# Segmenting and analyzing Chicago Schools

Aylin Mousavian

August 11, 2020

## 1. 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. State and national data systems cannot always provide the necessary data to evaluate schools on readiness. Many states are prioritizing efforts to link K–6 with postsecondary and state data systems in order to follow students from elementary school through high school and college or employment. In addition to state data sources, college data may be accessed from testing agencies such as ACT and the College Board, and from the National Student Clearinghouse, which calculates college attendance rates. In this section we focused on data from city of Chicago, to revise which community is the top and with communities are better for great schools. At data acquisition we prepare and clean data at first part, at the second part I will analyze it, at third part I called it methodology, I make a function as a weighted function to score each school based on influential features and rate school as mentioned score. Then we prioritize schools according to their gained score, at the final we cluster each community based on their features and we will end it with result and conclusion.

## 2. 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 to feel free to use it. Might be this dataset downloaded or scraped from multiple sources were combined into one table and because of these reasons, There were a lot of missing values from earlier seasons. 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. Base on definition of our problem, 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.

After analyzing, clustering has been done on schools to compare the results and see which communities are good to buy home based on schools ranking. Following data sources will be needed to extract/generate the required information:

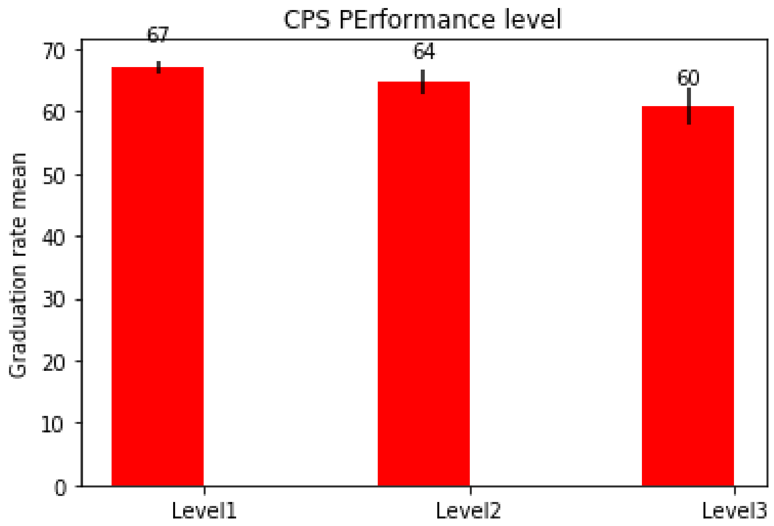
- ✓ Coefficients of every feature which make effect on score value.
- ✓ Apply function to measure score and rank the school.

## 3. Exploratory Data Analysis

One of the most visible components school accountability was the strengthening of Adequate Yearly Progress (AYP) requirements. In order to measure of the progress of all students, including major subgroups of students, towards reaching state proficiency standards is measuring student progress in term of graduation and college enrollment. Due to the statewide AYP public reporting requirements, and the consequences of missing AYP targets in the NCLB accountability system, AYP quickly became a *de facto* measure of public school quality. In this project AYP is calculated by correlation of each features with student progress as graduation and college enrollment.

### 3.1. 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. The CPS School Quality Rating Policy (SQRP) is the district's policy for measuring annual school performance. CPS provide a framework for goal-setting for schools, Identify schools in need of targeted or intensive support. CPS category level contains 3 level. Top level is 1, level 2 and level 3. Level 1 is highest performance, this type of school is a nationally competitive school with the opportunity to share best practices with others. Level 2 is average performance, additional support from the network team is needed to implement interventions and level 3 is below average performance. These schools status requires increased support from the network. In this section correlation between schools performance is revised and visualized, all three level schools have more than 50% percentage graduation, but the highest graduation belongs to schools with highest CPS performance policy.

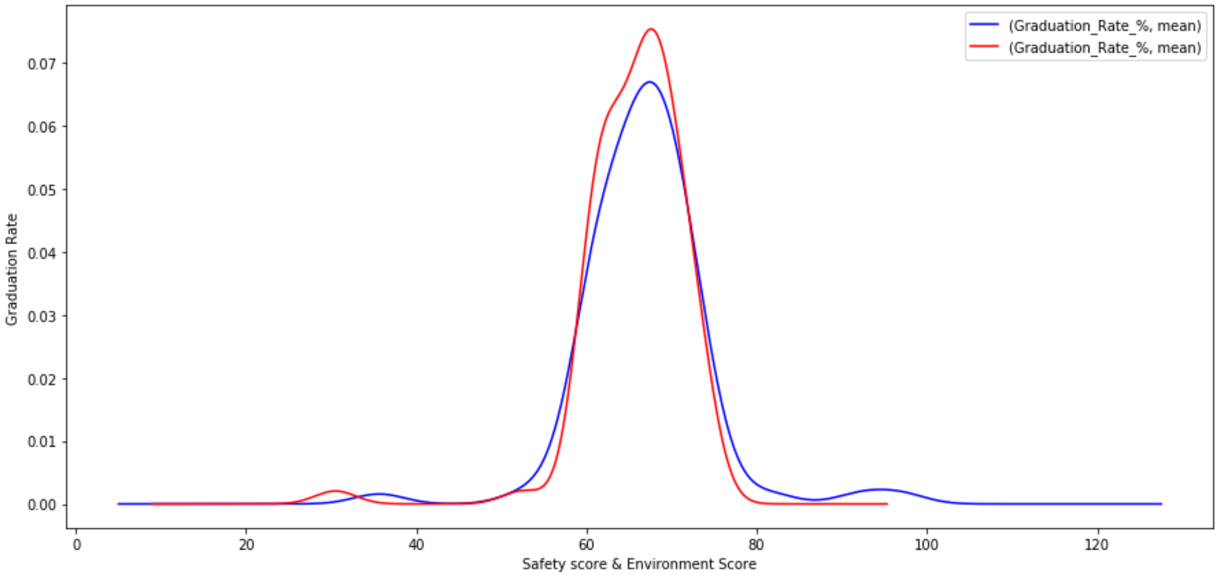


Relation between CPS performance level and Graduation levels of schools

As you can see, 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.

### 3.2 Relation between safety score, environment score and graduation rate

In evaluating school performance, registered voters say creating a safe and positive school environment is far more important than higher scores on standardized tests, according to a Berkeley IGS/Source poll. Voters also express considerable concerns about bullying, school fights and other forms of intimidation or violence on school campuses, along with harassment that students experience through social media. Therefore having schools with high safety and environment score and protecting students make them to focus on their aims and result progress in their fields. Poor indoor environments have been associated with a variety of health symptoms and a decline in student performance in reasoning, typing, and math. Several studies have found that health, attendance, and academic performance improve with increased maintenance of school facilities. This part of studies is dedicated to revise safety and environment with student performance on graduation rate. KDE type chart shows significant relation between these features, especially when scores of safe and environment getting high after 50, percentage of graduation get significantly high. It means high score safe and health schools have great impact on performance. The results is on average value of performance among all schools and grouped by environment and safety score.

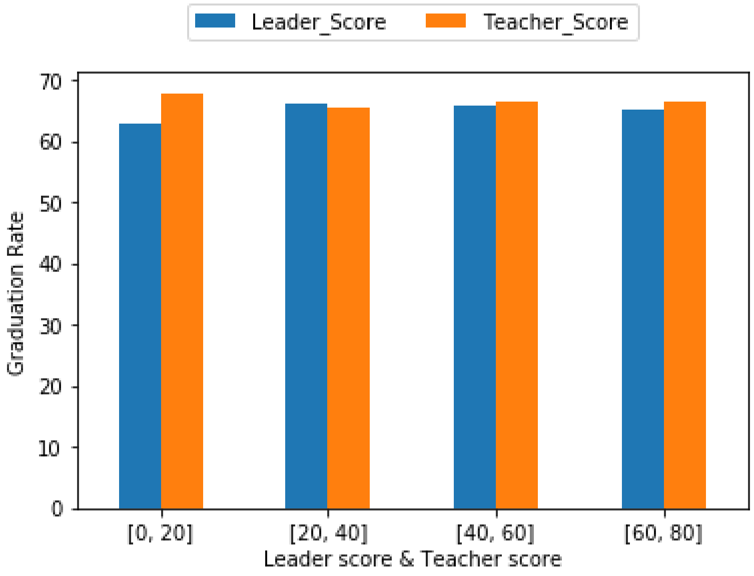


Relation between safety score and environment score with graduation percentage

As you can see 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.

### 3.3 Teacher score and leader score versus graduation rate

Students who go to schools where their teachers have a leadership role in decision making perform significantly better on state tests, new study finds. But some of the leadership elements that are most related to student achievement are the ones that are least often implemented in schools. According to new analysis, from the New Teacher Center's Teaching, Empowering, Leading, and Learning survey, which asks questions about teaching, learning, and working conditions in schools. Richard Ingersoll, a professor of education and sociology at the University Of Pennsylvania Graduate School Of Education and the report's lead author, studied responses from 2011 to 2015, which included data from nearly 1 million teachers from more than 25,000 schools, in 16 states. Schools with the highest levels of instructional and teacher leadership rank at least 10 percentile points higher in both math and English/language arts on state tests, compared to schools with the lowest levels, even after controlling for factors like school poverty, size, and location. Teachers are closest to students, so they know what students need to improve. It is widely believed that a good principal is the key to a successful school. No Child Left Behind encouraged the replacement of the principal in persistently low-performing schools, and administration has made this a requirement for schools undergoing federally funded turnarounds. Foundations have invested millions over the past decade in New Leaders for New Schools, an organization that recruits nontraditional principal candidates and prepares them for the challenges of school leadership. Studies provides new evidence on the importance of school leadership by estimating individual principals' contributions to growth in student achievement. This approach is quite similar to studies that measure leader's role in performance of schools and student. This section results indicate that highly effective principals and teachers raise the achievement of a typical student in their schools performance.

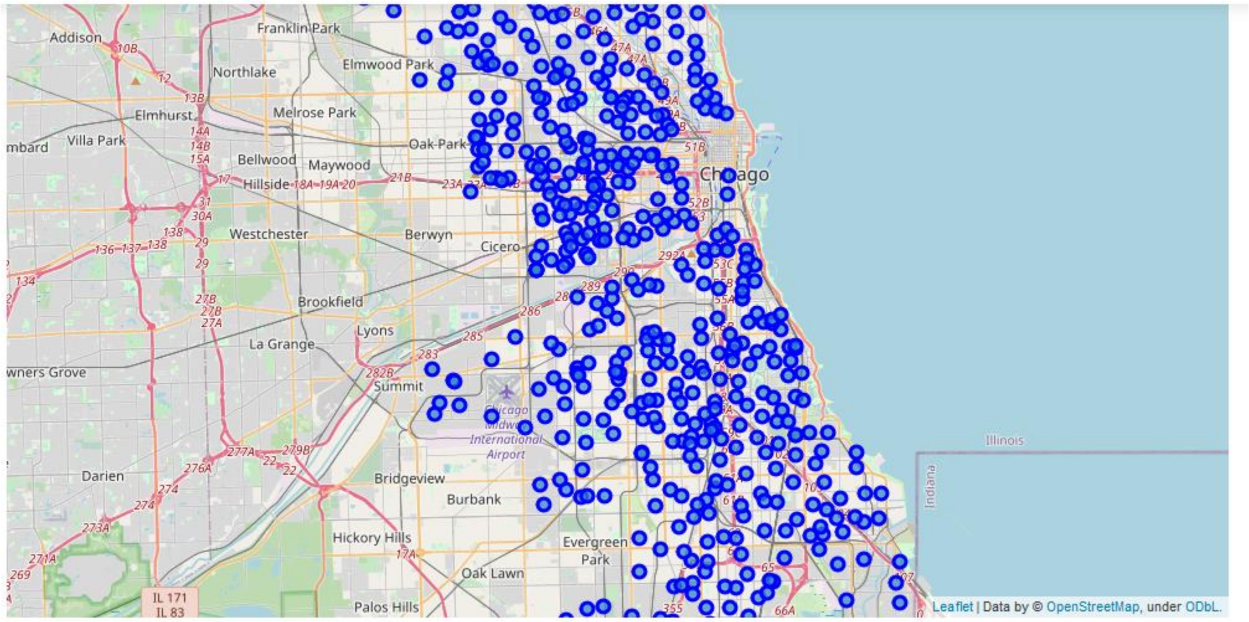


Teachers and leaders versus students' performance.

Graduation rate in this chart is calculated in average value of schools performance. It indicate great role of teachers and leaders on students achievement in a year of learning school.

### 3.4 Mapping schools location in Chicago city

In this section, location of each schools were plotted on map by their latitude and longitude. After getting latitude and longitude of Chicago city using geolocater of geopy library, mapping of schools have been specified by blue spots. This map plotting has been gained by folium library of python visualizing libraries. Folium makes it easy to visualize data that's been manipulated in python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/html visualizations as markers on the map. It is so fun to use it. Try by yourself to map some locations with this folium library.



Mapping schools of Chicago, using latitude and longitude of each school in folium map.

## 4. New function proposing

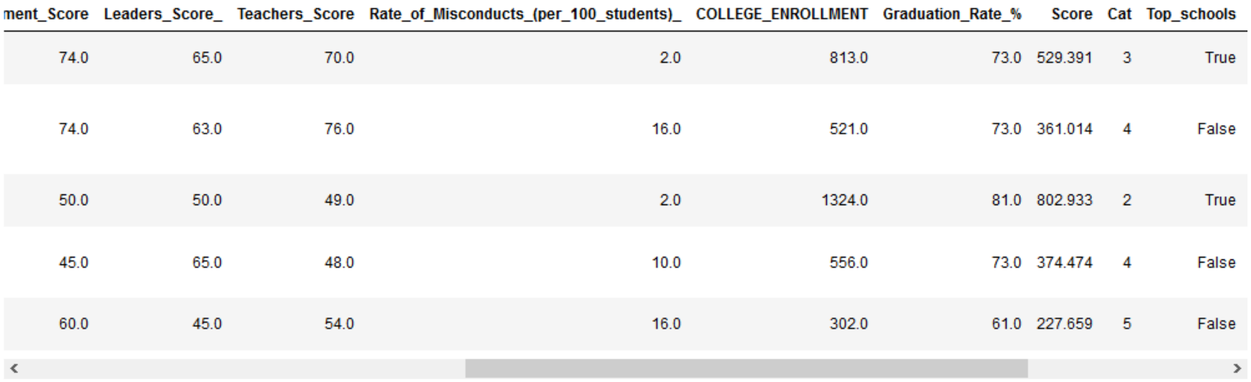
In this section I extend a function model to explicitly calculate score of each school using their attributes, each of which is either nominal or numerical. Numerical attributes is widely used to measure, the value of score for each school. Let $v_j$ be a value that attribute $v$ takes correlation value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

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

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)

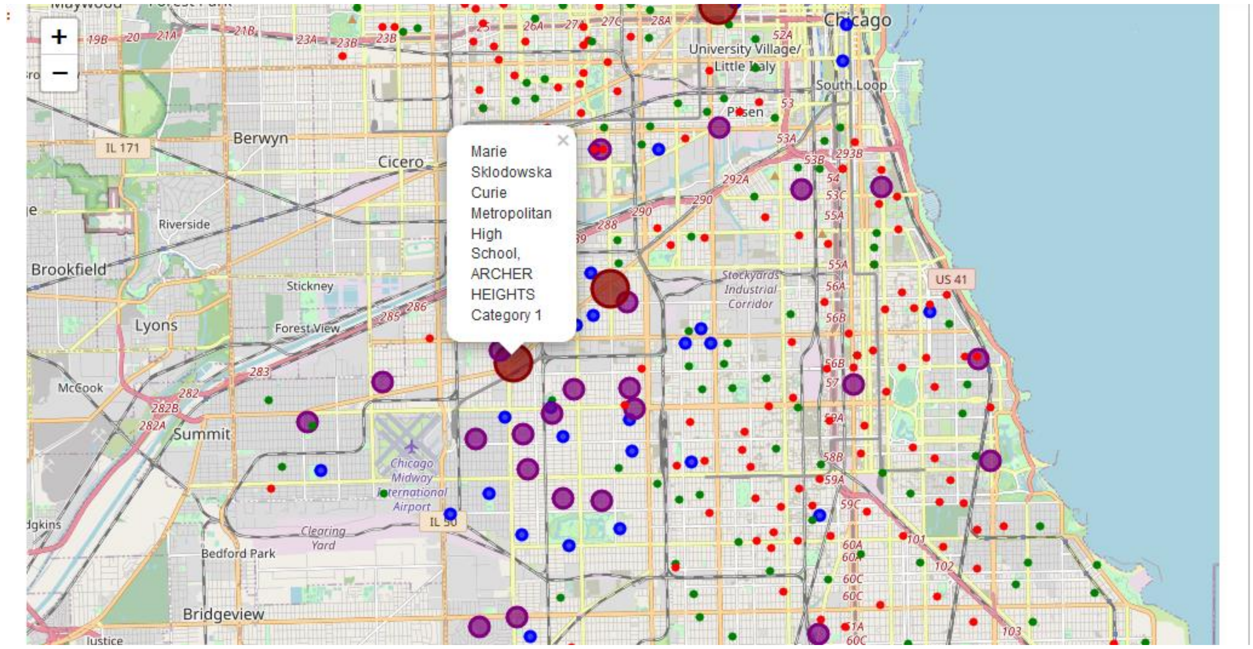
F function considers all significant attributes with their correlation to generate a number as score, this number based on value level specify a 5 scale level, in which level 1 as highest score school and level 5 as lowest level school. After using this function and assigned to each school, let’s map every school based on their latitude and longitude and also their category level associated with number in folium library. As it is shown on table below, Cat attribute determine category or level of each school among 566 schools in mentioned dataset.



| 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 |

Table shows category of school and score based on function value

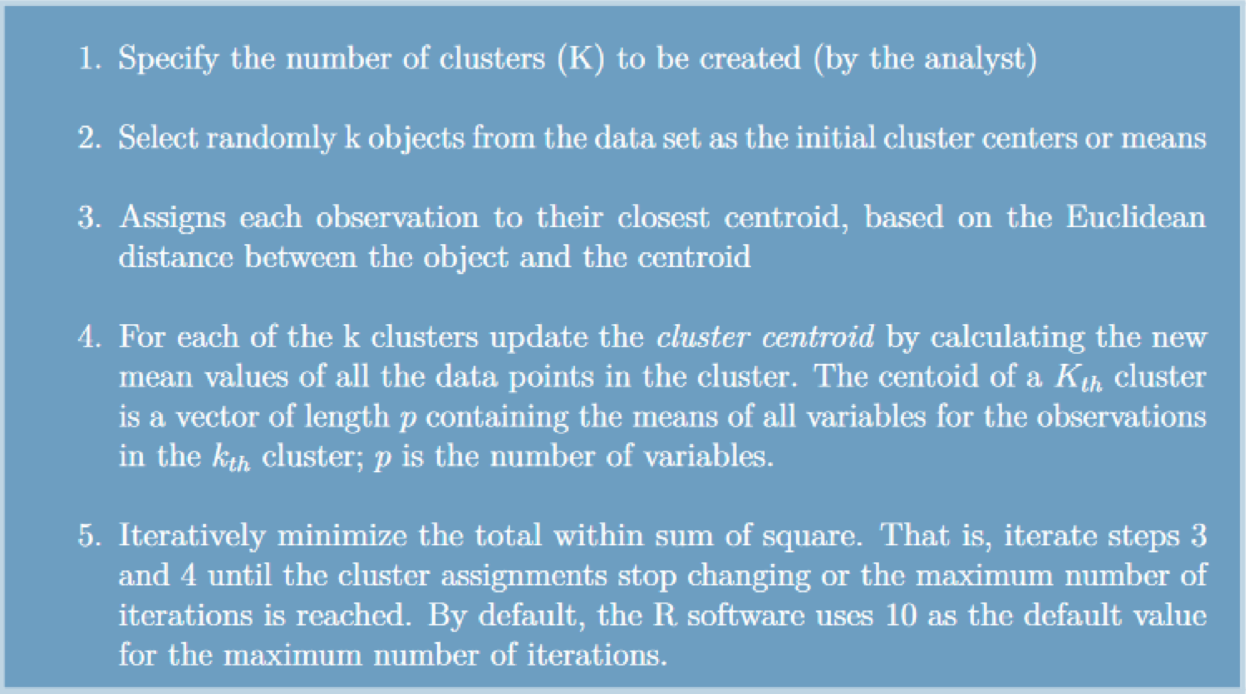
To experimentally and visually evaluate function results, let's take a look at folium map of schools with mentioned category. In this map, every school with highest rank is specified with large brown circle, second level schools as category 2 is shown by purple circle and as follow third category schools is shown with circles colored purple, category 4 schools is determined as green circle and finally category 5 schools is shown by red tiny circles.



Folium mapping of categorized schools of Chicago city

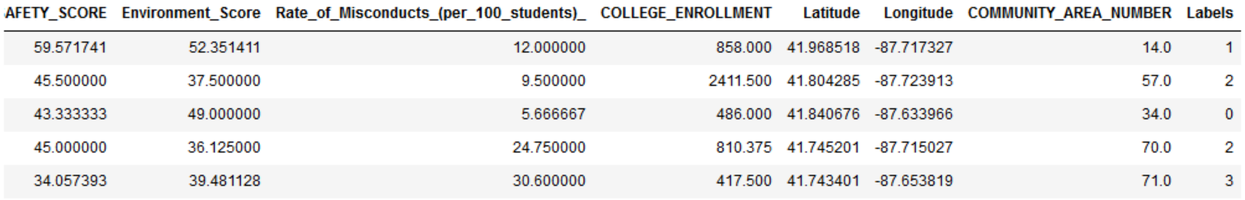
## 5. Clustering schools

So far we have only looked at function categorized method, but a cluster can also be generated to categorize schools. This type of clustering is called K-means clustering. Kmeans start using Euclidian distance to cluster each point of dataset into a one cumulative cluster. This procedure is applied recursively until each point of dataset is in its own singleton cluster. Each cluster is represented by the center or means of the data points belonging to the cluster. The K-means method is sensitive to anomalous data points and outliers. 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.

Before clustering schools, dataset has been grouped by their community_area_name. In fact clustering algorithm is applied to school dataset without any categorizing application before. This is due to compare results of categorized and clustered to see how these two methods work. Table below shows labels of schools which has been done by clustering algorithm.

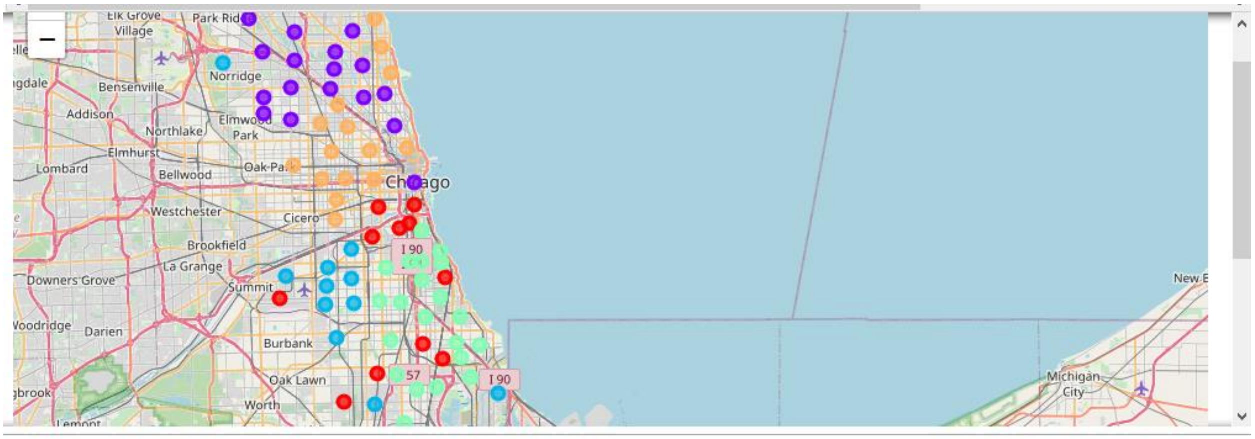


| AFETY_SCORE | Environment_Score | Rate_of_Misconducts_(per_100_students)_ | COLLEGE_ENROLLMENT | Latitude | Longitude | COMMUNITY_AREA_NUMBER | Labels |
|---|---|---|---|---|---|---|---|
| 59.571741 | 52.351411 | 12.000000 | 858.000 | 41.968518 | -87.717327 | 14.0 | 1 |
| 45.500000 | 37.500000 | 9.500000 | 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 |

Clustering methods has been shown by label attribute

Every community's schools is clustered and shown by label. Let's see the folium map of clustering by location of schools in Chicago city. Every cluster is shown with different color to determine each cluster.

Clustered result of each community's school in Chicago



## 6. Conclusions

In this study, schools of Chicago city analyzed the relationship between their attributes and performance and biographic data. Clustering results shows function is optimally labeled schools based on correlation of each features. If you notice, 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. So if parents have intent to buy house or reach good community to settle, it's better to invest both on money and future of their children to buy their own house at community for example ARCHER HEIGHT or other orange, purple or blue color communities. Function results which categorized schools shows, communities beside airplane symbol contain good schools like clustering results. It means analyzing and new proposed methodology works as better as clustering method.