Applied Data Science Capstone Project

Presentation

What's the best location to set up a high-end coffee shop in Chicago?

Business objectives

- When it comes to set up a new business, like a high-end coffee shop, in a city like Chicago, choosing the right location can be overwhelming. There are so many factors to take into account, like competition. Depending on who you want to cater to, you might want to set up your new venture in a touristic location, near a university or college, or nearby work or goverment places.
- This project proposes to identify the best location / neighborhood to set up a new coffee shop in Chicago. This will be done calling the Foursquare api to retrieve existing venues in all Chicago neighborhoods, and then by applying a scoring model to make the recommendations.

Data

We will be using different sources of data to deliver the project:

- Chicago list of neighborhoods: Data will be collected by scraping the Wikipedia page *https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago*
- OpenStreetMap Data: We'll use Nominatim, the search engine for OpenStreetMap data, to retrieve the GPS coordinates for the list of Chicago neighborhoods
- Foursquare data : We'll query the Foursquare APIs to extract venues of interest (ex: Coffee Shops, Offices, Universities…) and their coordinates.

Methodology

- Extracting all necessary data, and putting it in a dataframe
- First, we need to retrieve GPS coordinates for all of Chicago neighborhoods

	Neighborhood	Community area	Latitude	Longitude
0	Albany Park	Albany Park	41.971937	-87.716174
1	Altgeld Gardens	Riverdale	41.654864	-87.600439
2	Andersonville	Edgewater	41.977139	-87.669273
3	Archer Heights	Archer Heights	41.811422	-87.726165
4	Armour Square	Armour Square	41.840033	-87.633107

- Then, we append the counts of select venue categories

	Neighborhood	Community area	Latitude	Longitude	Universities	Offices	Bakeries	Coffee Shops
0	Albany Park	Albany Park	41.971937	-87.716174	64.0	1.0	15.0	6.0
1	Altgeld Gardens	Riverdale	41.654864	-87.600439	1.0	1.0	NaN	NaN
2	Andersonville	Edgewater	41.977139	-87.669273	15.0	6.0	25.0	12.0
3	Archer Heights	Archer Heights	41.811422	-87.726165	2.0	1.0	7.0	1.0
4	Armour Square	Armour Square	41.840033	-87.633107	54.0	6.0	7.0	5.0
5	Ashburn	Ashburn	41.747533	-87.711163	1.0	4.0	3.0	NaN
7	Auburn Gresham	Auburn Gresham	41.750474	-87.664304	NaN	4.0	4.0	1.0
8	Avalon Park	Avalon Park	41.745035	-87.588658	NaN	1.0	1.0	1.0

Methodology

- The counts of select venue categories are used to compute a score per neighborhood, where we assign a weight to venue categories

Define weight of each venue category in score determination

```
weight_offices = 3 # Having offices around is good
```

```
weight_universities = 2 # Having universities around is good
```

```
weight_bakeries = -1 # Having bakeries around is bad (competition)
```

```
weight_coffee_shops = -3 # Having coffee shops around is bad (competition)
```

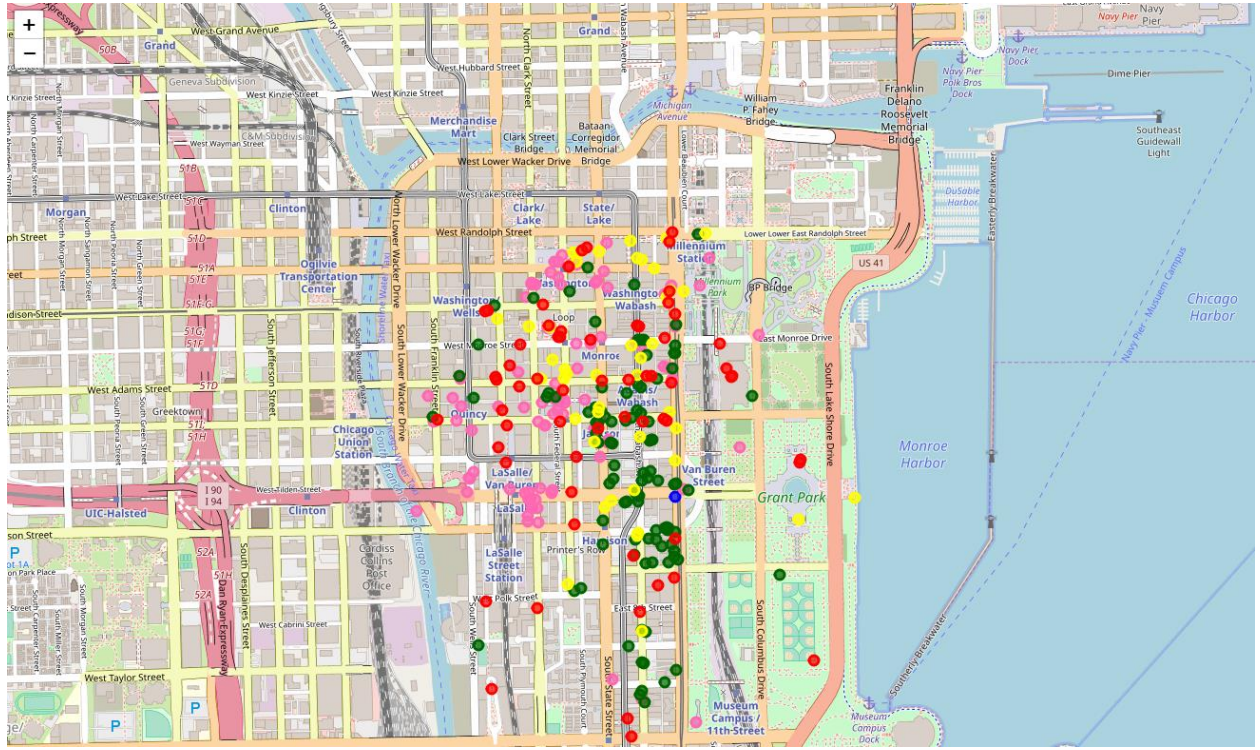
- Each neighborhood is assigned a score. Here are the 5 neighborhoods with the highest scores

```
chicago_score.head(5) # Top 5 recommendations
```

	Neighborhood	Score
96	The Island	235.0
173	Printer's Row	204.0
212	Tri-Taylor	193.0
214	Union Ridge	191.0
215	University Village	184.0

Results

- The Island is the clear winner
- Here’s how the different existing venues are scattered over the neighborhood



- Government Offices
- Universities
- Dessert shops
- Coffee shops

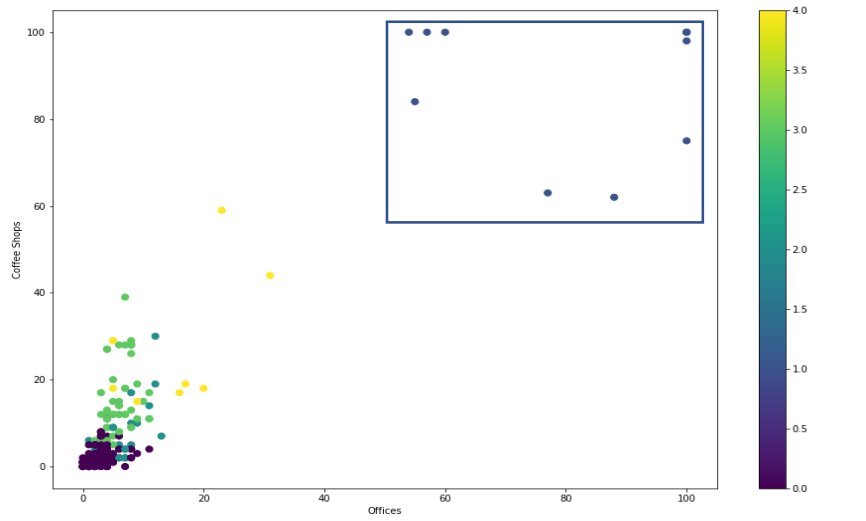
Additional analysis

- We ran a clustering algorithm (K-Means) to find neighborhoods that look similar to our winning neighborhood

	Cluster Labels	Neighborhood	Community area	Latitude	Longitude	Universities	Offices	Bakeries	Coffee Shops
0	2	Albany Park	Albany Park	41.971937	-87.716174	64.0	1.0	15.0	6.0
1	0	Altgeld Gardens	Riverdale	41.654864	-87.600439	1.0	1.0	0.0	0.0
2	3	Andersonville	Edgewater	41.977139	-87.669273	15.0	6.0	25.0	12.0
3	0	Archer Heights	Archer Heights	41.811422	-87.726165	2.0	1.0	7.0	1.0
4	2	Armour Square	Armour Square	41.840033	-87.633107	54.0	6.0	7.0	5.0

- The neighborhoods of interest all belong to cluster 1, where counts of coffee shops and of offices are > 60

```
fig.show()
```



[83]:

	Cluster Labels	Neighborhood	Community area	Latitude	Longitude	Universities	Offices	Bakeries	Coffee Shops
71	1	The Gap	Douglas	41.892357	-87.623588	100.0	60.0	80.0	100.0
96	1	The Island	Austin	41.875562	-87.624421	100.0	88.0	43.0	62.0
108	1	Lake View	Lake View	41.885382	-87.627908	100.0	100.0	46.0	100.0
121	1	The Loop	The Loop	41.881609	-87.629457	100.0	100.0	55.0	100.0
124	1	Magnificent Mile	Near North Side	41.894523	-87.624228	100.0	54.0	65.0	100.0
162	1	Park West	Lincoln Park	41.882557	-87.622500	100.0	100.0	60.0	98.0
173	1	Printer's Row	The Loop	41.873787	-87.628900	100.0	77.0	38.0	63.0
180	1	River North	Near North Side	41.888341	-87.617903	86.0	55.0	62.0	84.0
210	1	Streeterville	Near North Side	41.893365	-87.621997	88.0	57.0	68.0	100.0
214	1	Union Ridge	Norwood Park	41.878295	-87.638949	84.0	100.0	52.0	75.0
233	1	West Loop	Near West Side	41.881609	-87.629457	100.0	100.0	55.0	100.0

Conclusion

- We’ve demonstrated here how to build a low-complexity, highly-interpretable recommendation engine
- As a following step, we could impove the process by adding more variables to help the decision (other venue categories, demographics,…)