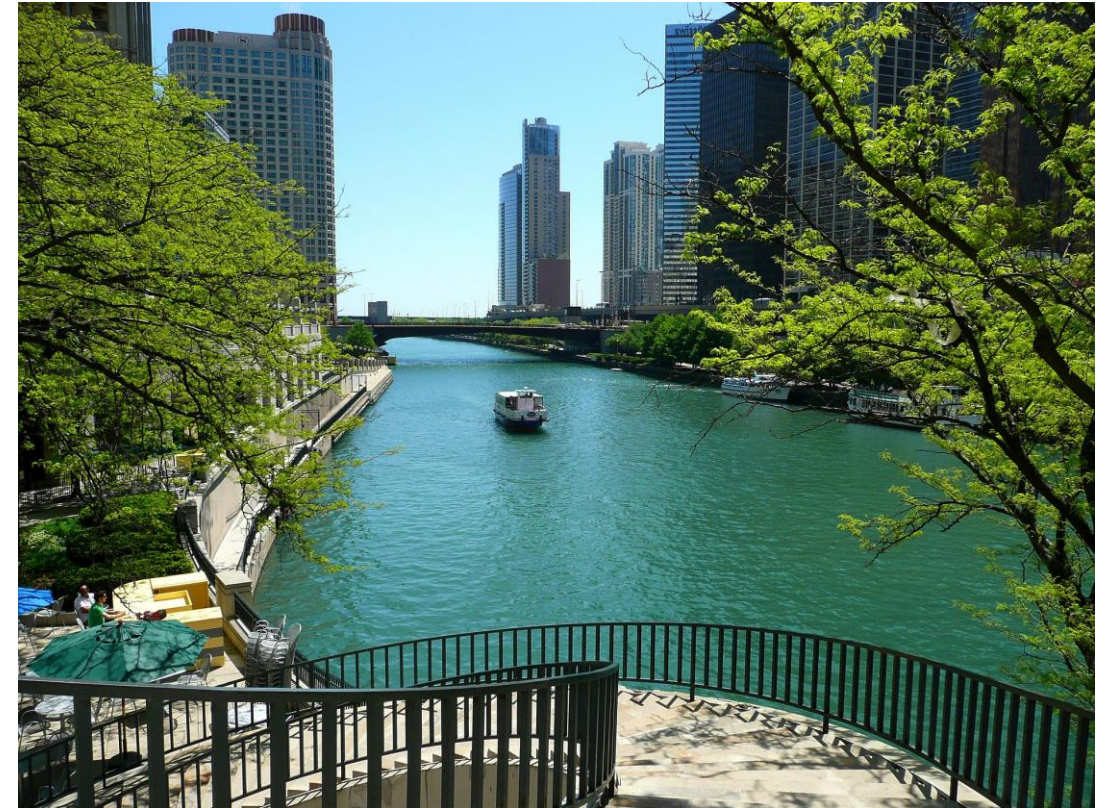


Applied Data Science Capstone
IBM Data Science Professional Certificate

The Battle of Neighborhoods



**In Search for a
safe place
to live in Chicago**



Monishankar Hazra

monishankar@gmail.com
linkedin.com/in/monishankar

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.



In Search for a
safe place
to live in Chicago



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.



In Search for a safe place to live in Chicago

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

1. Importing Datasets
2. Cleaning the Data
3. Data frame manipulation
4. Web-scrapping
5. Summarizing the Data
6. Visualization
7. Gathering GeoLocation using Geopy
8. Remote API call to FourSquare.com to get neighborhood-list
9. Clustering



Data Acquisition

Chicago Crimes

Data Source: Kaggle

No of Rows: 335670

Chicago Public Schools Progress Report Cards

Data Source: IBM

No of Rows: 566

Census Data Selected socioeconomic indicators

Chicago Data Source: IBM

No of Rows: 78

List of Community Areas in Chicago

https://en.wikipedia.org/wiki/Community_areas_in_Chicago

Table containing Community Number and Community Name

FourSquare.com

<https://api.foursquare.com>

API Calls to get Neighborhood Venue details

Crime Data:

Crime ID

Primary Type : Crime Types

Arrest : YES or NO

Community Area Number

Year : year of Crime

School Data:

School ID

COMMUNITY_AREA_NUMBER

SAFETY SCORE

Census Data:

COMMUNITY_AREA_NUMBER

PERCENT HOUSEHOLDS BELOW

POVERTY

PER_CAPITA_INCOME



A. Data Cleaning

1. Inspect the Crime data
2. Unnecessary cols dropped
3. Zero and Null Values to be managed
4. By Web-Scraping confirmed the correct Community Name as per the Community Number from Wikipedia page

B. Feature Selection

1. Considered Crime Data with Arrest = YES for Year = 2012
2. Considered Average School Safety Score for each Community
3. Considered Percent Households Below Poverty Level for each Community
4. Considered Per Capita Income for each Community

C. Run Correlation

1. Total Number of Crimes
2. Average School Safety Score
3. Percent Households Below Poverty Level
4. Per Capita Income for each Community

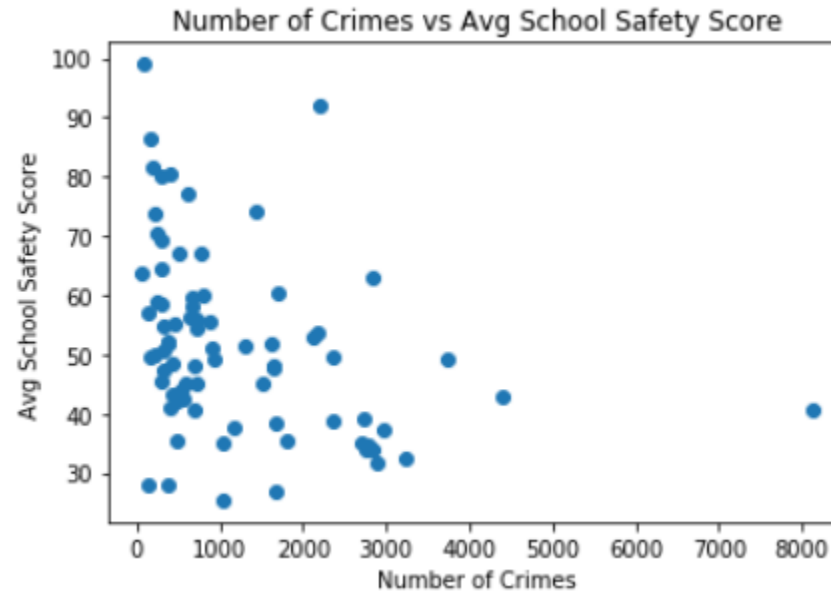
D. Identify the Safest Communities

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

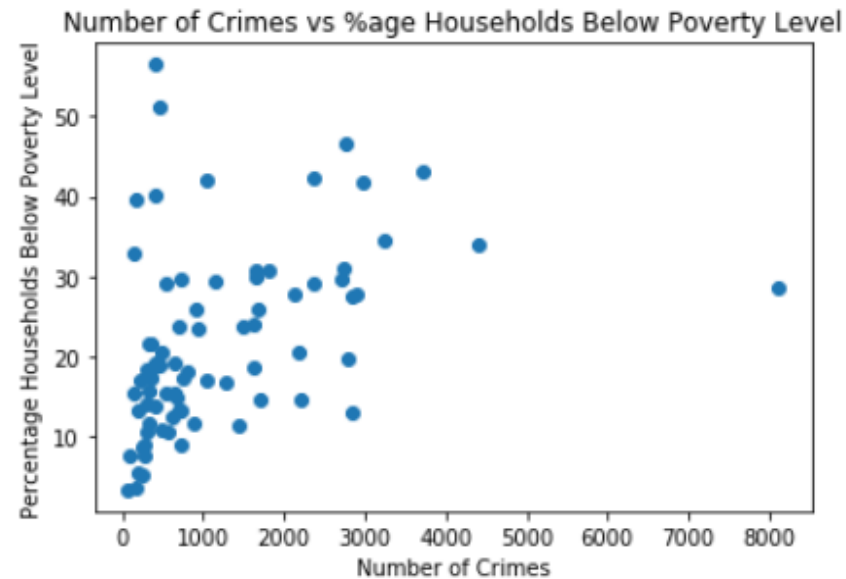
F. Classified Five Clusters using K-Means, which gives clustered Neighborhoods of similar Venues in Communities in Chicago



Data Visualization



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

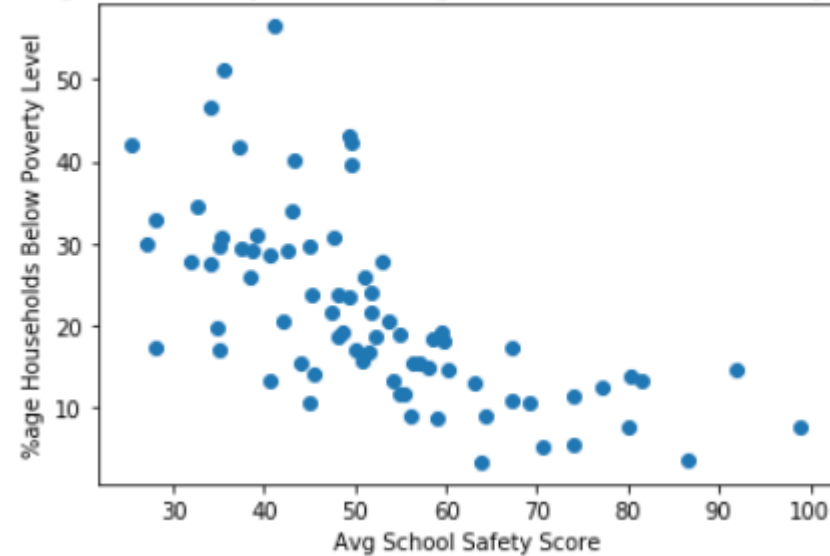


2. Relationship between (x) Total Number of Crimes and (y) Percent Households Below Poverty Level

Data Visualization

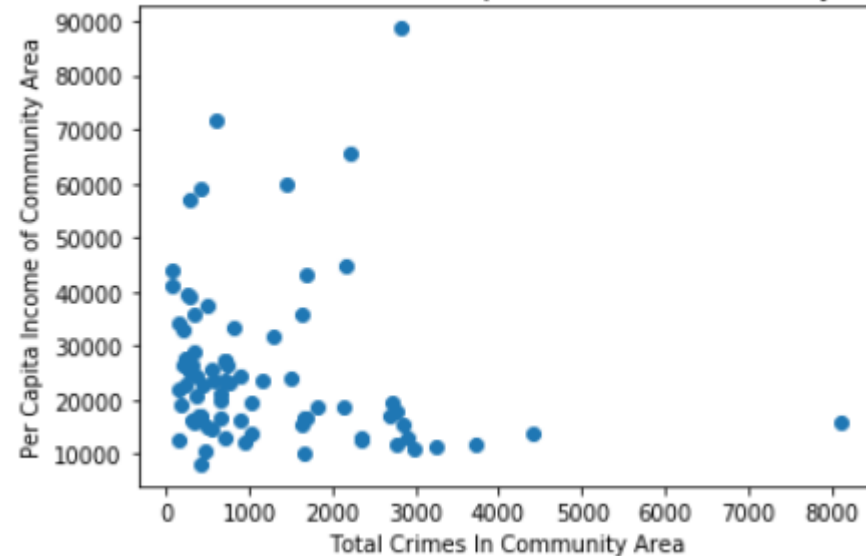


Avg School Safety Score vs %age Households Below Poverty Level



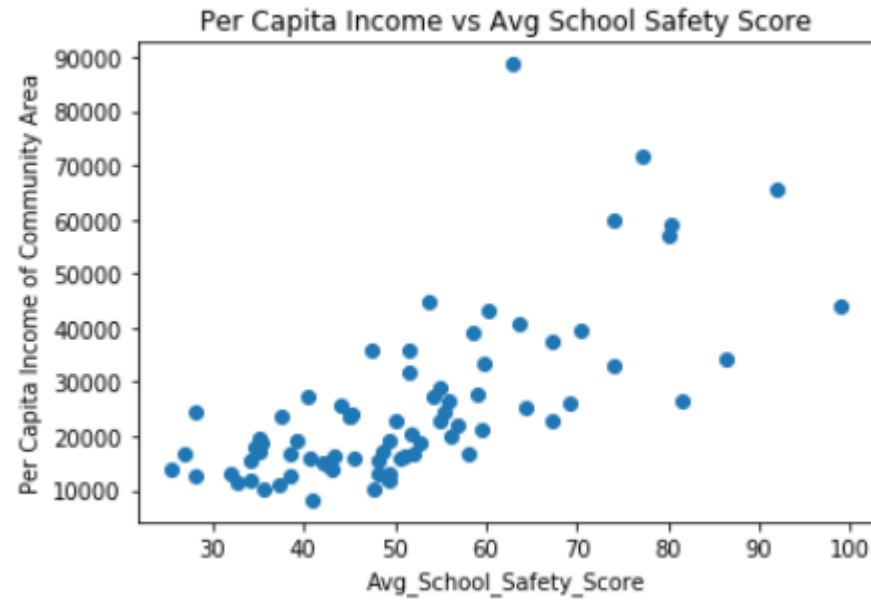
3. Relationship between (x) Average School Safety Score and (y) Percent Households Below Poverty Level

Number of Crimes vs Per Capita Income of Community Area



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

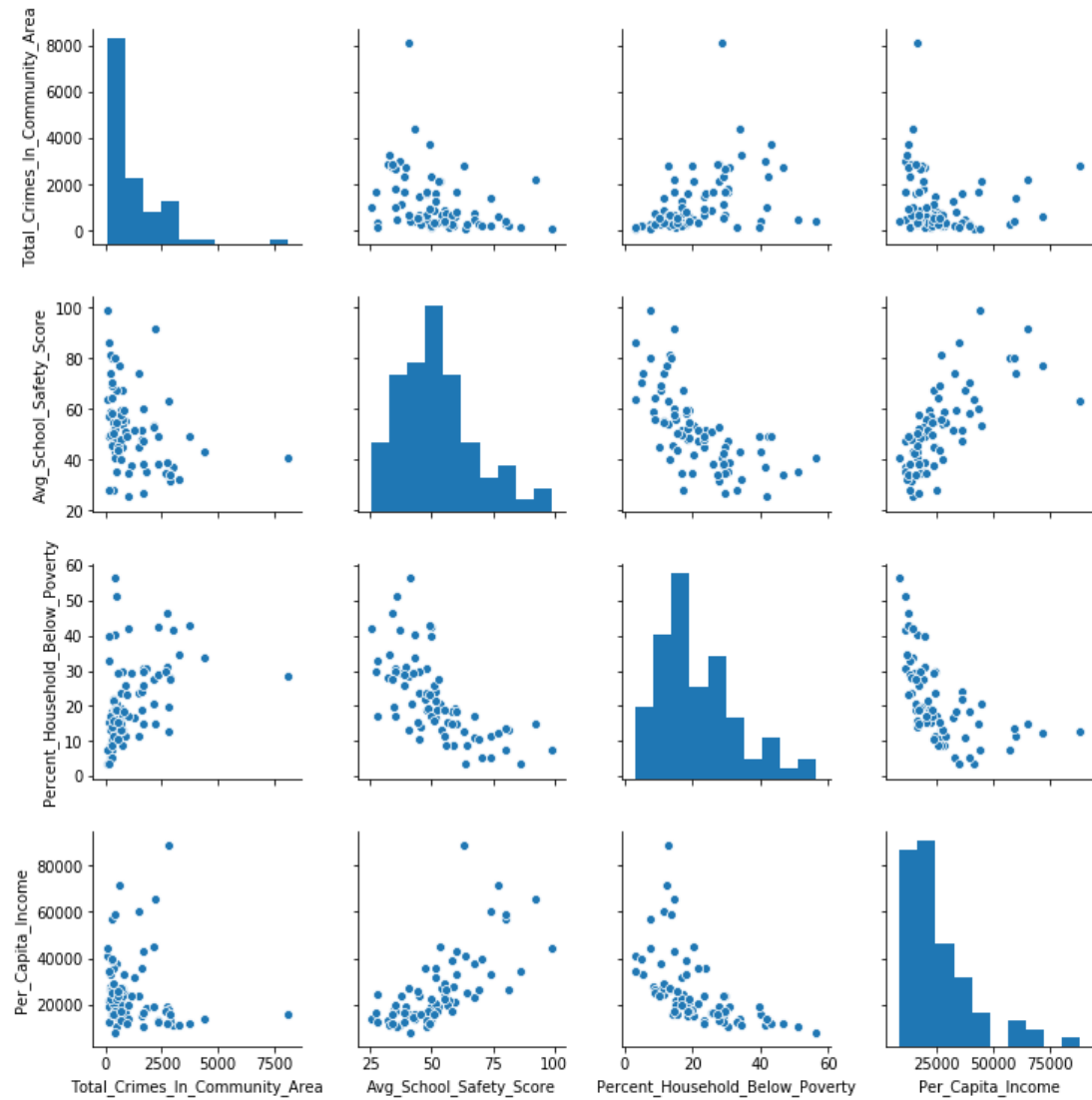
Data Visualization



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



Pair Plot



Heat Map



Correlations:

1. Total Number of Crimes
2. Average School Safety Score
3. Percent Households Below Poverty Level
4. Per Capita Income for each Community



Inferences

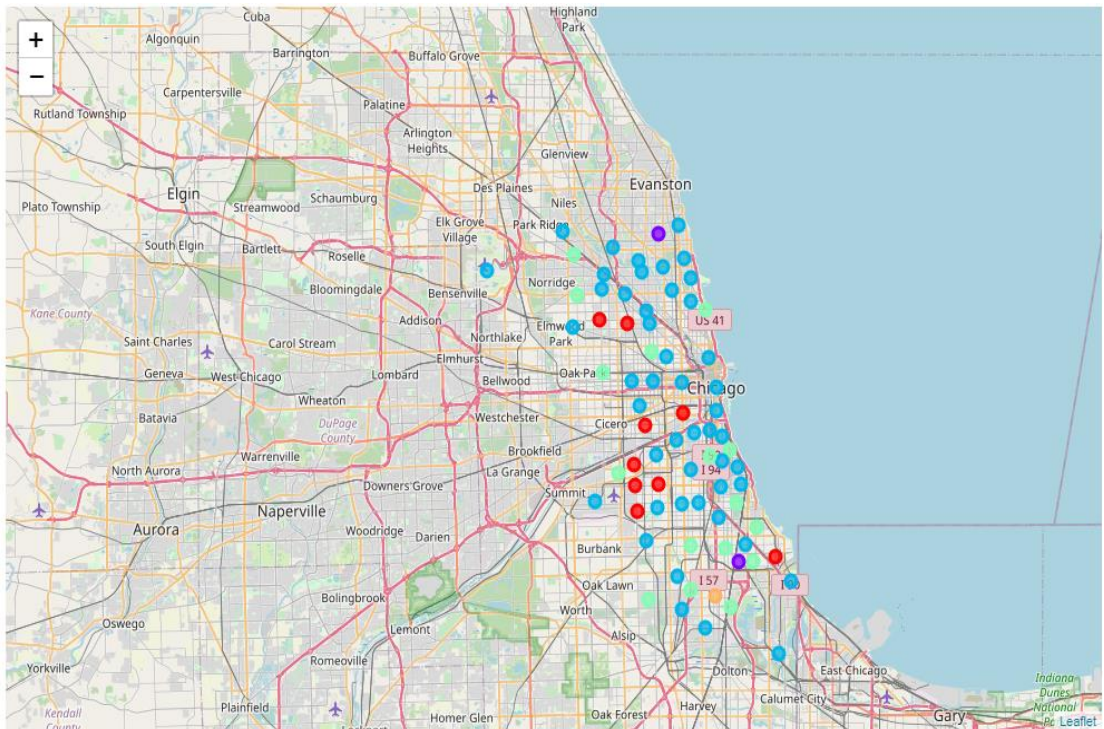
1. 'Number of Crimes' has Negative & Weak (0.33) Correlation with 'Average School Safety Score' of a Community Area

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

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

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

Classification



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**



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 technique which 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.

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



In Search for a
safe place
to live in Chicago

Thank You!



Monishankar Hazra

monishankar@gmail.com
[linkedin.com/in/monishankar](https://www.linkedin.com/in/monishankar)