SAVING CHICAGO WITH DATA SCIENCE

September 15, 2019

Emmanuel Akanbi

A. INTRODUCTION

Since the beginning of the 20th century, The Chicago Police Department's Bureau of Records have been tracking crimes in Chicago. In 2016 it was projected that the U.S. crimes would rise by 13%, but half would be due to the violence in the Chicago.[1] There are many factors that contributes to crime rate, but in this project education and overall income will be the main focus. Education gives people the opportunity to expand their scope in how to live an effective life. With the lack thereof, a person may be tempted to go the wrong path. Recent research suggested that lack of education can contribute to the rise of crime rate. For example, young adults dropping out of school due to personal or environmental factors.

In this project, we will determine the most optimal location to place a new nonprofit education center, Brighter Days Academy, in Chicago where they will be providing opportunities for young adults and adults to aid them in getting their GED, learn how to manage their finances, expose them opportunities with a college degree or trade school and more.

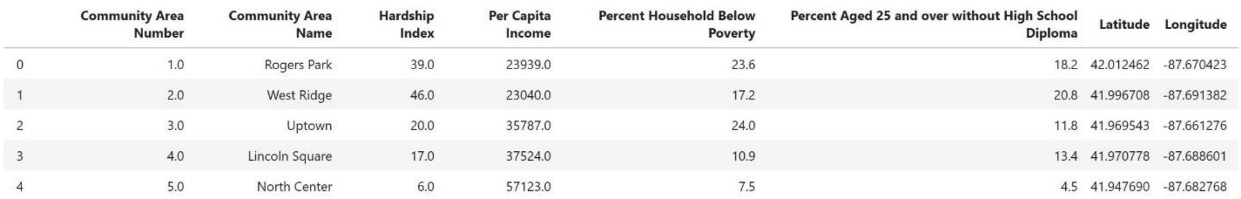
B. DATA

In order to select the optimal location for the educational center, the following data was gathered and analyzed:

- The location and name of Chicago public schools were obtained using Foursquare API this was used to confirm the location of the Chicago community areas.
- Chicago's crime data was gathered using the data from the Chicago Data Portal that recorded incidents that occurred in 2019. This data .json file will be cleaned to make a choropleth map to show the heavy occurrence of crimes in each region.
- Chicago's socioeconomic data was gathered from the Chicago Data Portal that contains a selection of six socioeconomic indicators of public health significance.

C. METHODOLOGY

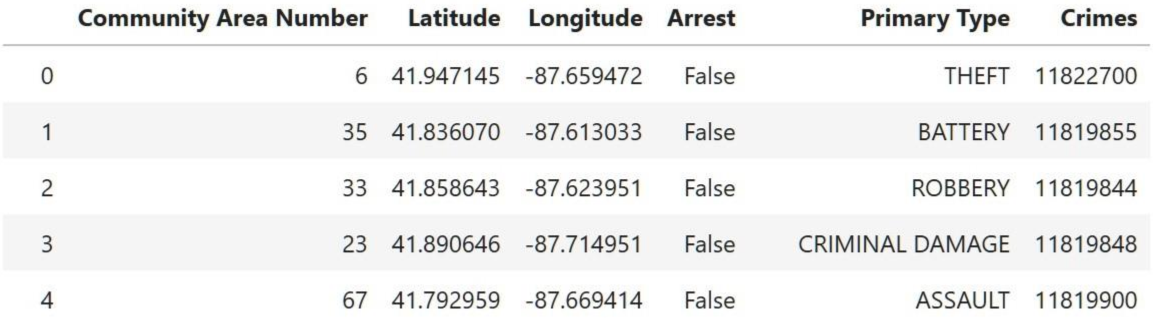
By using the Foursquare API, the names and location of Chicago public schools were gathered. This was then merged and cleaned with Socioeconomic data from the Chicago Data Portal.



	Community Area Number	Community Area Name	Hardship Index	Per Capita Income	Percent Household Below Poverty	Percent Aged 25 and over without High School Diploma	Latitude	Longitude
0	1.0	Rogers Park	39.0	23939.0	23.6	18.2	42.012462	-87.670423
1	2.0	West Ridge	46.0	23040.0	17.2	20.8	41.996708	-87.691382
2	3.0	Uptown	20.0	35787.0	24.0	11.8	41.969543	-87.661276
3	4.0	Lincoln Square	17.0	37524.0	10.9	13.4	41.970778	-87.688601
4	5.0	North Center	6.0	57123.0	7.5	4.5	41.947690	-87.682768

Figure 1: Socioeconomic Data merged with Foursquare

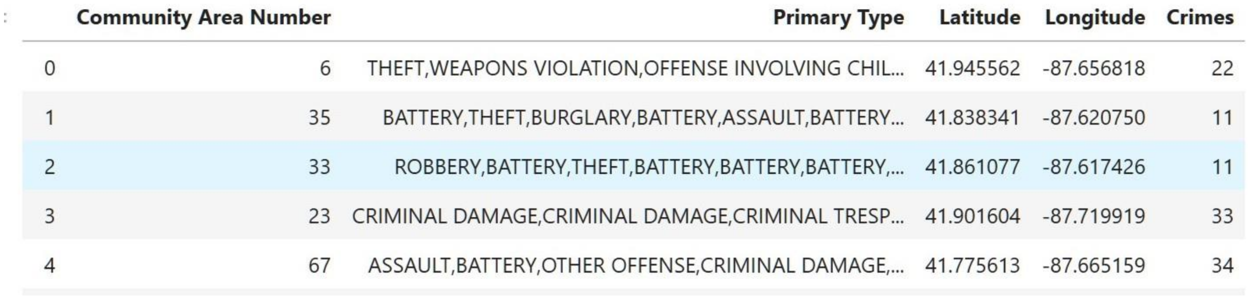
To examine the relation of the Socioeconomic data to crimes in Chicago, the Chicago crime data .json file was imported via the Chicago Data Portal. The date below provides information of the crimes such as the ID, type of crimes and location, which was in the form of coordinates and community area number.



	Community Area Number	Latitude	Longitude	Arrest	Primary Type	Crimes
0	6	41.947145	-87.659472	False	THEFT	11822700
1	35	41.836070	-87.613033	False	BATTERY	11819855
2	33	41.858643	-87.623951	False	ROBBERY	11819844
3	23	41.890646	-87.714951	False	CRIMINAL DAMAGE	11819848
4	67	41.792959	-87.669414	False	ASSAULT	11819900

Figure 2: Chicago crime data

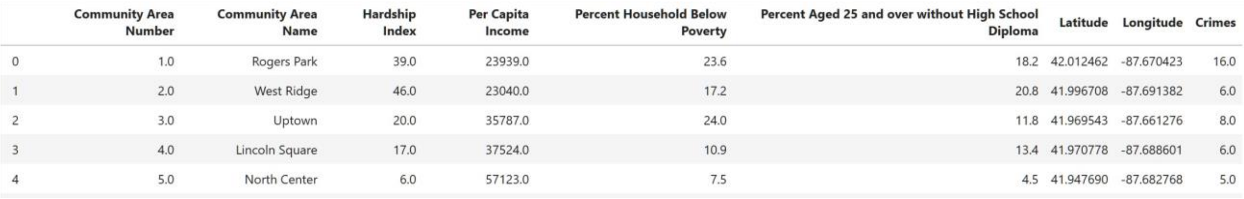
To relate the crime data to the socioeconomic data, the crime data was grouped by its community area number and by the number of crimes per group to obtain the total amount crimes that occurred per community area number.



	Community Area Number	Primary Type	Latitude	Longitude	Crimes
0	6	THEFT,WEAPONS VIOLATION,OFFENSE INVOLVING CHIL...	41.945562	-87.656818	22
1	35	BATTERY,THEFT,BURGLARY,BATTERY,ASSAULT,BATTERY...	41.838341	-87.620750	11
2	33	ROBBERY,BATTERY,THEFT,BATTERY,BATTERY,BATTERY,...	41.861077	-87.617426	11
3	23	CRIMINAL DAMAGE,CRIMINAL DAMAGE,CRIMINAL TRESP...	41.901604	-87.719919	33
4	67	ASSAULT,BATTERY,OTHER OFFENSE,CRIMINAL DAMAGE,...	41.775613	-87.665159	34

Figure 3: Grouped crime data

The socioeconomic data and the crime data were then merged by community area number in preparation for visualization.



	Community Area Number	Community Area Name	Hardship Index	Per Capita Income	Percent Household Below Poverty	Percent Aged 25 and over without High School Diploma	Latitude	Longitude	Crimes
0	1.0	Rogers Park	39.0	23939.0	23.6	18.2	42.012462	-87.670423	16.0
1	2.0	West Ridge	46.0	23040.0	17.2	20.8	41.996708	-87.691382	6.0
2	3.0	Uptown	20.0	35787.0	24.0	11.8	41.969543	-87.661276	8.0
3	4.0	Lincoln Square	17.0	37524.0	10.9	13.4	41.970778	-87.688601	6.0
4	5.0	North Center	6.0	57123.0	7.5	4.5	41.947690	-87.682768	5.0

Figure 4: Crime data and Socioeconomic data merged

To get an understanding of the data, the community areas were partitioned into groups that have similar features. To effectively segment the community area, the communities were clustered based on their location and hardship index using the Density-Based Clustering of Application with Noise (DBSCAN) algorithm. After running the DBSCAN algorithm the community areas were portioned into 3 mutually exclusive clusters. Using the matplotlib base map, the communities were visualized to respect to their clusters. This is shown below:

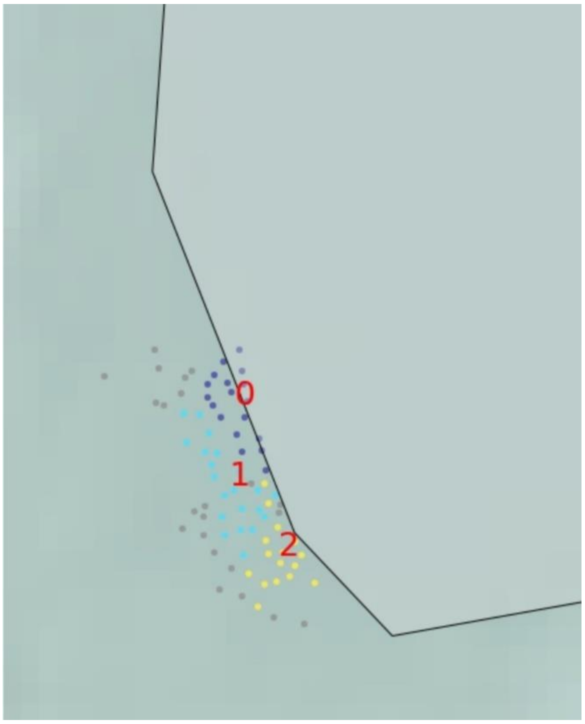


Figure 5: DBSCAN of communities based off their location and hardship index

A profile was then created for each cluster. The data was first normalized so that the respected features can be on the same scale.

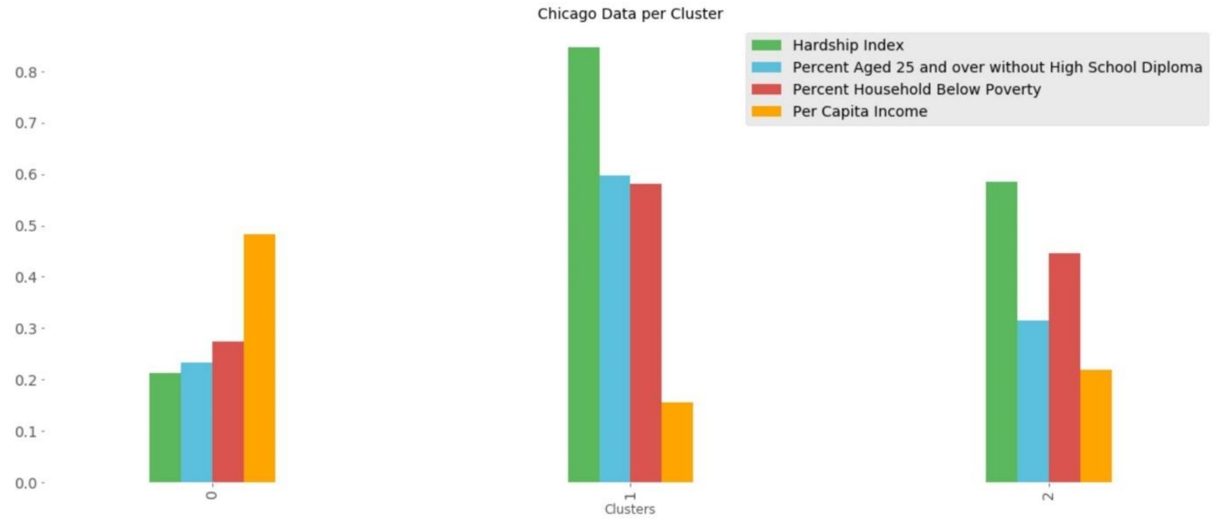


Figure 6: Bar chart of clusters

After examining the graph, each cluster was categorized as follows:

- Cluster 0: Low Hardship Index, High Education, High Per Capita Income
- Cluster 1: High Hardship Index, Low Education, Low Per Capita Income
- Cluster 2: Medium Hardship Index, Medium Education, Medium Per Capita Income

D. RESULTS

To visualize the correlation of data between the hardship index and the crimes in Chicago, a choropleth map was used. The markers correspond to the cluster from Figure 6.

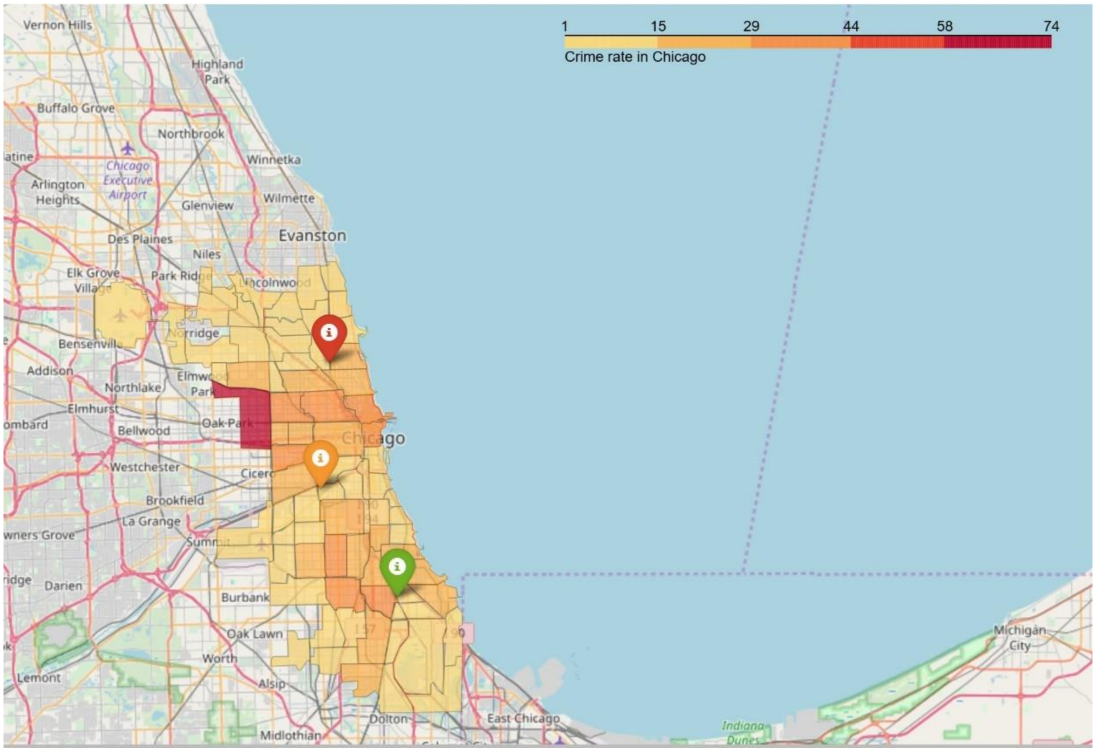


Figure 7: Choropleth map based off the cluster and crime rate

The markers correspond as follows:

- Red Marker, Cluster 0: Low Hardship Index, High Education, High Per Capita Income
- Orange Marker, Cluster 1 High Hardship Index, Low Education, Low Per Capita Income
- Green Marker, Cluster 2: Medium Hardship Index, Medium Education, Medium Per Capita Income

E. DISCUSSION

The DBSCAN algorithm was used to cluster the Chicago communities based on their location and hardship index. A bar chart was then created (Figure 6) for further visualization and understanding of the given features. Cluster 1 was determined to have the highest hardship index, lowest education rate, and lowest per capita income.

The clustered data was then combined with the crime data to the choropleth map shown in Figure7. From the choropleth map, it is apparent that the high crime rates are near Cluster 1. In the choropleth map it also show that the high crimes rates are also near Cluster 0 even though this cluster has the best result. This may be due to more factors than the factors that were provided such as pollical corruption and much more.

F. CONCLUSION

Based on the result, Brighter Days Academy should be in Cluster 1 due to its high hardship index, low education rate and low per capita income. This community area is the best candidate to place an educational center for young adults and adults to get their GED, gain some financial knowledge and much more.

G. REFERENCES

1. Chicago Crime
2. Prison Data
3. Chicago Data Portal