Categorizing Chicago Communities using Crime Data and Foursquare

Luis Alberto Reyes
May 3, 2020

Introduction / Business Problem

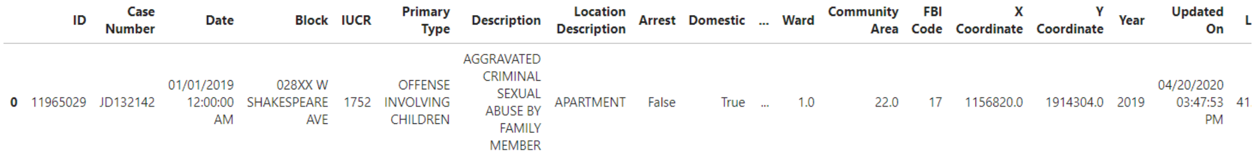
The objective of this project is to obtain the safest communities considering the crime rate. The result of this study is aimed at those interested in living in the city of Chicago.

Chicago is one of the largest cities of the united states with a population of over 2 and a half million people. Chicago has 77 communities grouped into 9 districts. The city has reported more than 7 million crimes of every category since 2001, 259 thousand just in 2019. Business Problem We will deal with the decision of "Which community has had the least crimes in 2019", finding the right community to move into or beginning a business entrepreneurship based on the security and venue(residence) density in each community.

Data

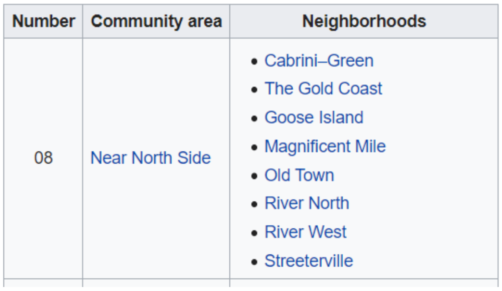
Our study will be based on the data extracted from:

- Crime Data(using the datasets provided by the city of Chicago)



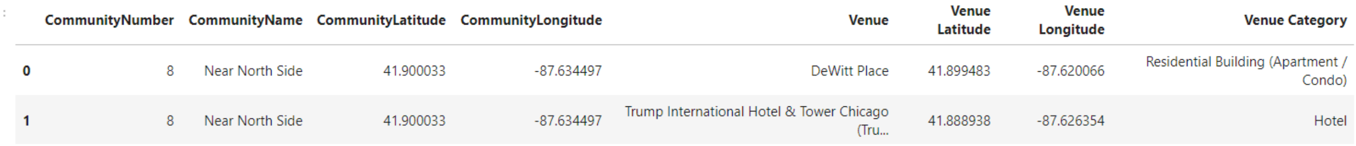
	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	Updated On	L
0	11965029	JD132142	01/01/2019 12:00:00 AM	028XX W SHAKESPEARE AVE	1752	OFFENSE INVOLVING CHILDREN	AGGRAVATED CRIMINAL SEXUAL ABUSE BY FAMILY MEMBER	APARTMENT	False	True	...	1.0	22.0	17	1156820.0	1914304.0	2019	04/20/2020 03:47:53 PM	41

- Community information(scraped from this wikipedia page)
  - Using geopy to get each coordinate



Number	Community area	Neighborhoods
08	Near North Side	• Cabrini–Green • The Gold Coast • Goose Island • Magnificent Mile • Old Town • River North • River West • Streeterville

- Foursquare api to explore residences near our communities



	CommunityNumber	CommunityName	CommunityLatitude	CommunityLongitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	8	Near North Side	41.900033	-87.634497	DeWitt Place	41.899483	-87.620066	Residential Building (Apartment / Condo)
1	8	Near North Side	41.900033	-87.634497	Trump International Hotel & Tower Chicago (Tru...	41.888938	-87.626354	Hotel

Data Cleaning

Cleaning Crime Data

I decided to download and use only 2019 data, that way the dataset would be easier to handle. Originally the data corresponding to 2001-2020 had more than 7 million records, using only 2019 the data subsided to about 259 thousand records.

The original dataset from the city of Chicago contained 22 rows, for this study most columns will not be needed, after removing said columns the dataset is as follows:

Field Name	**Field Type**
Community Area	Int64
Latitude	Float
Longitude	Float

To clean the crime data the following steps were needed:

- Remove all rows where latitude and longitude were empty
- Remove all rows where Community Area was empty
- Rename column Community Area as ‘CommunityArea’, removing the space provided ease of access
- Modifying the dataset to group by ‘CommunityArea’ and creating a new column called **CrimeTotal**

Now our dataset has 77 rows (1 row per community).

Cleaning Community Data

The data downloaded from the City of Chicago has every community Area number, now the Area name was needed. For this part scraped the Wikipedia page https://en.wikipedia.org/wiki/Community_areas_in_Chicago and created a new frame with the code and name of each of the 77 communities.

After having the name, the location data was needed. For this part I used the Geopy library and created a new frame with the following structure:

Field Name	Field Type
CommunityNumber	Int64
CommunityAreaName	String
Latitude	Float
Longitude	Float

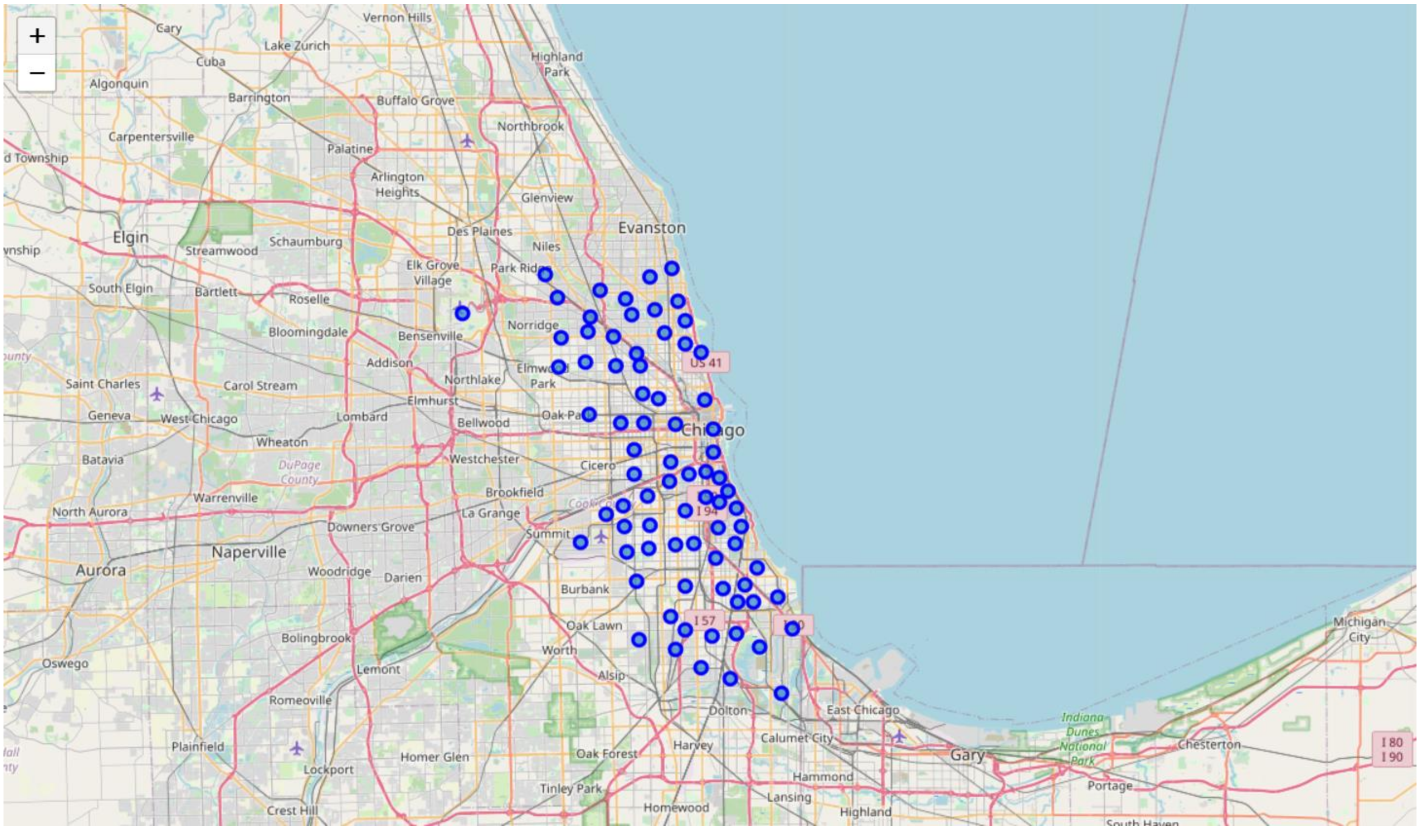


Figure 1 Chicago has 77 communities

Foursquare

Now that I had all my community and crime data, I needed the foursquare residency information. I used only the ‘Residency’ categoryID **‘4e67e38e036454776db1fb3a’**.

After processing each community Location into foursquare I obtained the top 100 Residences of reach location and created a new dataframe to store this information.

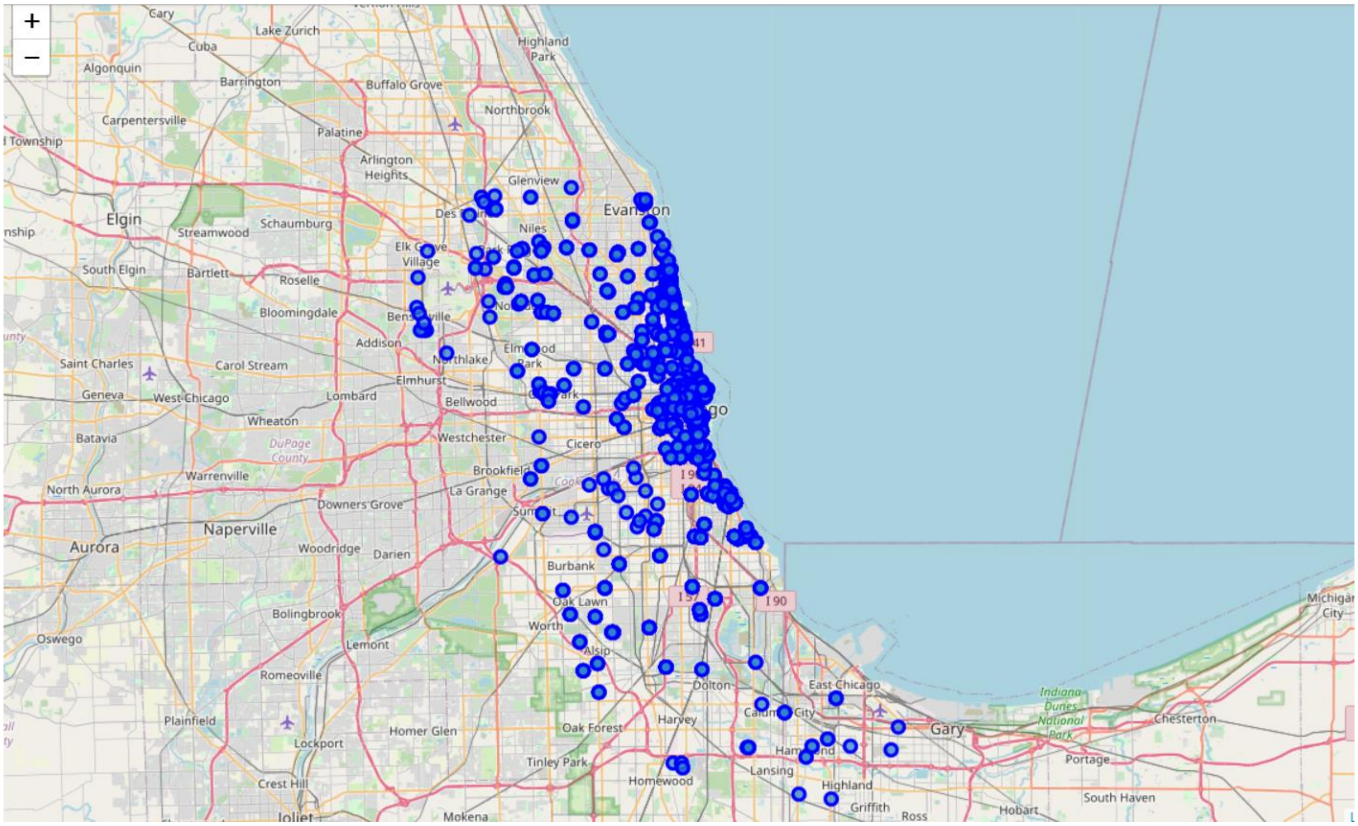


Figure 2Top 100 Residencies from all 77 Communities

Methodology

Now that we have our data, we will follow the following steps:

- We will be doing exploratory analysis on our dataframes.
- We will investigate whether the number of existing 'residency' locations in communities affects the crime rate.
- Our selected algorithm will be **K-means** which we will be using to cluster the different communities.

In order to provide the information as clear as possible, we will be using **scatter plots, heatmaps and overlay maps**.

Analysis

I started the analysis merging all previous datasets together, that way it is possible to have the following structure:

Field Name	Field Type
CommunityNumber	Int
CommunityName	String
CommunityLatitude	Float
CommunityLongitude	Float

VenueTotal	Int
CrimeTotal	Int

The next step is to do an exploratory analysis, since we already know the dimension of our new data frame and its description the next step is trying to find a correlation between each community,crime total and venue total.

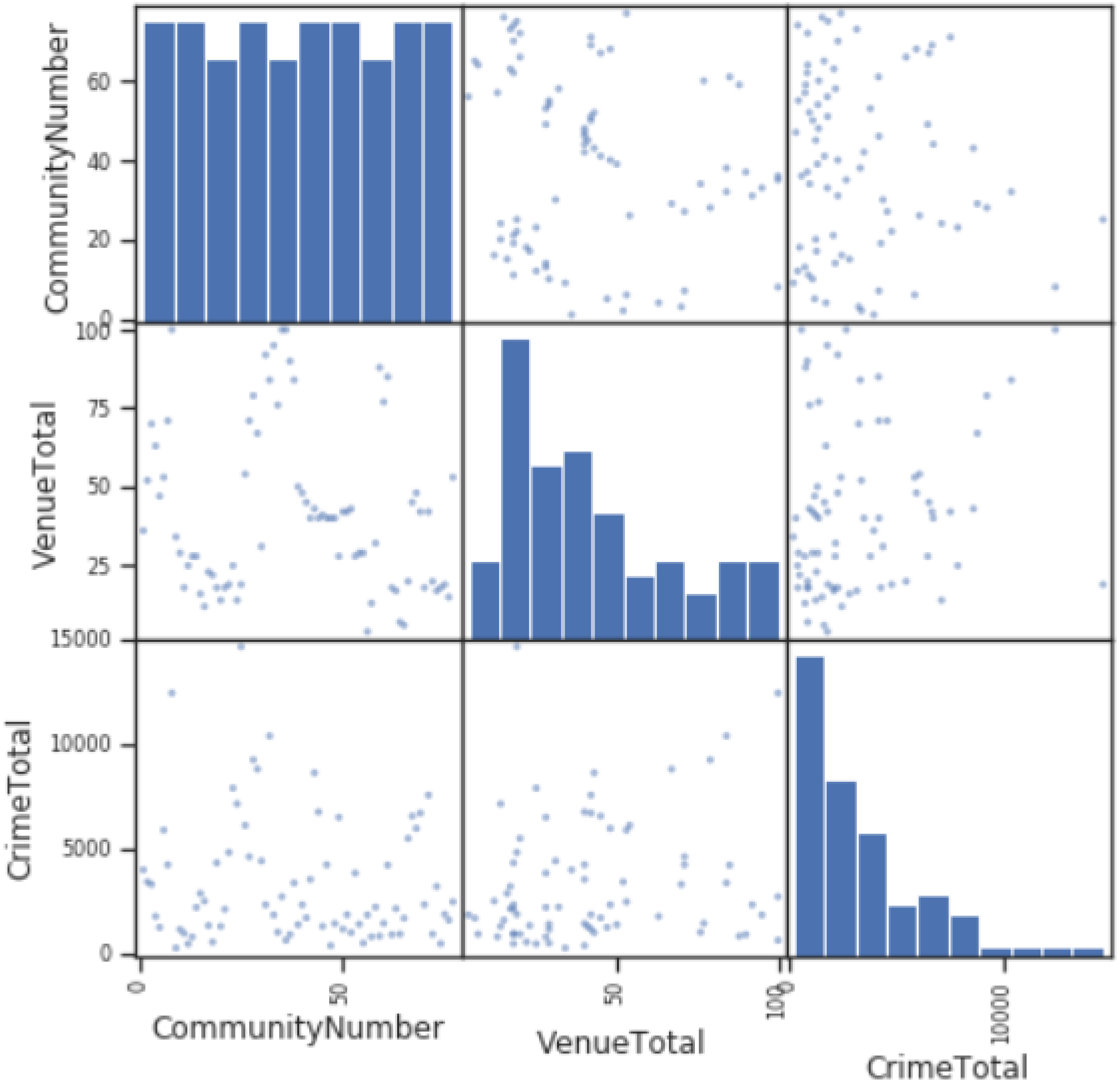


Figure 3 the lack of diagonal grouping suggests a low correlation

Next, we gather the communities with more crime and the communities with more residencies, this in order to try and establish a relation between crimes and residencies.

First, we create a heatmap with the communities with the top 15 crimes

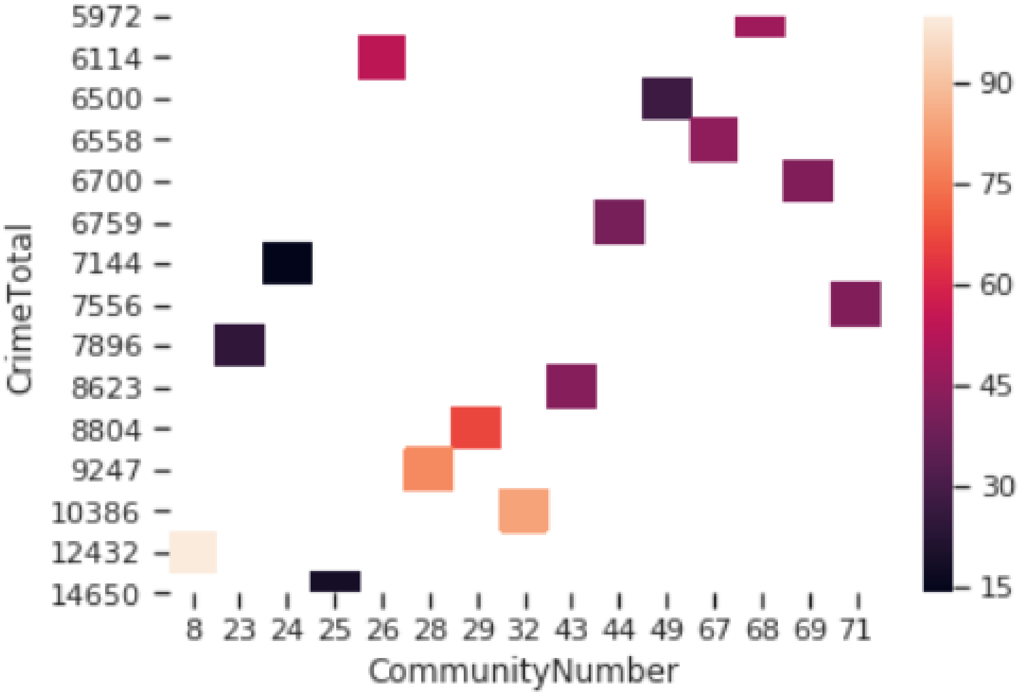


Figure 4 Communities with the TOP 15 crime rate value

Now, we create a heatmap with the communities with more number of residencies.

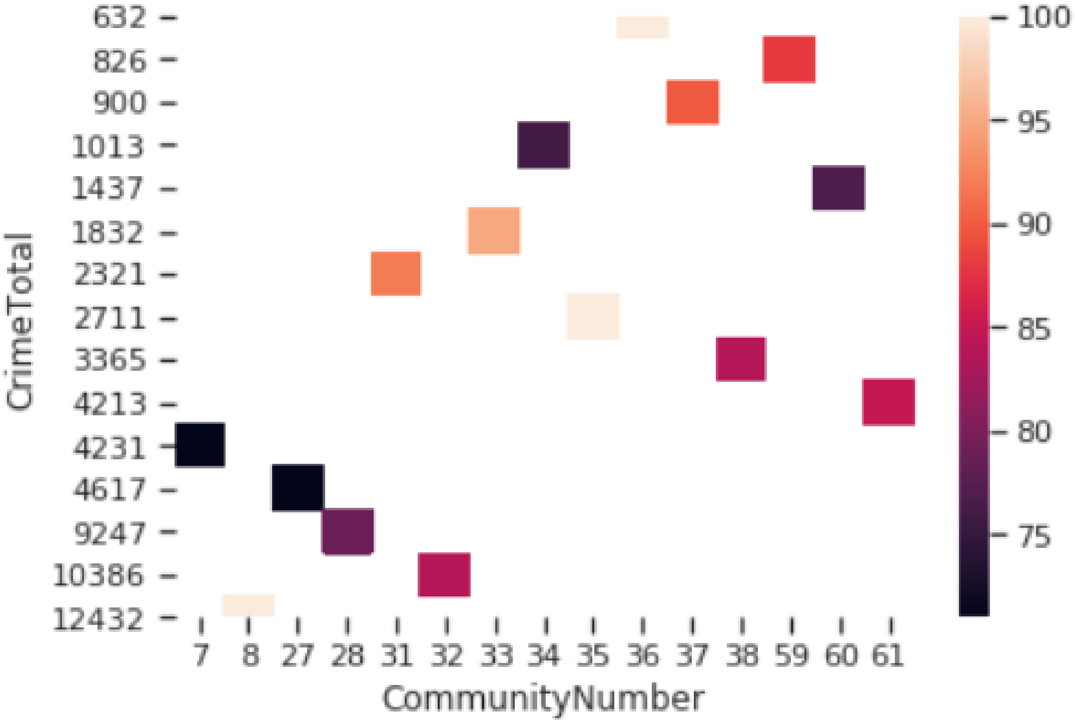


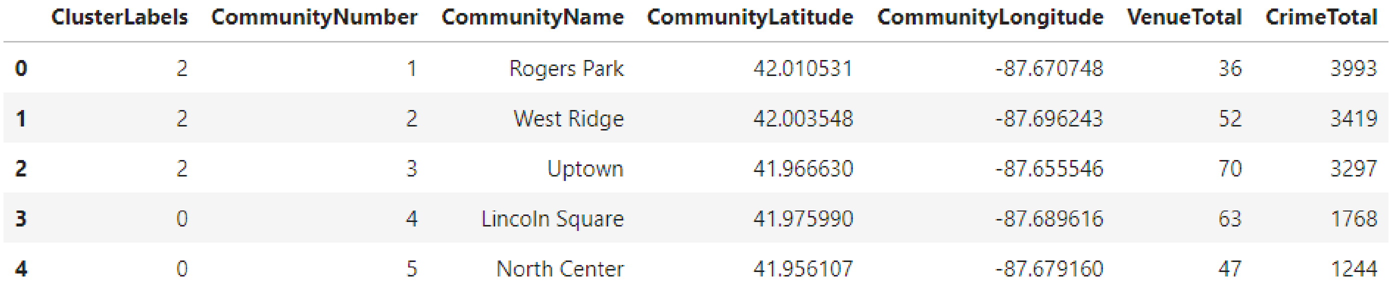
Figure 5 Communities with the TOP 15 residency values

As seen in the above heatmaps **there is no relation** between a community having a high number of residencies and having a high crime rate.

Clustering

For this study I decided to use k-means clustering which is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

I divided the data into 5 Clusters removing the CommunityName, after doing that I modified the dataframe to include the newly created cluster.



	ClusterLabels	CommunityNumber	CommunityName	CommunityLatitude	CommunityLongitude	VenueTotal	CrimeTotal
0	2	1	Rogers Park	42.010531	-87.670748	36	3993
1	2	2	West Ridge	42.003548	-87.696243	52	3419
2	2	3	Uptown	41.966630	-87.655546	70	3297
3	0	4	Lincoln Square	41.975990	-87.689616	63	1768
4	0	5	North Center	41.956107	-87.679160	47	1244

After dividing each cluster, it was easy to piece it together.

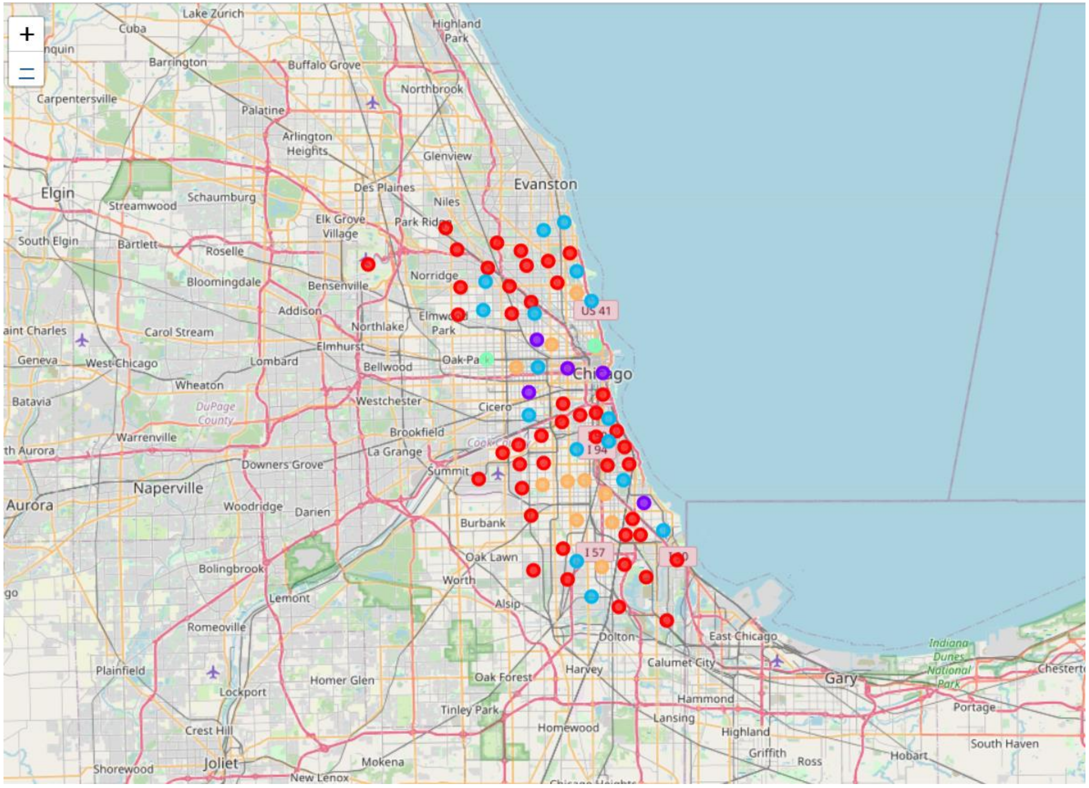


Figure 6 77 Communities in 5 Clusters

After examining each cluster, the data was divided as follows:
Cluster 1: 44 communities, Cluster 2: 5 communities, Cluster 3: 16 communities, Cluster 4: 2 communities, Cluster 5: 10 communities.

The most notorious clusters are (1,2,5). 1 has the least crimes as noted on figure 7, and cluster 2 and 5 have the highest crime rates, for this study I'll show the map of cluster 2 on figure 8.

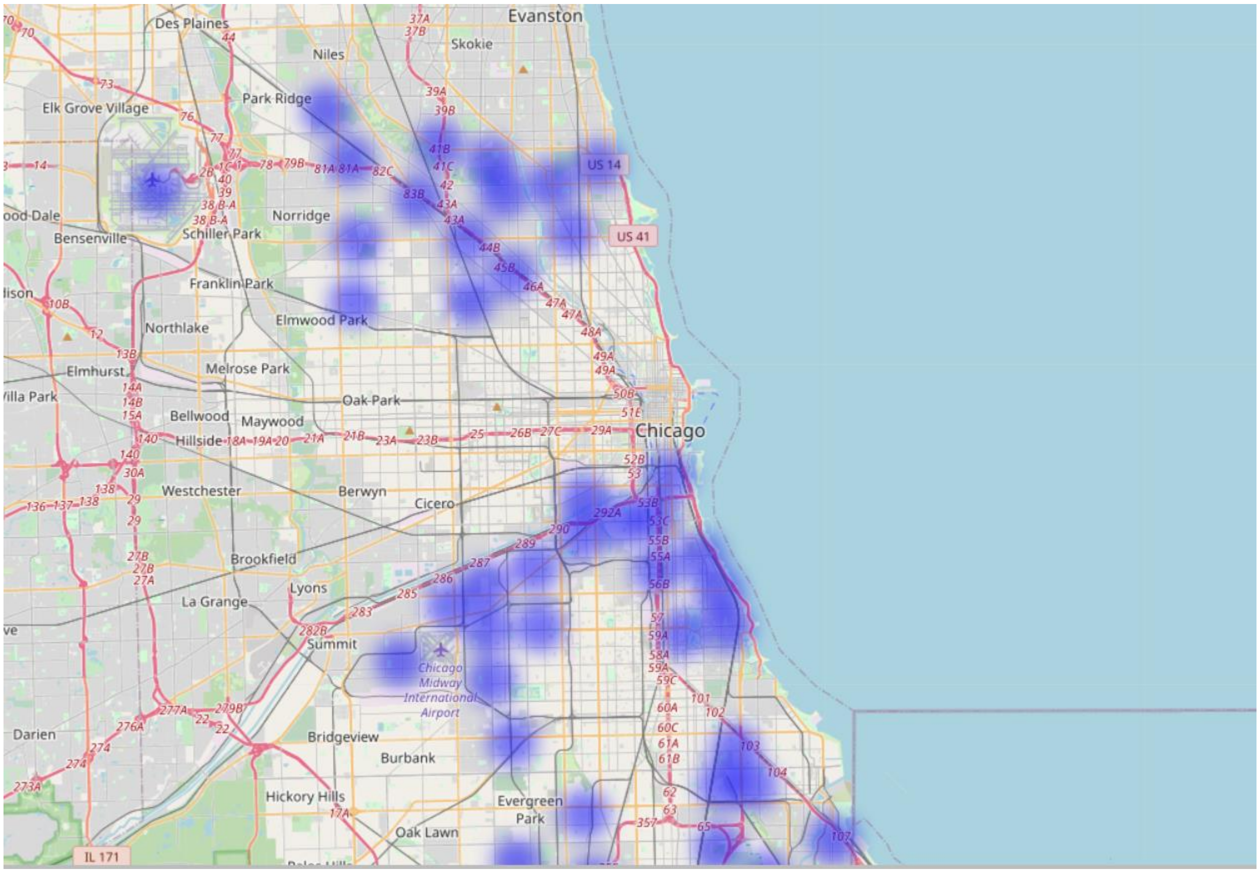
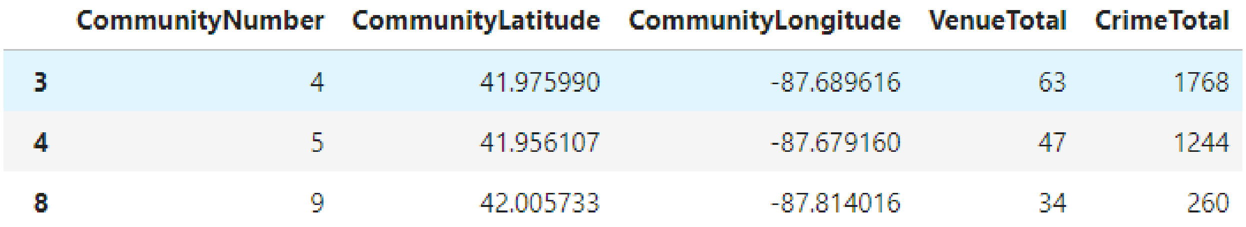


Figure 7 Heatmap with the communities that have the least amount of reported crimes

The heatmap above corresponds to the 44 communities on cluster 1, the crime rate of the sum of cluster 1 is barely above the 5 communities on cluster 2.



	CommunityNumber	CommunityLatitude	CommunityLongitude	VenueTotal	CrimeTotal
3	4	41.975990	-87.689616	63	1768
4	5	41.956107	-87.679160	47	1244
8	9	42.005733	-87.814016	34	260

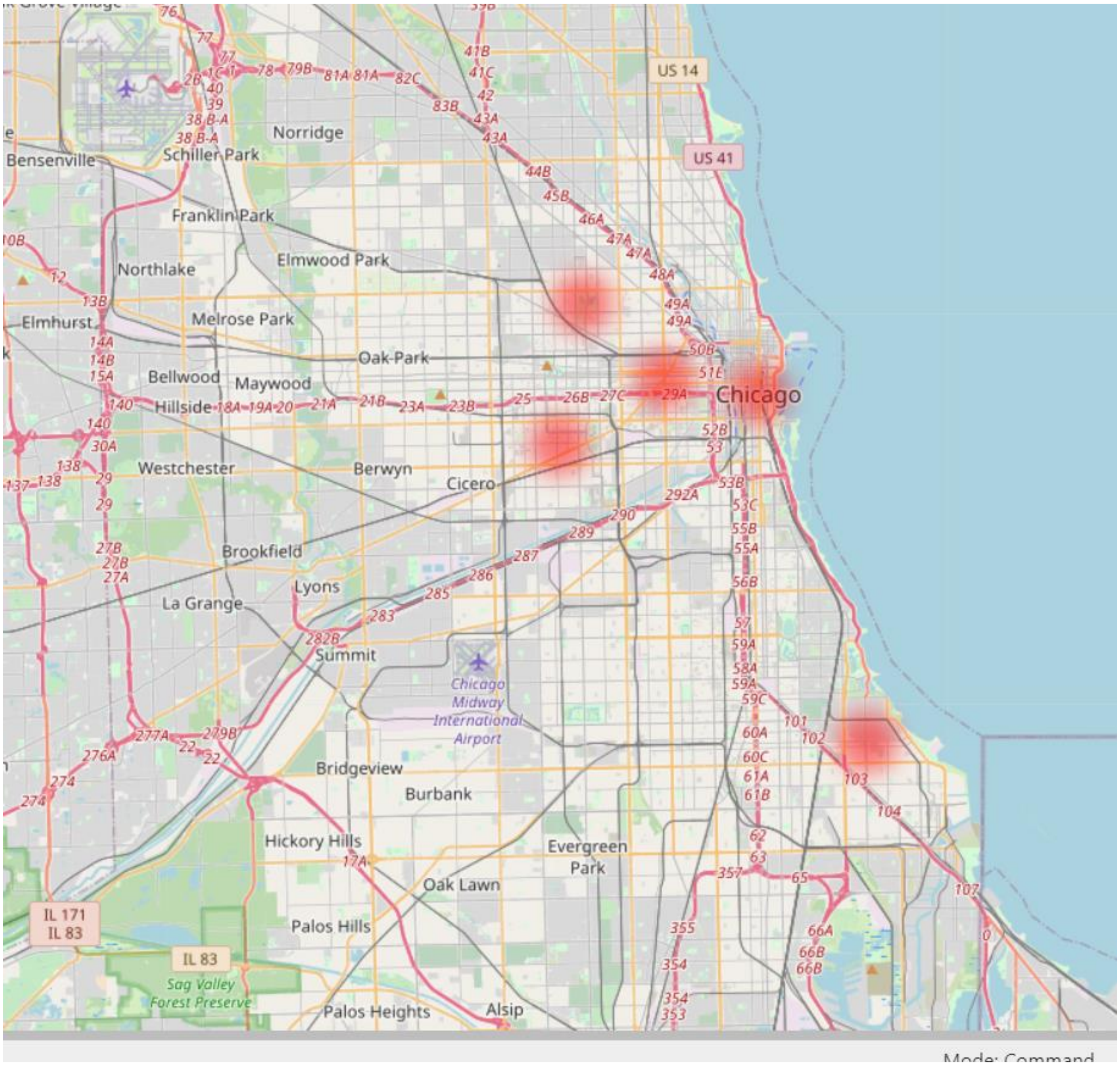
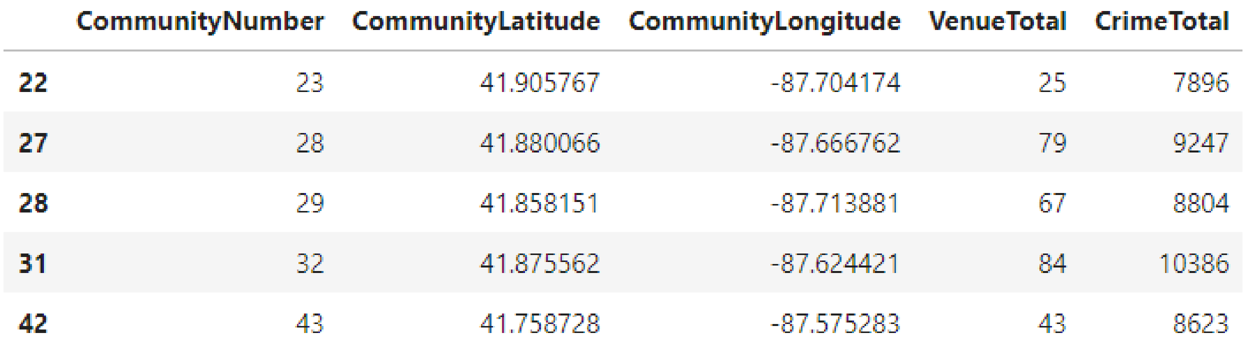


Figure 8 The least safe communities (for these we can choose between cluster 2/4/5), we are using cluster 2

Meanwhile cluster 2 even having only 5 communities has almost the same crime rate of cluster 1.



	CommunityNumber	CommunityLatitude	CommunityLongitude	VenueTotal	CrimeTotal
22	23	41.905767	-87.704174	25	7896
27	28	41.880066	-87.666762	79	9247
28	29	41.858151	-87.713881	67	8804
31	32	41.875562	-87.624421	84	10386
42	43	41.758728	-87.575283	43	8623

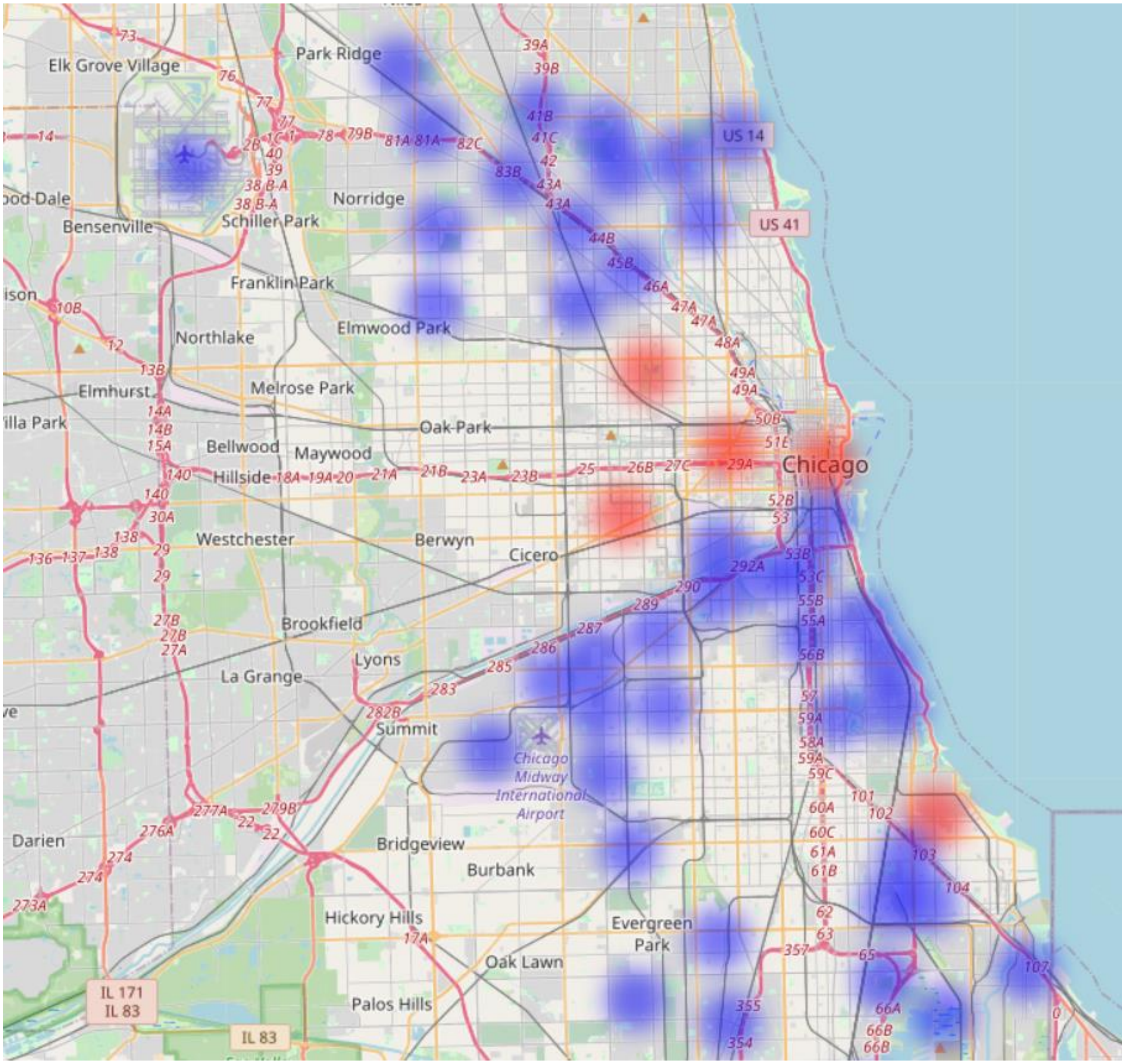


Figure 9 Blue heat shows communities with low crime rate, red heat shows communities with high crime rate.

Results and Discussion

Our study demonstrates that it does not matter how many residencies are in each community, the crime rate does not vary. Also our clusters have teached us the following:

- **The further from the city center the lower the crime rate** as shown in cluster 1, even though it has the most number of communities it has the least amount of crimes, in the map of cluster 1 we can see a diagonal pattern from city center to the outskirts.

- The communities that were in the center of the city have a high crime rate, as shown in group 2, analyzing the map of group 2 we can see that **most of the heat is concentrated near the center**, without taking into account the single outlier to the right.

Conclusion

The man objective of our study was to categorize communities as to show potential residents which areas are the best to live based on crime rate. And the **result shows clearly that everywhere not in the center is a good option**. Of course, this is made based on ***Crime rate*, it would be good to add more variables like: rent cost, average income and transportation.