Battle of the Neighbourhoods

Coursera Capstone, Final Assignment

Report by Lybomir Ahtapodov

1. Introduction: Business Problem. Sports are becoming an increasingly popular part of people's everyday life both as a means of improved wellbeing and personal health. It is only natural to expect that this increased popularity of regular training will lead to increased demand for sporting goods. Yet in big cities with a hectic everyday life such as Chicago, it is of crucial importance for commercial venues such as sporting goods shops to be visible and conveniently located in order to maximise their customer base. The present project explores the problem of finding the best location for opening a sporting goods shop in Chicago, based on a data-scientific approach. The outcome of this project will potentially be of significant interest for sporting goods retail businesses looking to either enter the market or expand their activities in the city of Chicago.

2. Data. The data used in this project has been obtained from two sources. Firstly, the precise geographical coordinates of multiple points along the boundary of each community area will be obtained from the City of Chicago's official website.[1] With some processing that will be detailed in the following section, this data was used to obtain the centre point of each community area, as well as determine the radius of a circular vicinity that best matches the borders of the community area. After that, venue information for the area around each community area centre point within a vicinity of radius as determined above will be requested from Foursquare.[2] Following this step, a venue shortlist containing venue types that tend to occur together with sporting goods shops was created, which was then used to perform cluster analysis of all Chicago community areas.

3. Methodology. In this section, I will explain the exploratory data analysis and modelling that was performed for the purposes of the present project.

3.1. Exploratory Data Analysis. As a first step, the coordinates for the boundary points of each community area were retrieved as a list and then converted into a numpy array. By averaging the latitude and longitude of all boundary points available for a given community area, the centre point of the community area was determined. An alternative approach would have been to inquire the centre point separately. After the coordinates of the centre point have been determined, the next step is to determine the most appropriate value for the radius of a circular vicinity to be used for inquiring venue types from Foursquare. This was done by calculating the distance from the centre point to each available boundary point, and then adopting the average of all distances as the radius of question. As community areas have diverse geographical shapes, departing considerably from a circle, there are inevitably some discrepancies between the circular vicinities arrived at in the above described fashion and the actual community areas, but I am confident the chosen method provides the best overall match. Fig. 1 shows an example for the North Park community area

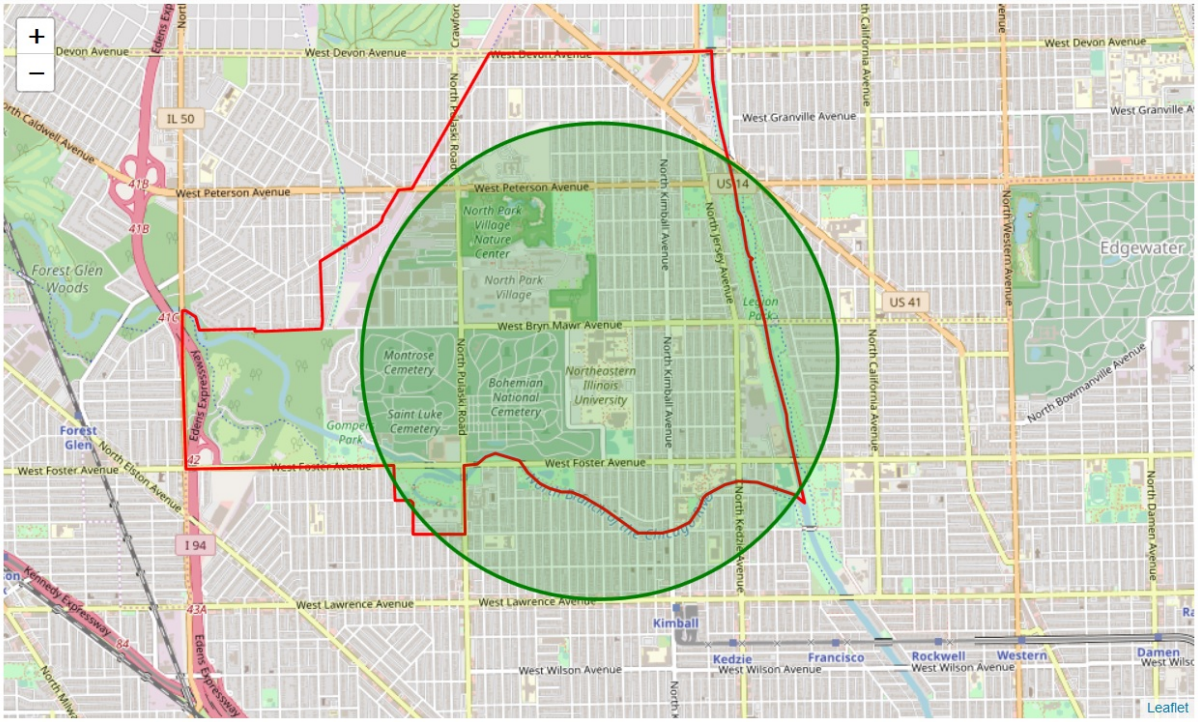

Fig. 1. North Park community area with the geographical boundary shown in red and the computed circular vicinity shown in green

The data obtained from the City of Chicago's official site did not require any cleaning, however it is worth mentioning that the order of the latitude and longitude coordinates is reversed in relation to the default in the folium and geopy packages.

The result of this first section of the exploratory data analysis is the following dataframe:

	CA Name	Latitude	Longitude	Vicinity Radius
0	DOUGLAS	41.836210	-87.616148	1148.147267
1	OAKLAND	41.822268	-87.600263	612.827418
2	FULLER PARK	41.809542	-87.632251	1012.599136
3	GRAND BOULEVARD	41.812351	-87.619113	1200.135664
4	KENWOOD	41.809096	-87.590145	730.719431
5	LINCOLN SQUARE	41.974304	-87.694233	1308.131415
6	WASHINGTON PARK	41.791200	-87.617740	1160.867231
7	HYDE PARK	41.794078	-87.584535	956.882038
8	WOODLAWN	41.777957	-87.584390	1364.614791

Fig. 2. Screenshot of the head of the dataframe resulting from the first part of the exploratory data analysis

3.2. Foursquare Venue Information. The next step is to obtain venue information from Foursquare for each of the 77 community areas using the centre points and the vicinity radii computed in 3.1. At this point it should be noted that it cannot be expected that the occurrence of all venues returned by Foursquare have a relation to the occurrence of a sporting goods shop, which is our venue type of interest. To tackle this, all venues were added to a dataframe containing the community area, the coordinates of its centre point, and the venue coordinates, name and type. This dataframe was further one-hot encoded by venue type (or category, which is equivalent), and then only the parent community area name was retained. The dataframe was then grouped by community area name and averaged, which resulted in the following:

	CA Name	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar
0	ALBANY PARK	0.000000	0.000000	0.000000	0.012658	0.000000	0.000000	0.00	0.000000	0.012658	...	0.000000	0.000000	0.0	0.000000
1	ARCHER HEIGHTS	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000
2	ARMOUR SQUARE	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.014493	...	0.000000	0.000000	0.0	0.000000
3	ASHBURN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.033333	...	0.000000	0.016667	0.0	0.000000
4	AUBURN GRESHAM	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	...	0.000000	0.000000	0.0	0.014085
5	AUSTIN	0.030000	0.000000	0.000000	0.000000	0.010000	0.000000	0.00	0.000000	0.020000	...	0.000000	0.000000	0.0	0.000000
6	AVALON PARK	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000
7	AVONDALE	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	...	0.000000	0.000000	0.0	0.010000

Fig. 3. Dataframe containing average occurrence rate of all retrieved venue types per community area

At this point, pairwise correlation was performed between the column “Sporting Goods Shop” and each of the other columns, and only venue types whose occurrence at least weakly correlates with that of a sporting goods shop were shortlisted based on the obtained correlation coefficients. The threshold value was selected to be 0.3.

	Venue Type	Correlation Coeff
128	Food Court	0.712272
131	Football Stadium	0.658410
148	Golf Course	0.632037
35	Big Box Store	0.428943
163	History Museum	0.425608
0	ATM	0.422418
294	Snack Place	0.410720
59	Cafeteria	0.400934
32	Bed & Breakfast	0.400934
72	Circus	0.400934

Fig. 4. Results of the pairwise venue type correlation analysis

3.3. K-Means Clustering. The next step is to perform clustering of all Chicago community areas based on the selected venue types whose occurrence correlates with that of a sporting goods shop, the latter included. The machine learning algorithm of K-Means Clustering with k=4 was chosen for this purpose. Fig. 5 shows results of the cluster analysis in a choropleth map and Fig. 6 presents a table with the number of community areas and sporting goods shops for each cluster.

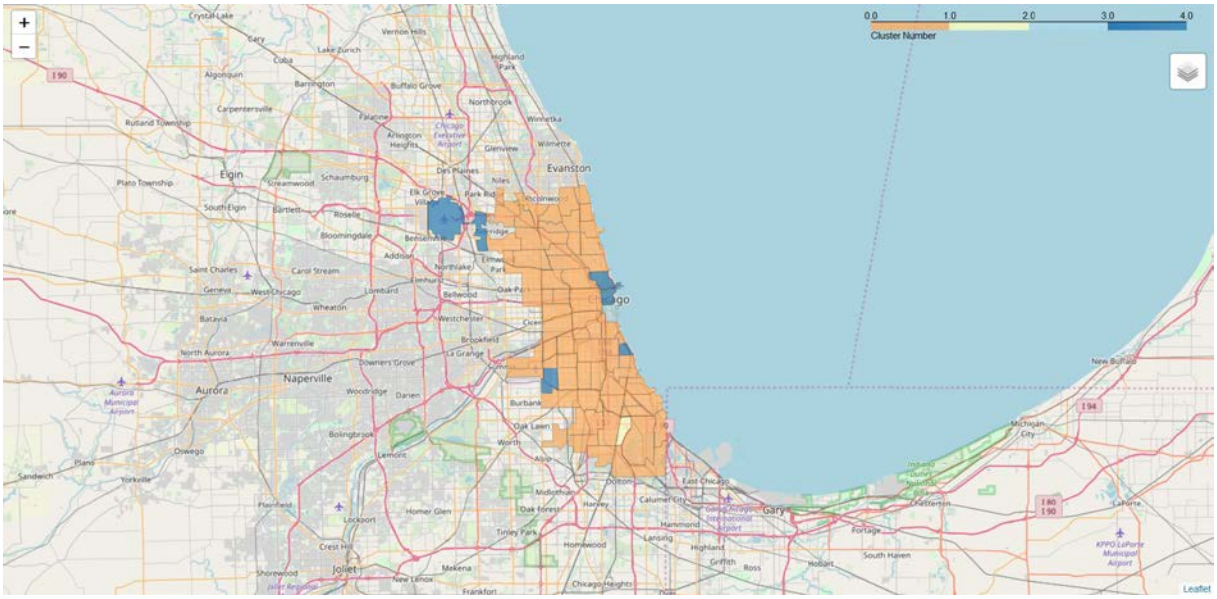

Fig. 5. Choropleth map of the cluster analysis of the Chicago community areas

	Cluster Labels	Sporting Goods Shop	CA Name
0	0	9	70
1	1	1	1
2	2	1	1
3	3	3	5

Fig. 6. Number of community areas and sporting goods shops per cluster

4. Results and Discussion. Each cluster was scored using the following formula:

$$\boldsymbol{Score} = \sum_{\boldsymbol{venues}} \boldsymbol{P_{avg}}(\boldsymbol{venue}).\boldsymbol{C_{corr}}(\boldsymbol{venue}),$$

where P_{avg} is the average occurrence rate of a particular venue type and C_{corr} is the computed pairwise correlation coefficient in relation to a sporting goods shop. The sporting goods shop venue type itself enters with a weight of 1.0. This is equivalent to stating that the occurrence of a venue type fully correlates with itself.

The analysis yields one large cluster, Cluster 0 containing 70 community areas and a total of 9 sporting goods shops, yielding 0.13 sporting goods shops per community area, which scores lowest, with a score of 0.0817. The other three clusters have an identical score of 0.2667, and consist of:

- Cluster 1: 1 community area with 1 sporting goods shop
- Cluster 2: 1 community area with 1 sporting goods shop
- Cluster 3: 5 community areas with 3 sporting goods shops

Thus, according to the introduced scoring system, Clusters 1, 2 and 3 should all be equally favourable locations for a sporting goods shop. The community areas corresponding to those clusters are: Kenwood, Loop, Near North Side, Near South Side, O'Hare, Pullman and West Lawn.

It should be pointed out that while being part of the most favourable locations according to the carried out analysis, following community areas do not have a sporting goods shop yet Kenwood, Loop, West Lawn.

5. Conclusion. In conclusion, the above analysis points at seven community areas as equally favourable locations for a sporting goods shop in Chicago: Kenwood, Loop, Near North Side, Near South Side, O'Hare, Pullman and West Lawn. Of these three do not yet have such a venue, which suggests that there might be significant potential for opening a sporting goods shop in one of them: Kenwood, Loop and West Lawn.

6. References

[1] City of Chicago's Official website:
https://www.chicago.gov/city/en/depts/doit/dataset/boundaries_-_communityareas.html

[2] Foursquare: https://foursquare.com