Save the Children, Save the City

Public Schools and Crime Data Analysis of Chicago

September 8, 2019

Jeremy Martin

A. Introduction

The city of Chicago's overall crime rate is significantly higher than the US average. In 2016, Chicago was responsible for nearly half the year's increase in homicides in the US. In 2010, more than 40% of females, and nearly 50% of males admitted to Cook County jail did not complete high school at the time of admission. Recent research has suggested that policies aimed to increase educational attainment and improve school quality can significantly reduce crime rates.

In this project, we will try to find an optimal location to place a new nonprofit community center, HERO (Helping Everyone Receive Opportunities), in Chicago. Specifically, this analysis will be intended for stakeholders interested in opening a non-profit educational community center to combat the high crime rates, provide aid to impoverished communities, and keep the youth of the perilous areas of Chicago off the streets.

B. Data

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

- The name and location of Chicago Public Schools was gathered via Foursquare API [1].
- Via the Chicago Data Portal [2] (CDP), the Chicago Public Schools dataset was employed to present all school level performance data used to create CPS School Report Cards for the 2011 academic school year. The .csv file provided scores from the 5Essentials Survey [3] which asked students about their experiences and feelings (such as how safe they felt at school and if they got support from their teachers). The dataset also contained each school's rate of misconducts per 100 students, graduation rate, and college eligibility rate.
- Chicago crime data was also gathered from the Chicago Data Portal that reported incidents of crime (with the exception of murders) that occurred in the City of Chicago in 2019. We will clean this .json file to make a choropleth map to examine the crime heavy community areas.
- Chicago socioeconomic data was collected from the Chicago Data Portal that contained a selection of six socioeconomic indicators of public health significance. We will clean this .json file to make a choropleth map to examine the community areas with the lowest per capita income.

C. Methodology

Using Foursquare API, I was able to gather the names of Chicago Public Schools along with their coordinates. I then cleaned and merged the data with the Chicago Public Schools data (via Chicago Data Portal) to attain further relevant information on the schools provided by the Foursquare API search.

	name of school	elementary or high school	community area name	community area number	school lat	school long	safety score	environment score	instruction score	rate of misconducts per 100 students	graduation rate	college eligibility
0	Abraham Lincoln Elementary School	ES	LINCOLN PARK	7	41.924497	-87.644522	99.0	74.0	66.0	2.0	NDA	NDA
1	Adam Clayton Powell Paideia Community Academy ...	ES	SOUTH SHORE	43	41.760324	-87.556736	54.0	74.0	84.0	15.7	NDA	NDA
2	Adlai E Stevenson Elementary School	ES	ASHBURN	70	41.747111	-87.731702	61.0	50.0	36.0	2.3	NDA	NDA
3	Agustin Lara Elementary Academy	ES	NEW CITY	61	41.809757	-87.672145	56.0	45.0	37.0	10.4	NDA	NDA
4	Air Force Academy High School	HS	ARMOUR SQUARE	34	41.828146	-87.632794	49.0	60.0	55.0	15.6	NDA	NDA

Figure 1: Head of the dataset containing the merged data from Foursquare and CDP

To examine the performance of the schools per neighborhood, I grouped the data above by community area name to attain the average school *safety score*, *environment score*, *instruction score*, *student misconduct rate*, *graduation rate*, and *college eligibility rate* in each community. By grouping the data by the community area name, I was also able to gather the central point of the schools in each community area. Using this information along with Python's folium library, I was able to visualize the geographic details of the Chicago public schools in each community shown in Figure 2.

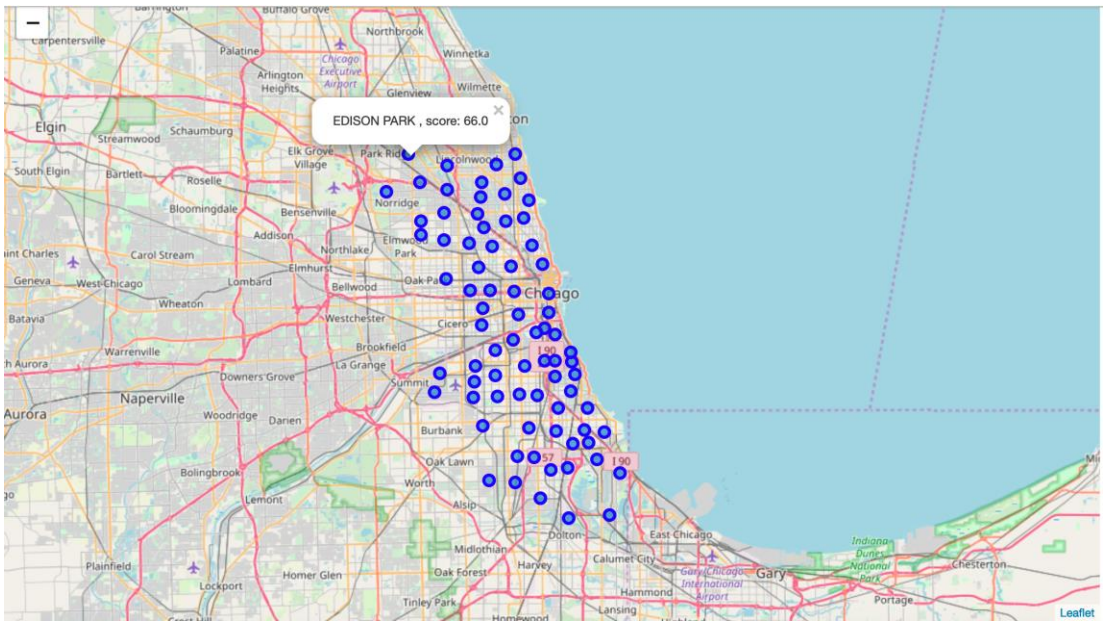

Figure 2: Folium map showing CPS per community along with their average school environment score.

In order to gain a better understanding of this unsupervised Chicago public school data, I partitioned a community area into groups of communities that have similar characteristics pertaining to the schools within those communities. In order to effectively segment the community areas, I clustered the communities based on their schools' locations and rate of misconducts per 100 students using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm.

DBSCAN from sklearn library is able to run DBCAN clustering from vector array or distance matrix. For this project, I passed it to a Numpy array to find core samples of high density and expanded clusters from them. Density, in this case, was defined as the number of points (minimum number of samples = 4) within the specified radius (epsilon) of 0.7.

After running the DBSCAN algorithm, I was able to partition the community areas into 4 mutually exclusive clusters. Using the matplotlib basemap toolkit, I was able to visualize the communities as well their respective clusters on basemap shown below:

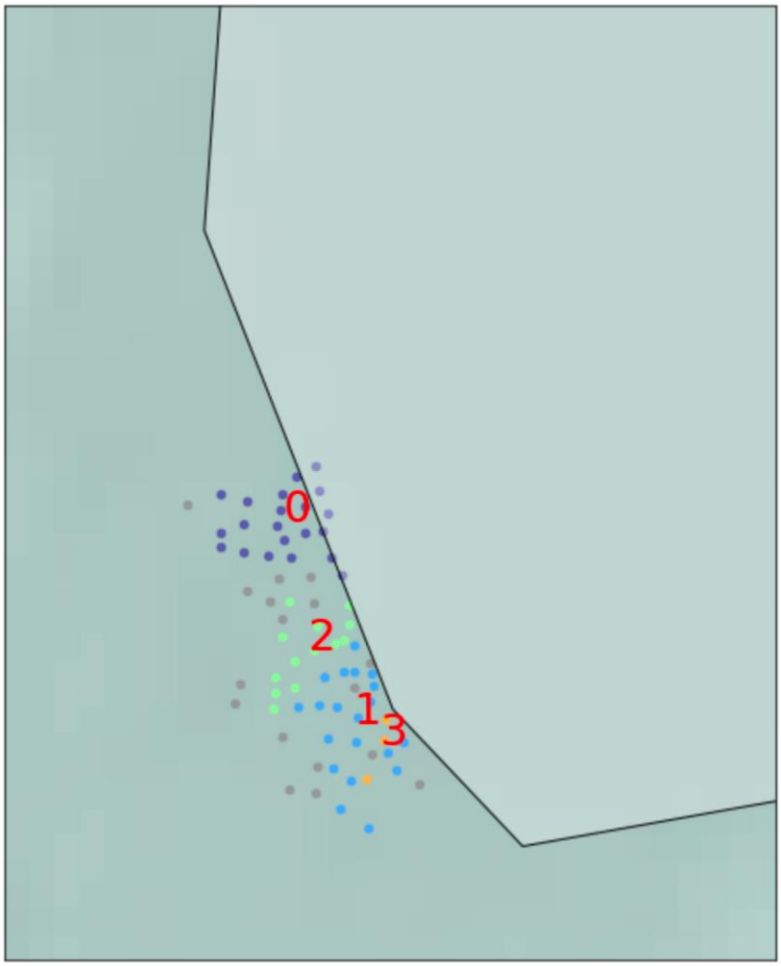

Figure 3: Clustering of station based on their location and rate of misconducts per 100 students

I was then able to create a profile for each cluster (excluding the outliers in grey), considering the common characteristics of each cluster shown in the bar chart below.

Clus_Db	index	Lat	Long	safety score	environment score	instruction score	misconducts_rate	graduation rate	college eligibility	
0	40.833333	41.955942	-87.656650	63.0	54.0	48.0	16.45	68.24	40.24	4
1	30.619048	41.758938	-87.621412	38.0	44.0	47.0	30.38	59.66	15.39	
2	45.366667	41.905100	-87.721843	60.0	49.0	47.0	7.81	60.54	25.61	4
3	25.333333	41.736302	-87.587802	38.0	41.0	46.0	41.25	42.97	9.93	5

Figure 4: Table of DB Clusters with their average school scores and rates

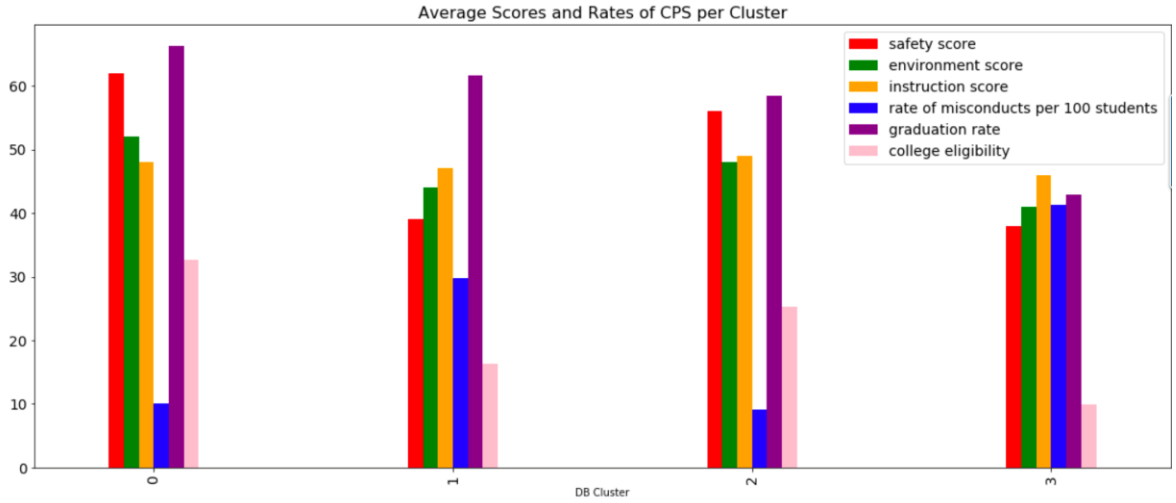

Figure 5: Bar chart displaying average scores and rates of CPS per DB Cluster

After examining the graph in Figure 5, I was able to classify each cluster as follows:

- Cluster 0: "High Safety, High Grad rate, High Eligibility, Low Misconducts per 100 students"
- Cluster 1: "Low Safety, Medium Grad Rate, Low Eligibility, Medium Misconducts per 100 students"
- Cluster 2: "Medium Safety, Medium Grad Rate, Medium Eligibility, Low Misconducts per 100 students"
- Cluster 3: "Low Safety, Low Grad Rate, Low Eligibility, High Misconducts per 100 students"

To examine the relation of the Chicago Public School data to crimes in Chicago, I had to import, clean and wrangle the Chicago crime data .json file provided by the Chicago Data Portal. The dataset below provides the crime's id, the type of crime, and the location (in the form of coordinates and community area) of which the crime occurred:

	crime id	primary type	community area number	crime lat	crime long
0	11813565	CRIMINAL DAMAGE	14	41.968276	-87.724912
1	11813507	NARCOTICS	67	41.793255	-87.664563
2	11813499	BATTERY	46	41.746321	-87.563113
3	11813910	BURGLARY	16	41.953953	-87.702439
4	11814209	ASSAULT	48	41.732485	-87.582776

Figure 6: Head of crime table showing Crime ID, Primary Type, and location of crimes in 2019

In order to relate the crime data to the clustered CPS data, I had to group the crime data by its community area number and count the *crime id* per group to attain a total number of crimes that occurred per community area. I will eventually use this data to create a choropleth map with the CPS clusters to visualization their correlation.

	community area number	crime id
0	1	26
1	2	8
2	3	13
3	4	4
4	6	26

Figure 7: Head of grouped crime table showing total number of crimes per community area in 2019

I imported, cleaned, and wrangled the Chicago socioeconomic .json file to examine the relation of the Chicago Public School data to per capita income in Chicago. I also grouped the socioeconomic data by *community area name* and took the mean of the per capita income to attain Chicago's per capita income by community.

	community_area_name	per_capita_income_
0	ALBANY PARK	21323
1	ARCHER HEIGHTS	16134
2	ARMOUR SQUARE	16148
3	ASHBURN	23482
4	AUBURN GRESHAM	15528

Figure 8: Head of socioeconomic data displaying per capita income by community in Chicago

D. Results

To effectively visualize the correlation of data between the performance of public schools and crimes in Chicago as well as the correlation of public schools and socioeconomic status in Chicago, I bound the crime data and socioeconomic data (from Figures 7 and 8 respectively) to their own maps for choropleth visualizations. I passed visualization as markers on the map to denote the clusters of the CPS data shown in Figure 4.

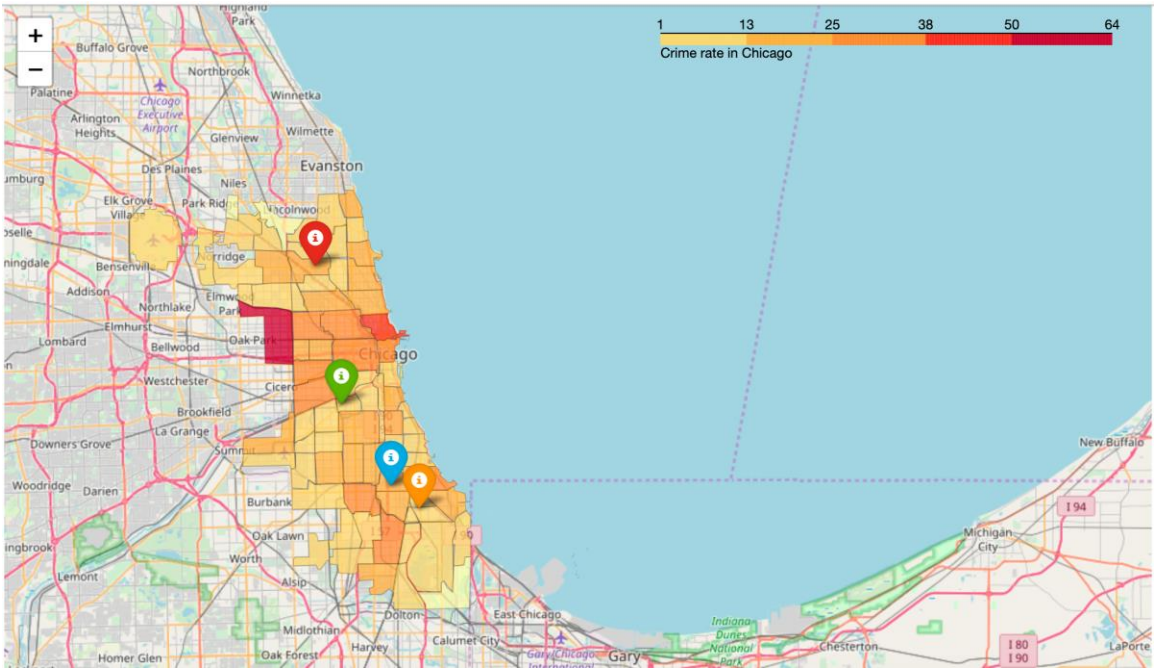

Figure 9: Choropleth map of crime rate in Chicago (2019) with markers representing CPS data clusters

In the map above, I incorporated markers representing clusters gathered from the DBSCAN algorithm. Each colored marker is denoted as the following:

- Red Marker/Cluster 0: "High Safety, High Grad rate, High Eligibility, Low Misconducts per 100 students"
- Green Marker/Cluster 2: "Medium Safety, Medium Grad Rate, Medium Eligibility, Low Misconducts per 100 students"
- Blue Marker/Cluster 1: "Low Safety, Medium Grad Rate, Low Eligibility, Medium Misconducts per 100 students"
- Orange Marker/Cluster 3: "Low Safety, Low Grad Rate, Low Eligibility, High Misconducts per 100 students"

The markers were used as a featured group on the choropleth map using the Python Library: Folium. With this map, we can visualize the areas with the highest crime rate in Chicago in relation to the clusters of schools.

The same principle was applied using the socioeconomic data to visualize how the CPS clusters/markers relate to the per capita income by community in Chicago.

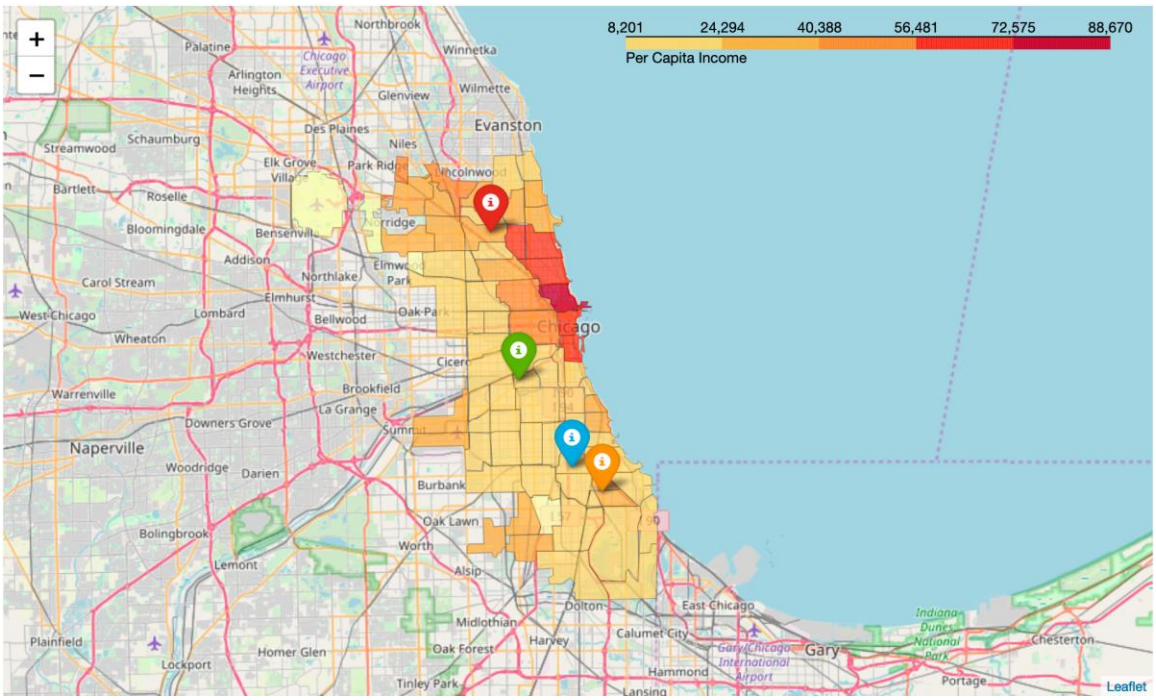

Figure 10: Choropleth map of per capita income by community with clustered CPS markers

E. Discussion

Using the DBSCAN algorithm, we were able to cluster the Chicago Public Schools data based on their location and rate of misconducts per 100 students. After examining the data within each of the four clusters displayed in the bar chart (in Figure 5), we were able to create a profile for each cluster. Cluster 0 and Cluster 2 contained schools on average with good standing: medium-high safety scores, medium-high graduation rates, medium-high college eligibility, and low rate of misconducts. The locations of these clusters were in the central and north side of the city. On the other hand, Cluster 1 and Cluster 3 contained schools on average with less desirable results: low safety scores, medium-low graduation rates, low college eligibility, and medium-high rate of misconducts. Both Cluster 1 and 3 were in the southern part of the city. So why are the schools on the southside of Chicago having a higher rate of misconducts and lower college eligibility then the schools in the central and northside?

Could the crime rate within these communities influence the poor scores and rate of misconducts of their respective schools? Well after binding the crime data onto the choropleth map (shown in Figure 9) along with the CPS clusters, we fail to see a direct correlation between the crime rate and rate of misconduct in schools in the community. Perhaps other factors or features of the crime dataset set needed to be implemented to drive the argument that high crime rates in the community drive poor results in the schools within the community.

To provide the stakeholders for H.E.R.O with a solid suggestion on where to place their community center, I decided to make another choropleth map binding socioeconomic data in Chicago along with the CPS clusters (displayed in Figure 10). Based on this map, we can see Cluster 2 and Cluster 0 are in community with high per capita income on average, whereas Clusters 1 and 3 are in communities with low per capita income on average (with Cluster 1 being the lowest).

F. Conclusion

Based on the results, H.E.R.O. should be in the Cluster 1 due to its schools' low rate of college eligibility, high rate of student misconducts, and community's low per capita income.

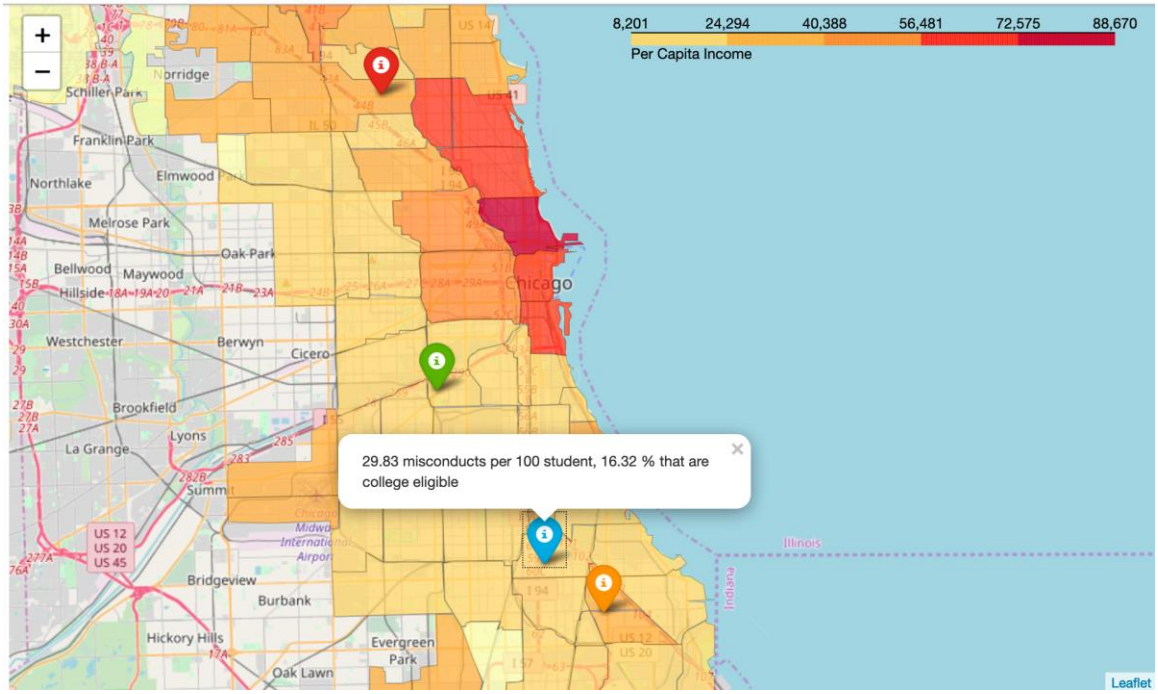

Figure 11: Choropleth map of per capita income by community with clustered CPS markers

This community area is a prime candidate to place the community center where it can greatly impact the community by providing it with free to affordable after school programs aimed at making sure every student is where they're supposed to be academically. The center will also provide a food shelf to help the families in unfortunate circumstances. That way the kids can worry less about what they are going to eat that night and more about what they are going to learn that day. The center will also offer mentoring programs aimed at preparing the kids to think about college and what they need to do and accomplish in order to be eligible to attend the colleges of their interest.

G. References

1. Foursquare API
2. Chicago Data Portal
3. 5Essential Survey