Capstone Project- (Week 2)
Applied Data Science Capstone by IBM/Coursera

Table of Contents

1. Business Problem

Chicago is a very diverse city and one of the largest in the US. When it comes to food, the city has a lot to offer. For instance, every year in July, the city hosts a food festival called “Taste of Chicago”. It is held for five days in Grant Park, one of the biggest park downtown. The event is the largest festival in Chicago. During this event, one can experience some local specialties as well as some ethnic food.

However, during the remainder of the year, one might be wondering where to find the best ethnic restaurant in city. The city has about 200 neighborhoods and about 77 community areas. With a population of more than 2.7 million, it may not be easy to decide where to open a new restaurant if one wanted to.

In this project, I will build a map of Chicago showing the population density per area, then I will explore the different type of restaurants available and finally see where one could open a new “ethnic” restaurant.

2. Data

After a few hours of research, I was able to find the following; it is a link of some US city with their local areas, geolocations and population density.

https://geo.nyu.edu/catalog/stanford-xq082nw3443

I will exploit this file to extract the relevant data pertaining to Chicago.

We will also use the Foursquare API, with the credentials we created during the lab session on foursquare to explore the different areas and cluster them.

Finally, we may use geocoder if needed as we move along with the project.

After a quick look at the sample data, here are the relevant data we will use:

- The State (we will later filter IL as this analysis is for Chicago)
- The City, here we will analyze "Chicago"
- The neighborhood, which can be found in the property "name"
- The population density, in the "popdensity" property.

We will extract these properties as well as the latitude and longitude of each neighborhood to construct the data frame to be used.

3. Methodology

I will use GitHub to store the final work.

Here are the steps we will take:

- We will download the data, analyze it to extract the relevant information.
- Next we will construct a data frame with all cities represented in the raw data set, then filter on Chicago as it is our city of interest.
- Use the Foursquare API to explore venues in the city.
- Make a chloropleth map of the city to visualize the denser neighborhoods
- Explore venues in theses dense neighborhoods and see if there are many “exotic” restaurants.
- From there, we will be able to conclude.

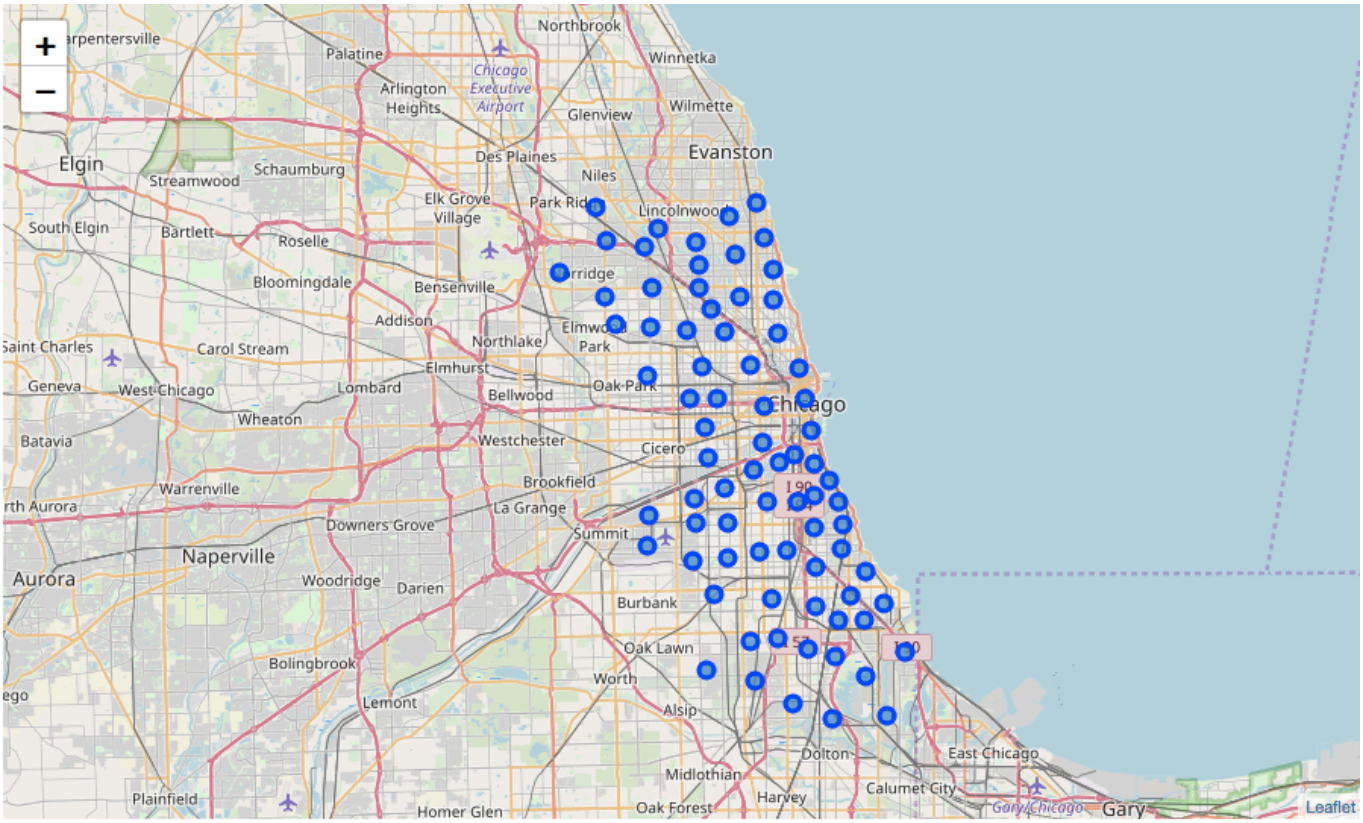
Sample from the data set

[27]:

	State	City	Area	PopDensity	Latitude	Longitude
0	CA	Long Beach	Airport Area	5741.323529	33.8167	-118.154496
1	CA	Long Beach	Alamitos Heights	7060.266667	33.7738	-118.125871
2	CA	Long Beach	Belmont Heights	15536.411111	33.7639	-118.151191
3	CA	Long Beach	Belmont Shore	13146.320000	33.7589	-118.137396
4	CA	Long Beach	Bixby Area	9901.688000	33.8405	-118.176421

I used the visualization tools of python to visualize geographic details of Chicago and its neighborhoods. The map of the City then had the community areas superposed on top. I used latitude and longitude values to get the visual as below:

Map of Chicago with its community areas



quick look at the venue in the loop area:

	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Loop	41.8784	-87.625333	The Art Institute of Chicago	41.879579	-87.623909	Art Museum
1	Loop	41.8784	-87.625333	Symphony Center (Chicago Symphony Orchestra)	41.879275	-87.624680	Concert Hall
2	Loop	41.8784	-87.625333	Chicago Architecture Foundation	41.878556	-87.624550	Museum
3	Loop	41.8784	-87.625333	Auditorium Theatre	41.876058	-87.625303	Theater
4	Loop	41.8784	-87.625333	Thorne Miniature Rooms	41.879532	-87.623680	Museum

The Loop area in like the business center of the town. With little residential home. This may explain why the first few venues are in the category of museum and theater.

4. Analysis

At this point, I was ready to start utilizing the Foursquare API to explore the neighborhoods and segment them.

I designed the limit as 100 venues and the radius 1500 meter for each area from their given latitude and longitude information.
Here is a list of the five first values of the venues name, category, latitude and longitude information from Foursquare API.

	name	categories	lat	lng
0	The Art Institute of Chicago	Art Museum	41.879579	-87.623909
1	Symphony Center (Chicago Symphony Orchestra)	Concert Hall	41.879275	-87.624680
2	Chicago Architecture Foundation	Museum	41.878556	-87.624550
3	Auditorium Theatre	Theater	41.876058	-87.625303
4	Thorne Miniature Rooms	Museum	41.879532	-87.623680

As expected for a large city like Chicago, too many venues were returned, 5305 venues.

I then created a data frame with the top 10 most common venues in each area. Here is how the data looks:

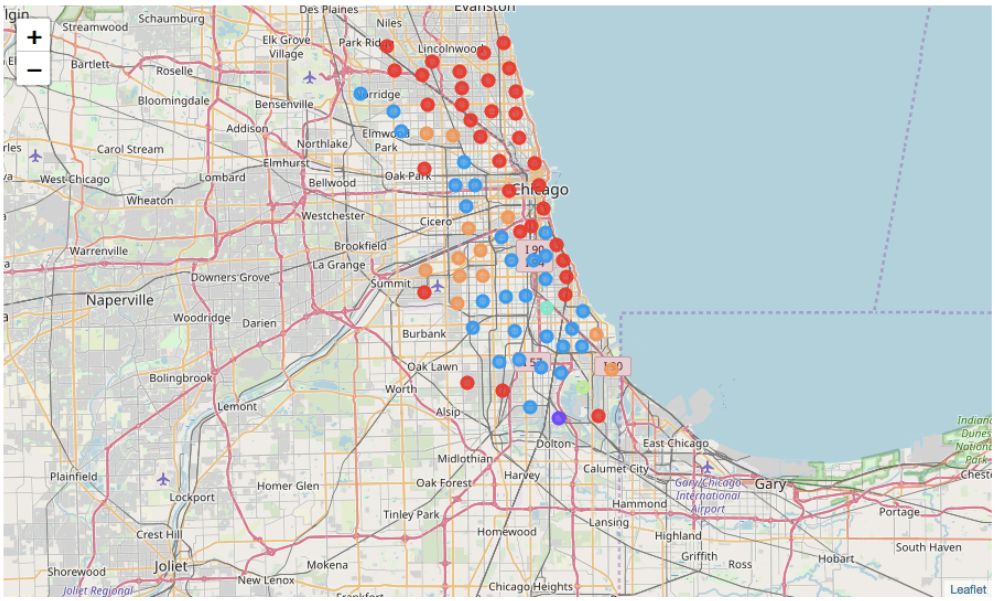
Top 10 venues per area

	Area	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	Coffee Shop	Chinese Restaurant	Pizza Place	Asian Restaurant	Korean Restaurant	Supermarket	Park	Mexican Restaurant	Mobile Phone Shop	Pub
1	Archer Heights	Mexican Restaurant	Mobile Phone Shop	Bar	Sandwich Place	Discount Store	Fast Food Restaurant	Bank	Coffee Shop	Cosmetics Shop	Pizza Place
2	Armour Square	Chinese Restaurant	Bar	Pizza Place	Bakery	Park	Mexican Restaurant	Korean Restaurant	Coffee Shop	Hot Dog Joint	Sandwich Place
3	Ashburn	Park	Fast Food Restaurant	Fried Chicken Joint	Mexican Restaurant	Seafood Restaurant	Pizza Place	Discount Store	Intersection	Pharmacy	Liquor Store
4	Auburn Gresham	Discount Store	Seafood Restaurant	Lounge	Fast Food Restaurant	Pharmacy	Sandwich Place	Southern / Soul Food Restaurant	Nightclub	Bar	Train Station

I will run K-Means to cluster the boroughs into 6 clusters. Below are the first few rows of the table data with clusters

	State	City	Area	PopDensity	Latitude	Longitude	Cluster label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
730	IL	Chicago	Loop	17486.450000	41.8784	-87.625333	0	Hotel	Theater	Park	Coffee Shop
738	IL	Chicago	Chatham	12681.364151	41.7385	-87.616624	2	Fast Food Restaurant	Sandwich Place	Women's Store	Discount Store
739	IL	Chicago	North Center	17814.894231	41.9473	-87.684523	0	Bar	Pizza Place	Mexican Restaurant	Coffee Shop
740	IL	Chicago	O'hare	6591.975000	41.9633	-87.847436	2	Sandwich Place	Donut Shop	Grocery Store	Convenience Store
741	IL	Chicago	Washington Park	13106.814286	41.7916	-87.617580	2	Fast Food Restaurant	Coffee Shop	Train Station	Sandwich Place

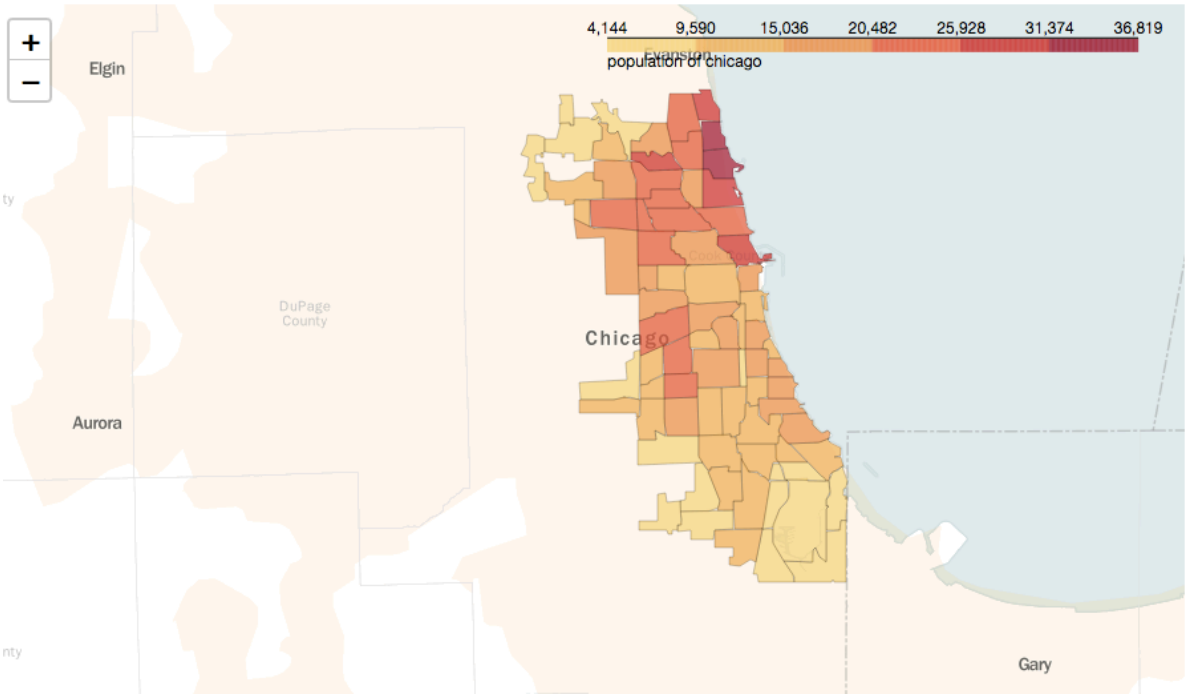
Here is a map of the city with different clusters:



At this point, we have too many venues and it is not obvious what the relevant information should be.

We will now look at the population density, then select the dense areas and perform the venue exploration there.

Here is a choropleth map of the city show the density of the population:



As we can see, some areas are denser than others. Since we are looking at the data for a restaurant business, it makes sense for us to consider only areas with a certain threshold of density.

5. Results and discussion

We will arbitrary take 25,000 as our threshold.

	State	City	Area	PopDensity	Latitude	Longitude
760	IL	Chicago	Uptown	34483.802381	41.9659	-87.654329
766	IL	Chicago	Edgewater	36499.086047	41.9869	-87.662182
785	IL	Chicago	Albany Park	27295.919149	41.9682	-87.721486
788	IL	Chicago	Rogers Park	30763.919048	42.0101	-87.669898
792	IL	Chicago	Logan Square	25074.923656	41.9237	-87.698424

```
df_chi_new.shape
```

```
(7, 6)
```

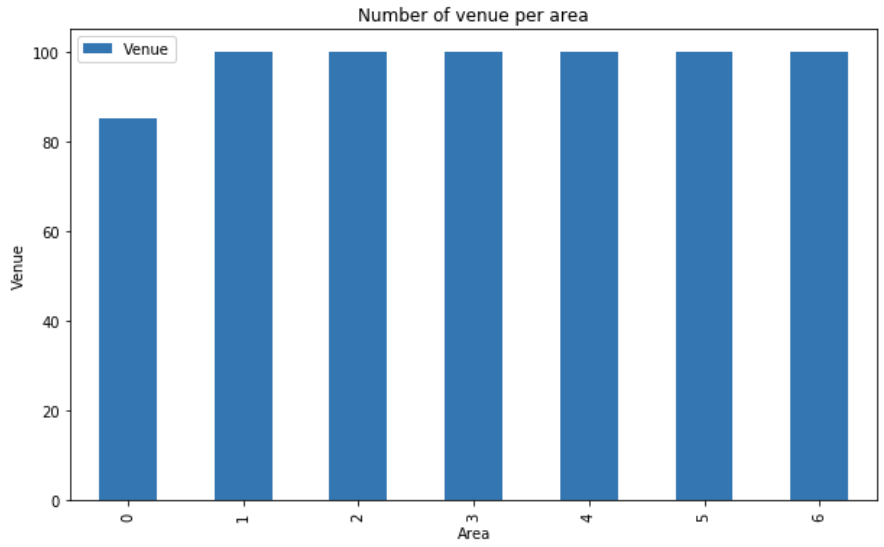
We have now trimmed the selection to just seven areas with the highest density. We can now explore these locations.

Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Albany Park	85	85	85	85	85	85
Edgewater	100	100	100	100	100	100
Lake View	100	100	100	100	100	100
Logan Square	100	100	100	100	100	100
Near North Side	100	100	100	100	100	100

A quick look at the most common venues in each area shows that restaurants category dominates. This is expected in high density areas. We also see that these areas have almost all 100 venues in their vicinity.

	Area	Venue
0	Albany Park	85
1	Archer Heights	65
2	Armour Square	100
3	Ashburn	58
4	Auburn Gresham	59

Here is a bar chart, showing the number of venues in the 7 areas.



In each of the areas, the top 10 venues do include restaurants, most of which are Asian restaurants (Chinese, Thai, Japanese)

```
-----Albany Park----
                venue  freq
0         Coffee Shop  0.05
1  Chinese Restaurant  0.05
2         Pizza Place  0.05
3                Park  0.04
4   Korean Restaurant  0.04
5  Mexican Restaurant  0.04
6    Asian Restaurant  0.04
7         Supermarket  0.04
8     Thai Restaurant  0.02
9      Shipping Store  0.02


-----Edgewater-----
                venue  freq
0         Coffee Shop  0.04
1    Asian Restaurant  0.04
2               Café  0.04
3                Bar  0.03
4  Mexican Restaurant  0.03
5        Antique Shop  0.03
6      Sandwich Place  0.03
7  Italian Restaurant  0.03
8     Thai Restaurant  0.03
9        Burger Joint  0.03


-----Lake View-----
                venue  freq
0          Pizza Place  0.05
1       Sandwich Place  0.04
2   Italian Restaurant  0.04
3                  Gym  0.04
```

```
0      American Restaurant  0.08
1                    Hotel  0.08
2              Pizza Place  0.04
3               Steakhouse  0.04
4       Italian Restaurant  0.04
5                      Bar  0.03
6       Salon / Barbershop  0.03
7            Grocery Store  0.03
8  New American Restaurant  0.03
9                      Gym  0.03


-----Rogers Park-----
                  venue  freq
0                 Beach  0.08
1        Sandwich Place  0.06
2                   Bar  0.04
3           Coffee Shop  0.04
4           Pizza Place  0.04
5                  Park  0.04
6                 Café  0.04
7  Fast Food Restaurant  0.03
8               Bakery  0.03
9                 Bank  0.03


-----Uptown-----
                          venue  freq
0          Vietnamese Restaurant  0.09
1                    Coffee Shop  0.06
2             Chinese Restaurant  0.04
3               Sushi Restaurant  0.04
4                Thai Restaurant  0.04
5                    Pizza Place  0.03
6                Breakfast Spot  0.03
7                 Grocery Store  0.03
8  Vegetarian / Vegan Restaurant  0.03
9                           Park  0.03
```

Chicago is the third largest city in the US, divided into 77 community areas. We have seen that seven of those areas have a density of 25000 or more, and all packed with hundreds of venues. Clustering and classify these venues can prove challenging.

In this analysis, the K-Mean method was used for the clustering. I chose to cluster the city using k=6. For more accuracy, we could have used a function to determine a better value for k. However, this will not add much since we are trying to spot the best neighborhood far a business based on the population.

I added the visualization and clustering information on a map. It can be improved by adding additional parameters and looking at the correlation between restaurants and other types of venues where people usually go before or after restaurants.

6. Conclusion

In this study, we had the opportunity Chicago remains a city of choice if one wants to set up a new ethnic restaurant. Seven of the areas are most suited, besides being densely populated, they almost all include more that 100 venues each, which is a good indicator of trending area.

There is room for additional restaurants with an exotic flavor. **I would recommend Rogers Park,** since this area is a dense neighborhood and has only one fast food restaurant.

Setting up a new restaurant in this area is likely to yield significant traffic.