Applied Data Science Capstone

IBM Data Science Professional Certificate

Week 5 Submission

Project Name:

In Search of a Safe place to Stay with Family in Chicago, IL

(The Battle of Neighborhoods)

Chicago

Chicago, officially the City of Chicago, is the most populous city in the U.S. state of Illinois, and the third-most-populous city in the United States. With an estimated population of 2,693,976 in 2019, it is also the most populous city in the Midwestern United States. Located on the shores of freshwater Lake Michigan, Chicago was incorporated as a city in 1837 near a portage between the Great Lakes and the Mississippi River watershed. Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. Chicago is home to several Fortune 500 companies, including Allstate, Boeing, Caterpillar, Exelon, Kraft Heinz, McDonald's, Mondelez International, Sears, United Airlines Holdings, US Foods, and Walgreens.

The Problem

The city's overall crime rate, especially the violent crime rate, is higher than the US average. Chicago was responsible for nearly half of 2016's increase in homicides in the US, though the nation's crime rates remain near historic lows. The reasons for the higher numbers in Chicago remain unclear. Crime in Chicago has been tracked by the Chicago Police Department's Bureau of Records since the beginning of the 20th century.

In this situation, a family with a school-going kid, needs to move to Chicago, and **find out few safe Community Areas to live with good amenities in the Neighborhoods**. The project aims to analyze the Community Areas of Chicago and suggest appropriate Community Area for Safe living. The project uses 'K-Means Clustering', which is an unsupervised Machine Learning technique of Data Science, to segment all 77 community areas of Chicago, IL to prepare five clusters as per neighborhood venue categories.

The project aims to do :

a. Importing Datasets
b. Cleaning the Data
c. Data frame manipulation
d. Web-scrapping
e. Summarizing the Data
f. Visualization
g. Gathering GeoLocation using geopy
h. Remote call to FourSquare.com to get neighborhood-list
i. Clustering

2. **Data Acquisition and Cleaning**

2.1 Data Sources:

a. Chicago_Crimes_2012_to_2017
   Data Source: Kaggle
   No of Rows: 335670

b. Chicago_Public_Schools_-_Progress_Report_Cards__2011-2012
   Data Source: IBM
   No of Rows: 566

c. Census_Data_Selected_socioeconomic_indicators_Chicago_008_2012
   Data Source: IBM
   No of Rows: 78

d. List of Community Areas in Chicago
   https://en.wikipedia.org/wiki/Community_areas_in_Chicago
   Table containing Community Number and Community Name

e. FourSquare.com
   https://api.foursquare.com
   API Calls to get Neighborhood Venue details

Below Details have been considered from the Datasets:

1. Crime Data
   ID : Crime ID - unique
   Primary Type : Crime Types
   Arrest : YES or NO
   Community Area : Community Area Number
   Year : year of Crime

2. School Data:
   School ID : Unique
   COMMUNITY_AREA_NUMBER
   SAFETY SCORE

3. Census Data:
   COMMUNITY_AREA_NUMBER : Unique
   PERCENT HOUSEHOLDS BELOW POVERTY
   PER_CAPITA_INCOME

A. Data Cleaning:

1. Inspect the Crime data
    a. Unnecessary cols dropped
    b. Zero and Null Values to be managed

2. By Web-Scraping confirmed the correct Community Name as per the Community Number from Wikipedia page

B. Feature Selection

a. Considered Crime Data with Arrest = YES for Year = 2012
b. Considered Average School Safety Score for each Community
c. Considered Percent Households Below Poverty Level for each Community
d. Considered Per Capita Income for each Community

C. Run Correlation, for each Community Area, among:

- Total Number of Crimes
- Average School Safety Score
- Percent Households Below Poverty Level
- Per Capita Income for each Community

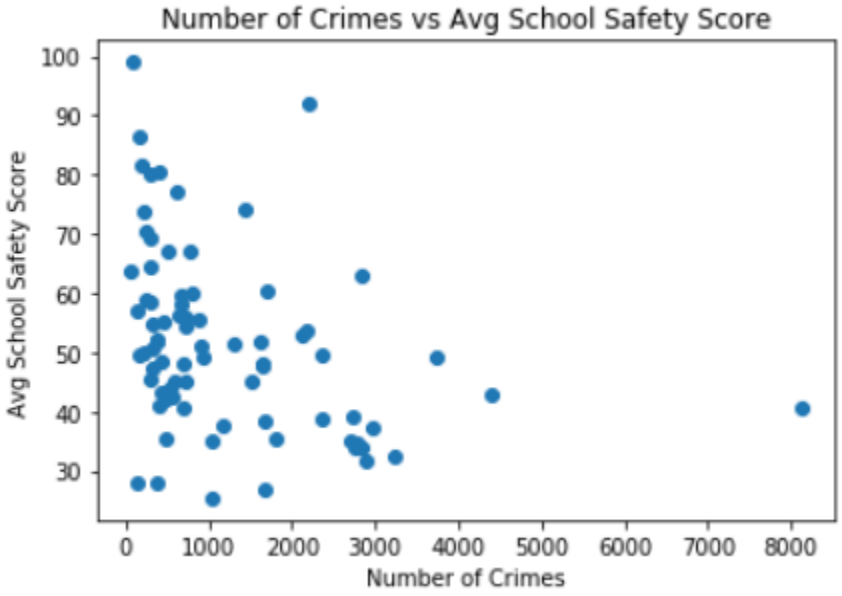
D. Identify the Safest Communities with above features

E. Send calls to Foursquare.com and found out 1463 venues across 77 Community Areas

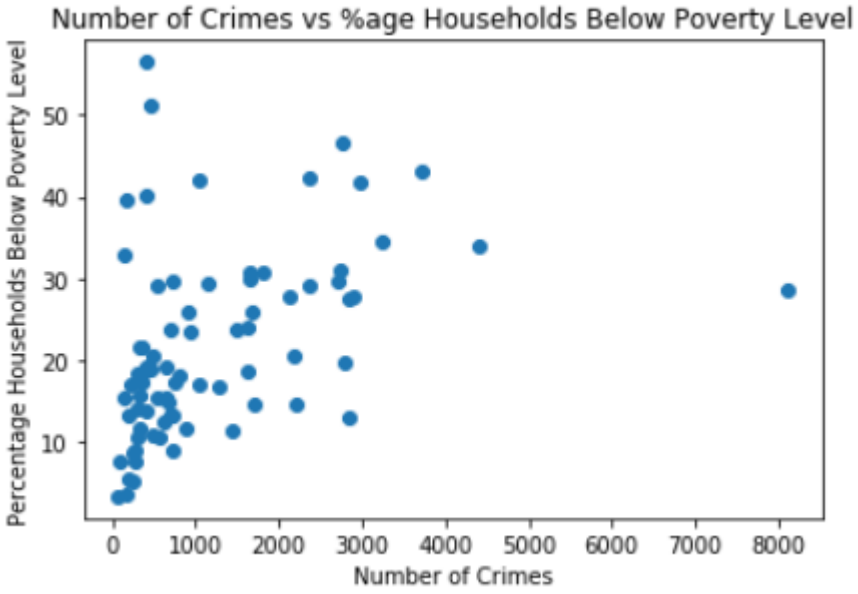
F. Classified Five Clusters using K-Means, which gives clustered Neighborhoods of similar Venues in Communities in Chicago

Data Visualization

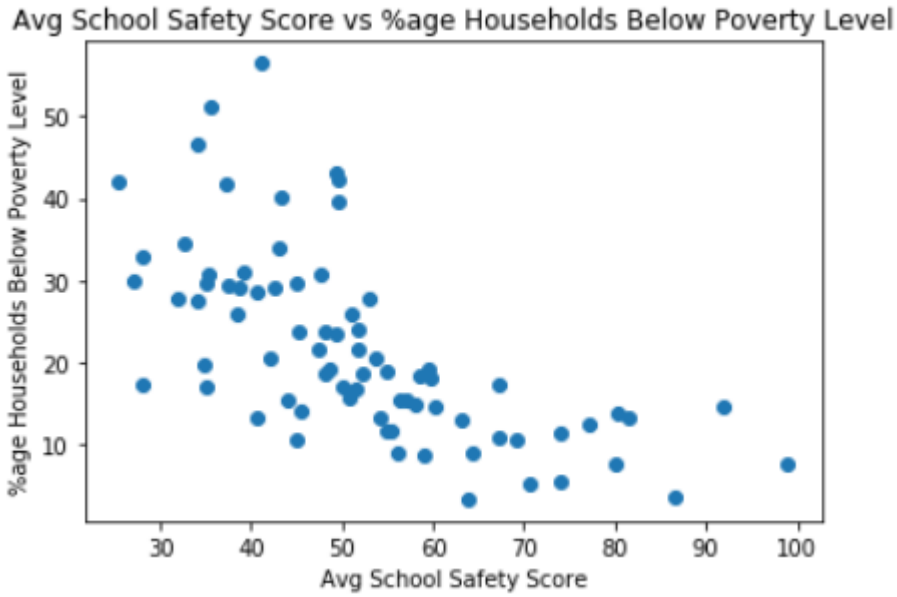
1. Relationship between (x) Total Number of Crimes and (y) Average School Safety Score



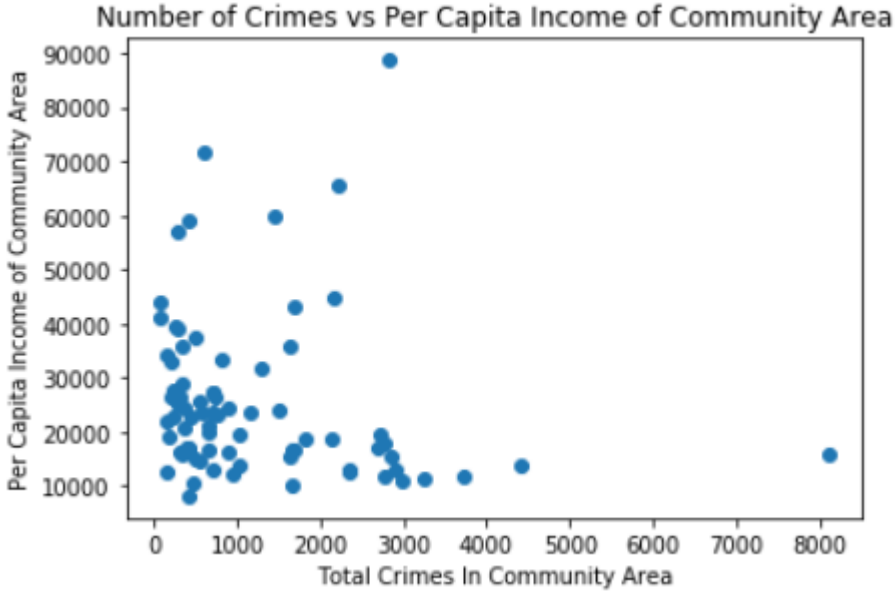
2. Relationship between (x) Total Number of Crimes and (y) Percent Households Below Poverty Level



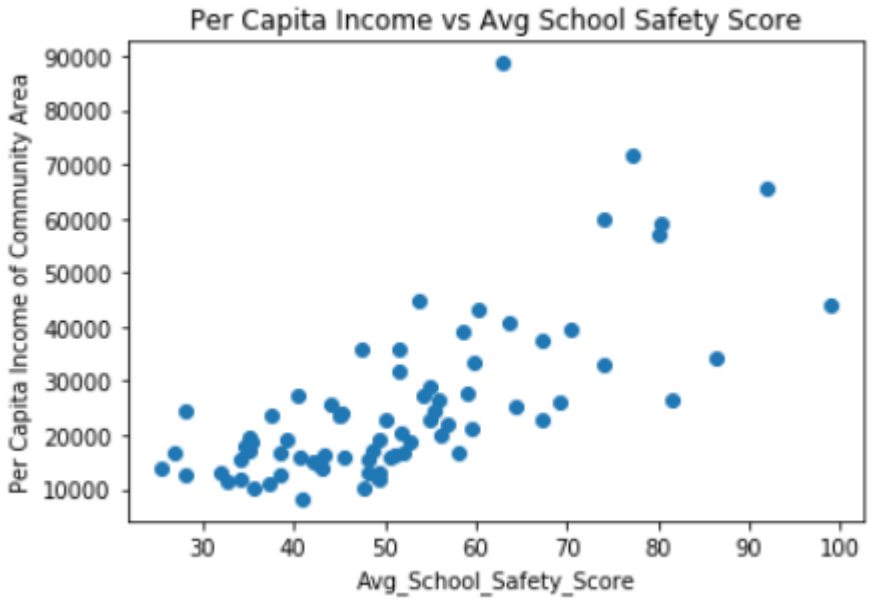
3. Relationship between (x) Average School Safety Score and (y) Percent Households Below Poverty Level



4. Relationship between (x) Total Crimes In Community Area and (y) Per Capita Income



5. Relationship between (x) Average School Safety Score and (y) Per Capita Income



Inferences:

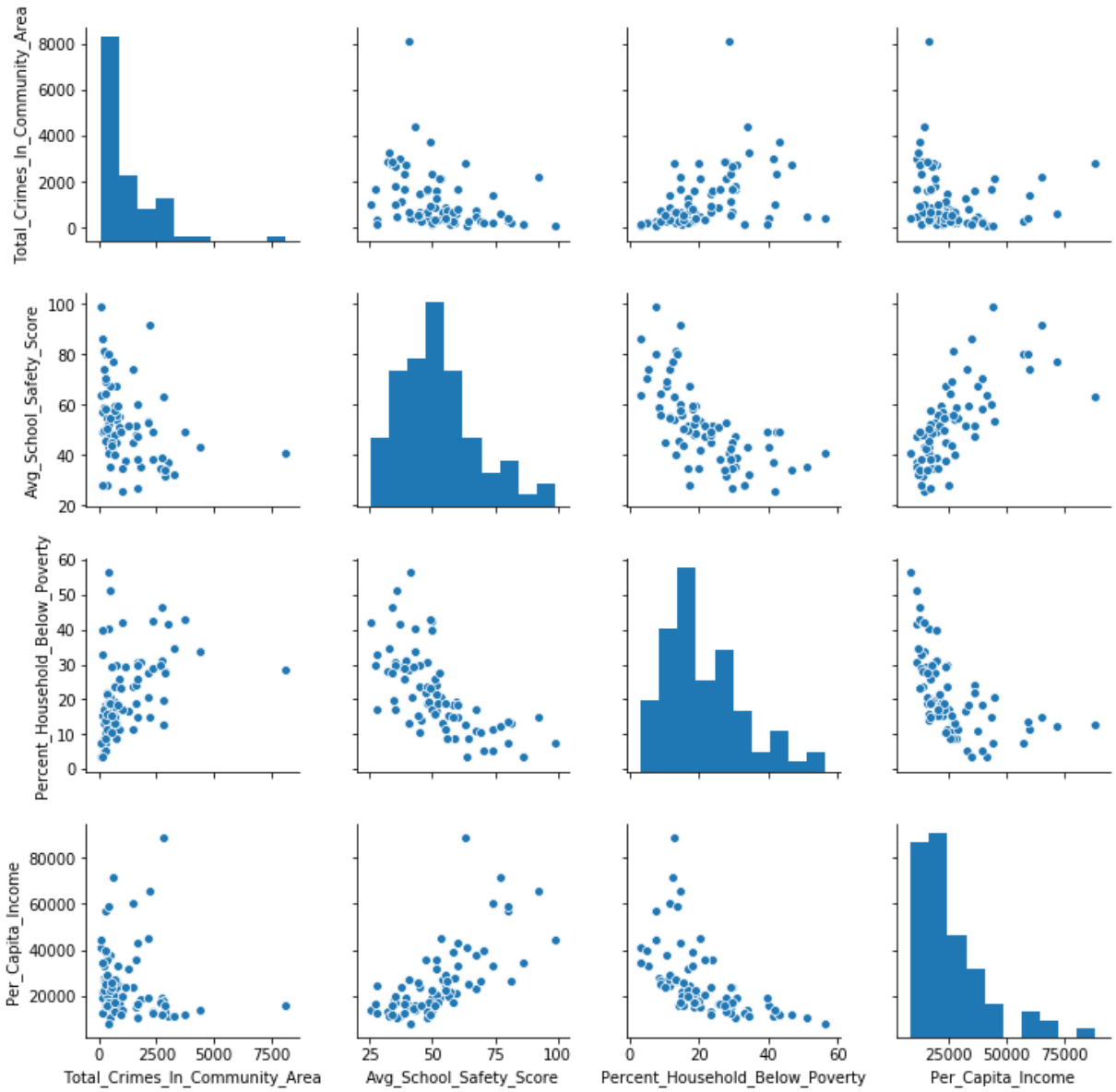
A. Number of Crimes' has Negative & Weak (0.33) Correlation with 'Average School Safety Score' of a Community Area

B. 'Number of Crimes' has Positive & Medium (0.40) Correlation with 'Percent Household below poverty' of a Community Area

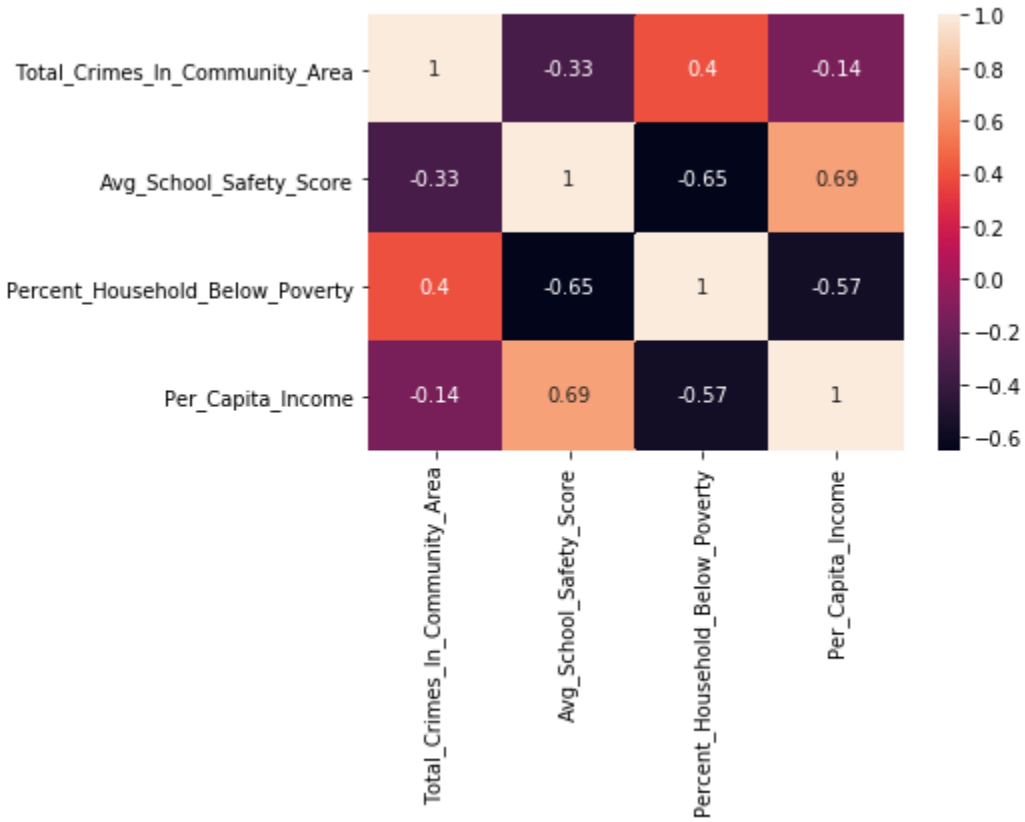
C. 'Average School Safety Score' has positive (0.69) correlation with 'Per Capita Income' of a Community Area

D. 'Average School Safety Score' has negative (0.65) correlation with 'Percent Household below poverty' of a Community Area

6. Pair-Plot


Total_Crimes_In_Community_Area
Avg_School_Safety_Score
Percent_Household_Below_Poverty
Per_Capita_Income
0
2000
4000
6000
8000
20
40
60
80
100
0
10
20
30
40
50
60
20000
40000
60000
80000
0
2500
5000
7500
25
50
75
100
0
20
40
60
25000
50000
75000

7. Heat-Map showing the Correlation among:
    - Total Number of Crimes
    - Average School Safety Score
    - Percent Households Below Poverty Level
    - Per Capita Income for each Community



8. **Inferences**:

E. Number of Crimes' has Negative & Weak (0.33) Correlation with 'Average School Safety Score' of a Community Area

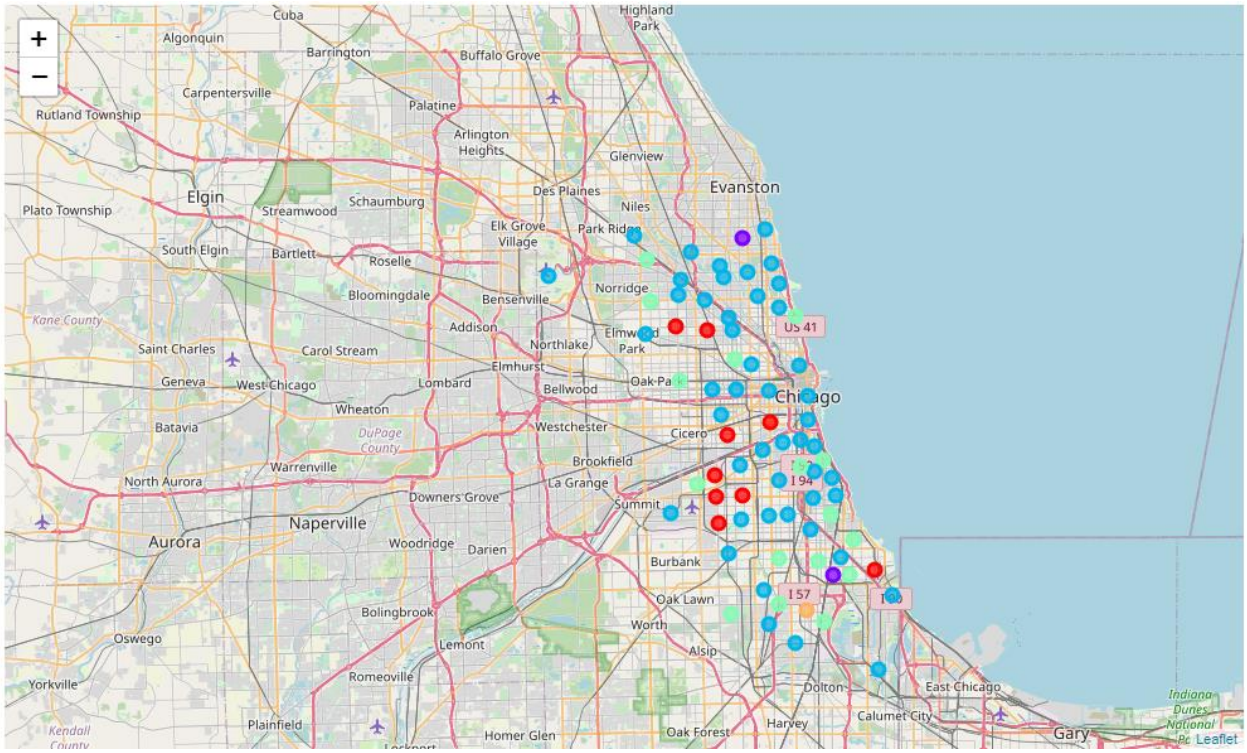
F. 'Number of Crimes' has Positive & Medium (0.40) Correlation with 'Percent Household below poverty' of a Community Area

G. 'Average School Safety Score' has positive (0.69) correlation with 'Per Capita Income' of a Community Area

H. 'Average School Safety Score' has negative (0.65) correlation with 'Percent Household below poverty' of a Community Area

Classification

I have used K-Means to prepare clusters with similar neighborhood venues. K-Means Clustering is used for unsupervised machine learning based on the intra-cluster similarity. There are various types of clustering algorithms such as partitioning hierarchical or density based clustering. K-means is a type of partitioning clustering that divides the data into K non-overlapping subsets or clusters without any cluster internal structure or labels, thus, it's an unsupervised algorithm to ensure that the objects within a cluster are very similar and objects across different clusters are very different or dissimilar.



Clusters:

With K-Means Clustering, below Five Clusters, with segmented Communities were identified

- Cluster 1: 09 communities
- Cluster 2: 03 communities
- Cluster 3: 47 communities
- Cluster 4: 16 communities
- Cluster 5: 01 community

Result and Conclusion

The objective of this project has been to help any family (with a school going kid) who would like to relocate to Chicago. The project will help to find out low crime & safe Cluster Areas, and also surrounding amenities and venues.

The Cluster 3 which has highest number of Community Areas, has the most appropriate Community Area (Forest Glen), in terms of (a) low crime, (b) good school safety score and (c) low Percentage of Household Below Poverty Level