Segmenting and analyzing Chicago schools

Aylin Mousavian

October 2020

Table of content:

Introduction

- Although growing national attention has focused on high rates of required participation in college admission, I reviewed Chicago schools to find best schools which features included college enrollments rates, CPS Performance Policy Level, SAFETY_SCORE, Environment Score, Leaders Score, Teachers Score, Parent Engagement Score, Parent Environment Score and Rate of Misconducts (per 100 students) in its rating systems. Admission and enrollments rates are generally collected state by state through statewide college reporting databases from each schools.

- Although there is increased nationwide focus on college and career readiness, so the influence of college enrollment would be higher than other factors

Data acquisition and cleaning

- This dataset belongs to Chicago schools, and the features contain their important items that specify their score among other schools. This dataset has been prepared by IBM during data analyst course for development.
- Slots with blank and NDA cells would be cleaned using average value to be replaced.
- Data frame has 566 columns which every column define each school properties of Chicago city.

factors that will influence our decision are:

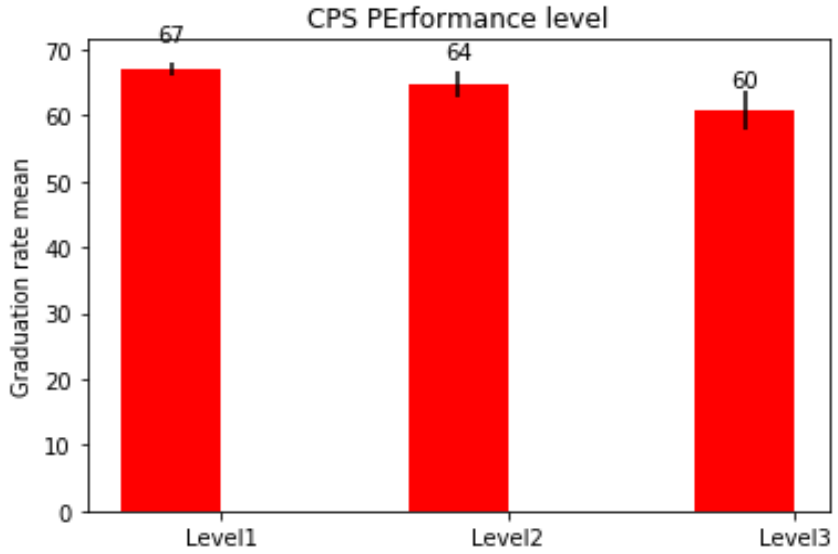
- Influential features which has effect on schools performance.
- Function which made for calculating the score of school.
- How other feature influence on college preparation in this problem.

Exploratory Data Analysis

- **Relationship between CPS performance level and graduation rate**.

  It is widely accepted that CPS performance level policy is one of important feature among schools quality.

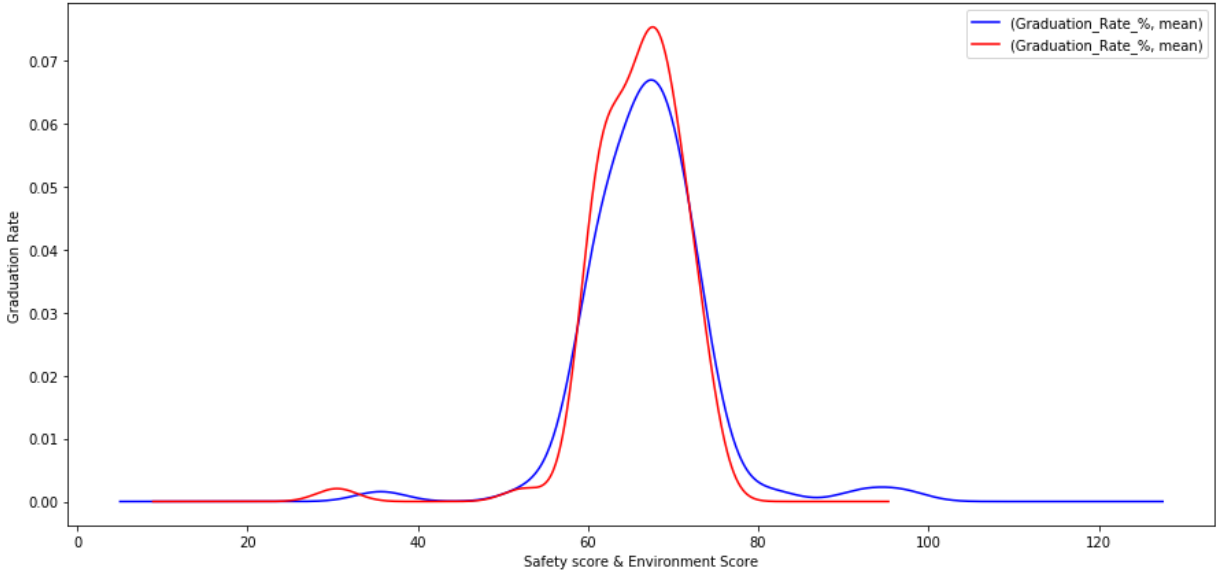
  there is a direct relation between CPS performance level and Graduation rate, like every school has a good level of CPS has good graduation rate at all.



- **Relation between safety score, environment score and graduation rate**

creating a safe and positive school environment is far more important than higher scores on standardized tests, according to a Berkeley IGS/Source poll.

when safety & environment score are between [50, 80], high graduation in average is gained so, safety and environment score has a direct impact on graduation rate of schools. Blue line shows safety score and red line shows environment score impact on performance.

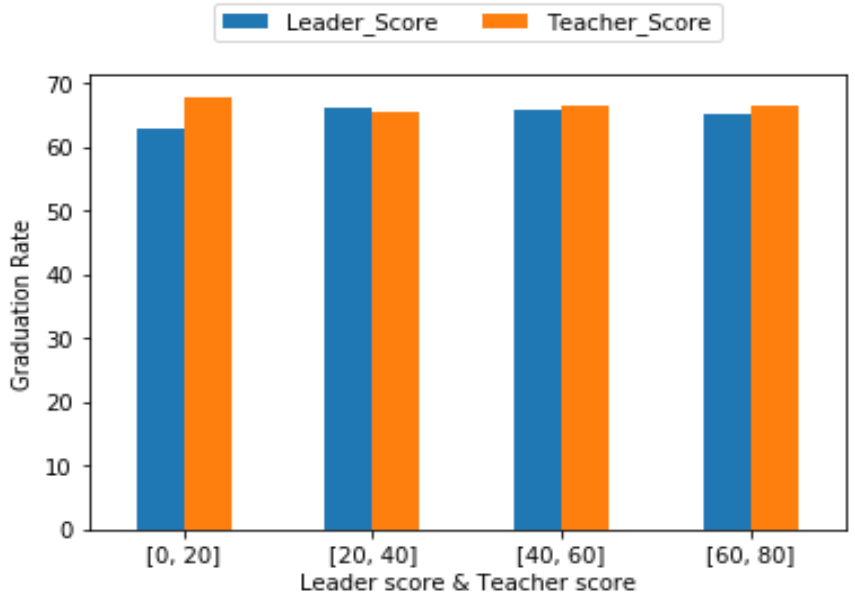


➢ Teacher score and leader score versus graduation rate

It is widely believed that a good principal is the key to a successful school.

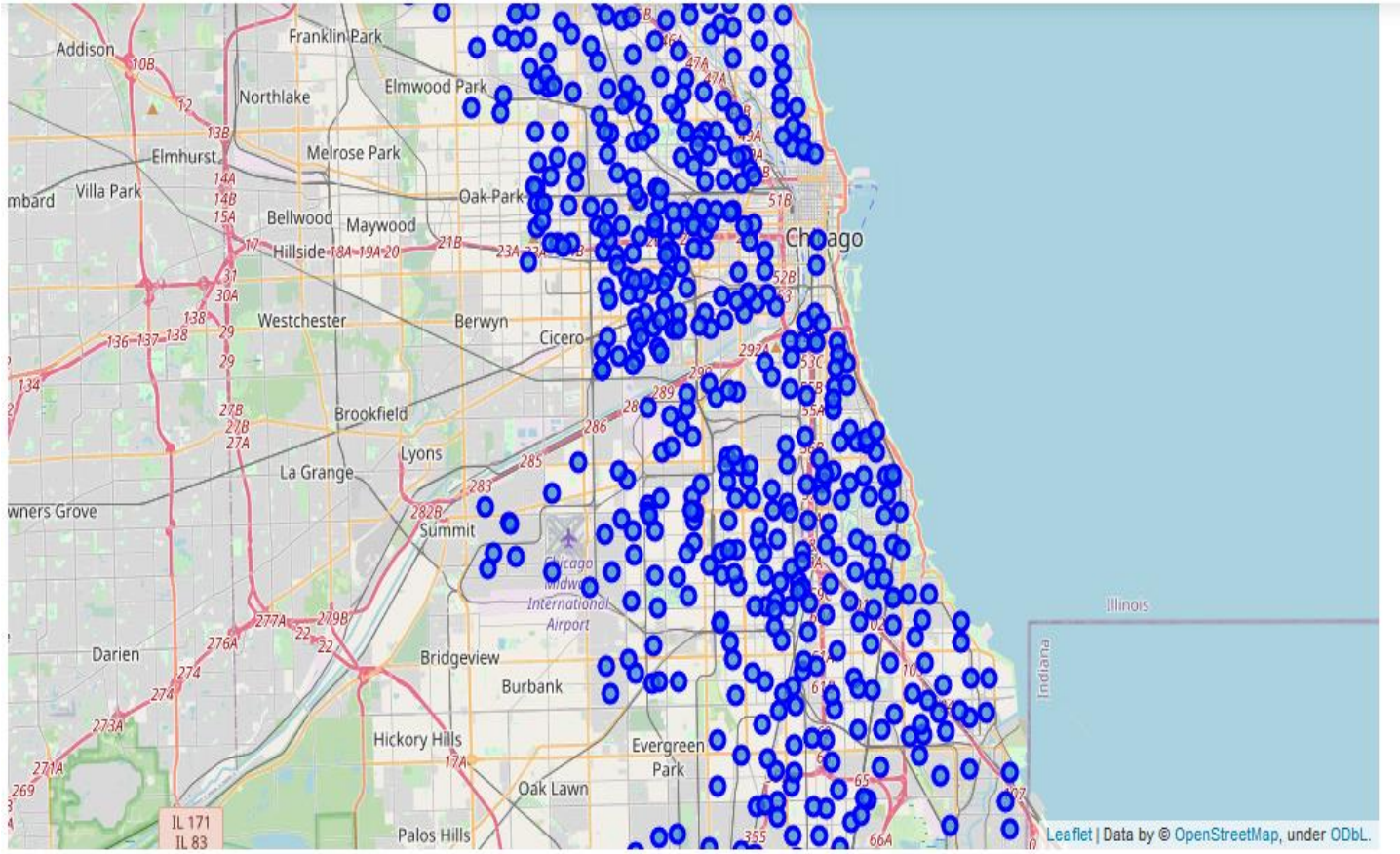
Students who go to schools where their teachers have a leadership role in decision making perform significantly better on state tests, new study finds.

There is highly effective principals and teachers raise the achievement of a typical student in their schools performance.



➢ Mapping schools location in Chicago city

In this section, location of each schools were plotted on map by their latitude and longitude.



New function proposing

- In this section I extend a function model to explicitly calculate score of each school using their attributes, each of which is either nominal or numerical. Numerical attributes is widely used to measure, the value of score for each school.
- For each link attribute, we can consider the *J*-polynomial function. Each element of which is calculated by function of u_j and v_j, $f\,(uj, vj)$.

$$f(u_j, v_j, x_j, \ldots) = \alpha u_j + \beta v_j + \gamma x_j + \cdots$$

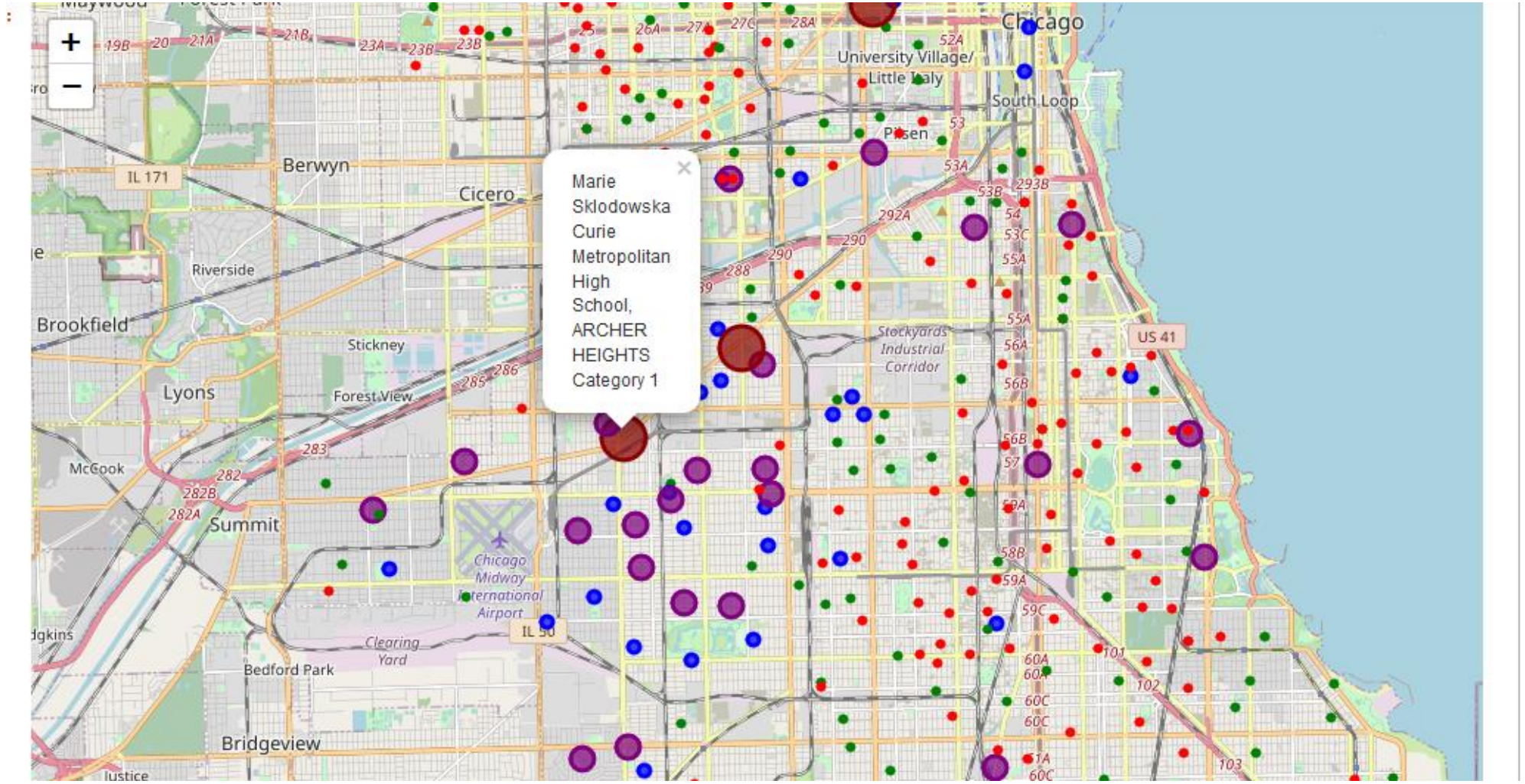
- So, according to this equation we have this function which calculate score using attributes:

F=CPS_Performance_level(0.28)+Environment_Score(0.172)+Leaders_Score(0.038)+Teachers_Score(0.044)+ Safety_Score*(0.177)+Graduation_rate*(0.64)+ *College_Enrollment**(0.55)

➢ Function results shown as Cat in table

ment_Score	Leaders_Score_	Teachers_Score	Rate_of_Misconducts_(per_100_students)_	COLLEGE_ENROLLMENT	Graduation_Rate_%	Score	Cat	Top_schools
74.0	65.0	70.0	2.0	813.0	73.0	529.391	3	True
74.0	63.0	76.0	16.0	521.0	73.0	361.014	4	False
50.0	50.0	49.0	2.0	1324.0	81.0	802.933	2	True
45.0	65.0	48.0	10.0	556.0	73.0	374.474	4	False
60.0	45.0	54.0	16.0	302.0	61.0	227.659	5	False

➢ Mapping result of proposing function on folium



school with highest rank is specified with larger circle

Clustering schools

K-means clustering method is used to categorized schools. Kmeans start using Euclidian distance to cluster each point of dataset into a one cumulative cluster.

K-means algorithm can be summarized as follow:

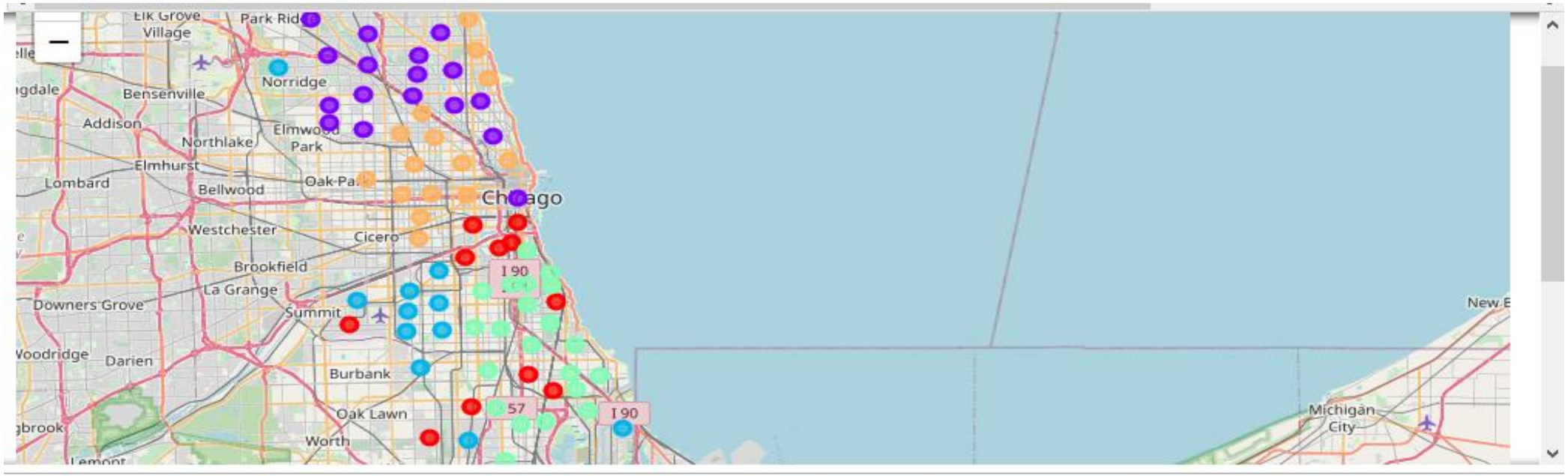
1. Specify the number of clusters (K) to be created (by the analyst)
2. Select randomly k objects from the data set as the initial cluster centers or means
3. Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
4. For each of the k clusters update the *cluster centroid* by calculating the new mean values of all the data points in the cluster. The centoid of a K_{th} cluster is a vector of length p containing the means of all variables for the observations in the k_{th} cluster; p is the number of variables.
5. Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations.

- Labeled dataset using cluster method

AFETY_SCORE	Environment_Score	Rate_of_Misconducts_(per_100_students)_	COLLEGE_ENROLLMENT	Latitude	Longitude	COMMUNITY_AREA_NUMBER	Labels
59.571741	52.351411	12.000000	858.000	41.968518	-87.717327	14.0	1
45.500000	37.500000	9.500000	2411.500	41.804285	-87.723913	57.0	2
43.333333	49.000000	5.666667	486.000	41.840676	-87.633966	34.0	0
45.000000	36.125000	24.750000	810.375	41.745201	-87.715027	70.0	2
34.057393	39.481128	30.600000	417.500	41.743401	-87.653819	71.0	3

- Every community's schools is clustered and shown by label, label attributes are Shown as last attribute created by Kmeans Clustering method

- folium map of clustering by location of schools in Chicago city



Conclusions

cluster with label 1(circles with color of purple on map) and respectively, cluster with label 2 (which are ordinary blue colors) are better schools, and label 4 contains good schools (which have been shown with orange color points), then cluster 3 with bright sky green color and the last and not very good are cluster 0 with red colors in Chicago city which are categorized by clustering algorithm. Clustering results shows function is optimally labeled schools based on correlation of each features. Popup info of folium map is available in Jupyter notebook loaded on Github account.