandwich Place	Video Store	Donut Shop	Grocery Store	Taco Place	Pizza Place	Ice Cream Shop
55	Bridgeport	Mexican Restaurant	Grocery Store	Art Gallery	Diner	Soccer Field	Bakery	Fast Food Restaurant	Coffee Shop	Pet Store	Furniture / Home Store
57	West Elsdon	Mexican Restaurant	Pizza Place	Taco Place	Grocery Store	Bakery	Video Store	Bar	Discount Store	Gas Station	Bank
61	Chicago Lawn	Pizza Place	Mexican Restaurant	Fast Food Restaurant	Sandwich Place	Pharmacy	Discount Store	Taco Place	American Restaurant	Grocery Store	Bakery

Figure 9.: *Data of the Cluster 1*

Cluster 2 ("The City")

This cluster only contains six districts but they are in the heart of Chicago. The most common venues are Hotels, Steakhouses and Italian, and American Restaurants. There are also Theatres, Bars and Coffee shops in the districts. However, there is a shortage of fine cuisine restaurants, apart from seafood restaurants. All important factors can be found, which is a condition for opening a restaurant (Fig. 10).

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	West Ridge	Hotel	Theater	Steakhouse	Snack Place	Coffee Shop	New American Restaurant	Italian Restaurant	Gym	Park	Cuban Restaurant
5	Lake View	Hotel	Steakhouse	Bar	Theater	Mediterranean Restaurant	New American Restaurant	Italian Restaurant	Seafood Restaurant	Mexican Restaurant	Park
7	Near North Side	Steakhouse	Hotel	Italian Restaurant	Pizza Place	Bar	New American Restaurant	Gym / Fitness Center	Mexican Restaurant	Gym	Coffee Shop
25	Near West Side, Near South Side, Lower West Side	Steakhouse	Hotel	Italian Restaurant	Pizza Place	Bar	New American Restaurant	Gym / Fitness Center	Mexican Restaurant	Gym	Coffee Shop
28	Loop	Hotel	Steakhouse	Theater	Park	Snack Place	Bar	Italian Restaurant	Coffee Shop	Donut Shop	Museum
69	Morgan Park	Italian Restaurant	New American Restaurant	Pizza Place	Coffee Shop	Restaurant	Hotel	Burger Joint	Grocery Store	Café	Yoga Studio

Figure 10.: *Data of the Cluster 2*

Cluster 3

This cluster only contains one district South Deering. The explanation of this is that South Deering is mainly an industrial district with very few population and big area of harbor with industrial tracks to commercial port. The only things are related to gastronomy is the relative common Greek and Eastern European Restaurants (Fig. 11).

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
46	South Deering	Canal Lock	Harbor / Marina	Greek Restaurant	Nature Preserve	River	Dry Cleaner	Duty-free Shop	Eastern European Restaurant	Electronics Store	Elementary School

***Figure 11**.: Data of the Cluster 3*

Cluster 4 ("Citadel of fast foods")

The districts in this cluster deserve to be marked by the " Citadel of fast foods " marker, since most common venue categories are related to fast foods. Beside fast foods some Chinese, American or Mexican Restaurant are also can be found in these districts. Train stations are also among the most common venues which is important from our perspective. Different candy shops (Donut store, Ice cream shop) and grocery stores are also frequently occurring (Fig. 12).

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
15	Dunning	Fast Food Restaurant	Clothing Store	Pizza Place	Ice Cream Shop	Deli / Bodega	Hot Dog Joint	Sandwich Place	Italian Restaurant	Bakery	Sports Bar
16	Montclare	Fast Food Restaurant	Grocery Store	Pizza Place	Bakery	Donut Shop	Discount Store	Italian Restaurant	Park	Mexican Restaurant	Ice Cream Shop
22	West Town, West Englewood	Fast Food Restaurant	Gas Station	Seafood Restaurant	Park	Intersection	Light Rail Station	Fried Chicken Joint	Supermarket	Mexican Restaurant	Bookstore
23	Austin	Golf Course	Fast Food Restaurant	Gym	Grocery Store	Southern / Soul Food Restaurant	Seafood Restaurant	Convenience Store	Cosmetics Shop	Hobby Shop	Sandwich Place
24	West Garfield Park, East Garfiled Park	Train Station	Fast Food Restaurant	Park	Food	Café	Botanical Garden	Discount Store	Fried Chicken Joint	Sandwich Place	Pet Service
31	Oakland	Bar	Pharmacy	Fast Food Restaurant	Bank	Sandwich Place	Cosmetics Shop	Pizza Place	Diner	Discount Store	Convenience Store
37	Woodlawn	Fried Chicken Joint	Lounge	Fast Food Restaurant	Sandwich Place	Discount Store	Bank	Grocery Store	Chinese Restaurant	Supermarket	Pizza Place
39	Chatham	Lounge	Sandwich Place	Fried Chicken Joint	Fast Food Restaurant	Bar	Donut Shop	BBQ Joint	Discount Store	Chinese Restaurant	American Restaurant
40	Avalon Park	Fast Food Restaurant	Sandwich Place	Discount Store	Pharmacy	Chinese Restaurant	Lounge	Video Store	Pizza Place	Diner	Donut Shop
42	Burnside	Fast Food Restaurant	Fried Chicken Joint	Rental Car Location	Shoe Store	Liquor Store	Athletics & Sports	Optical Shop	Gas Station	Discount Store	Nightclub
44	Roseland	Fast Food Restaurant	Sandwich Place	Intersection	Donut Shop	Grocery Store	Liquor Store	Breakfast Spot	Light Rail Station	Baseball Field	Fried Chicken Joint
				Fried Chicken							Mexican

***Figure 12**.: Data of Cluster 4*

7. Conclusions and Recommendations

Based on the evaluation of the results, the following conclusions and recommendations can be made:

1. The most suitable districts for opening a new fine cuisine restaurant are in the city center (Cluster 2). There are a lot of hotels that indicate the presence of many tourists. The relatively large number of cafés and bars suggests a lively nightlife. The relative lack of fine cuisine-type restaurants reduces the disadvantage of competitive pressure.
2. Another group of districts suitable for opening the restaurant is located in Cluster 0. These districts are made attractive locations by their diversity. However, not all the districts prove to be appropriate here. Districts with lots of tourist attractions and/or lots of cafes or bars, but no vegetarian restaurants, may be suitable candidates.
3. The districts of Clusters 1, 2 and 3 are unlikely to be suitable for opening a new fine-cuisine restaurant. Although the monotonous gastronomic venues in these districts would be influenced by a different type of restaurant, but the lack of tourist attraction sites calls into question the feasibility of such a project.

8. References

The Jupyter notebook of the analysis can be found on the GitHub: https://github.com/chegeo/Coursera_Capstone/blob/main/IBM_Capstone_Project_Final.ipynb