The City of Chicago

Prakshi Raval

May 20, 2021

1. Introduction Section

- Background

The City of Chicago, is the most populous city in the state of Illinois, and the third most populous city in the United States. Lots of people are migrating to different states of U.S., they need lots of research to find the best community area/neighborhood out of all of them. The key aspects they look for in a particular community area are the average housing prices, the ranking of schools for children, the crime rate of the community area and the ease of access of the neighborhood to Cafes, Schools, Super Markets, Medical Shops, Grocery Shops, Malls, Theatres, Hospitals, etc. Therefore, this Project will help people in exploring better facilities around their community area and it will help them in making smart as well as an efficient decision on selecting a great community area out of a number of other community areas in Chicago, Illinois.

- Problem

Data that might contribute in determining the facilities available in the surroundings of the community area/neighborhood might include distance from the place to Schools, Bus Station, Hospitals, Grocery Stores, and Medical Shops as well as availability of Transportation facilities, Median Housing Prices and Low Crime Rate that describes the appropriateness of the community area. This project aims to predict the most appropriate community area to migrate to based on these data.

- Interest

Obviously, the people who are planning to migrate to the city of Chicago in the state of Illinois would be very interested in the prediction of most appropriate community area, as it will help them to get awareness of the area before the move to a new neighborhood, city, state or country for their work or to begin a new life. Others who might be interested would be a builder or a construction company who wants to build a housing society/apartment/bungalow in Chicago, Illinois but are confused about which community area would be the most appropriate for their construction project.

2. Data Section

- Data Sources

The Chicago community area and neighborhood data can be scraped from Wikipedia.

The dataset for average housing prices could be found on Zillow. The dataset for obtaining the crime rate for all the neighborhood is obtained from Chicago Data Portal. The dataset for rankings of Schools in Chicago could be found on Great Schools. Lastly, the data about different venues in various neighborhoods of a particular community area. In order to gain the information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all types of venues and events within an area of interest. Such information includes venue name, location, menu, photos and even ratings. As such, the foursquare location platform will be used as the sole data source since all the required information can be obtained through the API.

- Data Cleaning

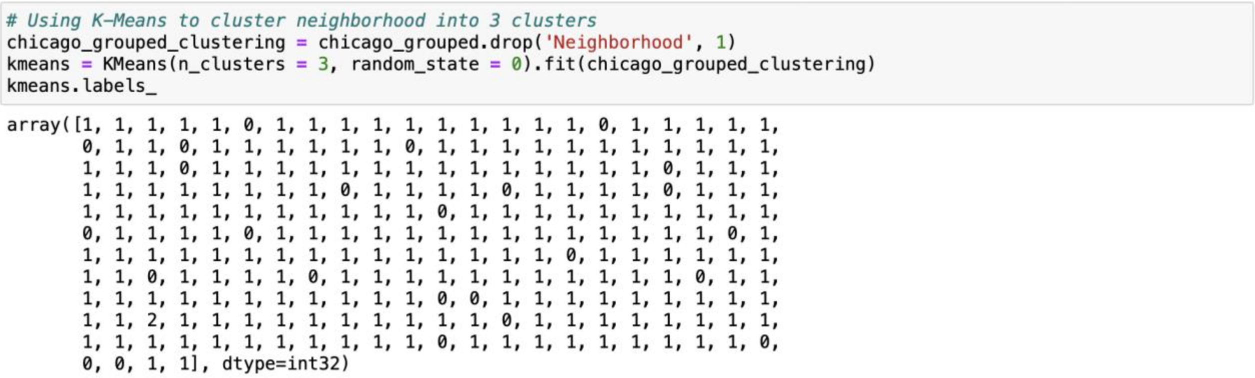
Data downloaded or scraped from multiple sources were combined into one table. There were a few missing values which only accounted for ~1% of the data. Therefore, the missing values were removed from the dataset. There are several problems with the datasets. First, the Zillow API has average housing prices for all the states and cities of the U.S. Therefore, expect the values belonging to Chicago, Illinois all others were cleaned from the dataset. Second, the dataset for calculating crime rate did not contain the names of community areas instead it just had community area id. Therefore, an SQL join operation was performed. After fixing all these problems, I checked for outliers in the data. I found that there were some extreme outliers, mostly caused by some types of small sample size problem. For example, there were community areas in the city of Chicago with extremely small or high population compared to the others. Therefore, I normalized the values for calculating the crime rate.

3. Methodology Section

- Clustering Approach

To compare the similarities of neighborhoods, we decided to segment them, and group them into clusters to find the appropriate neighborhoods will all the amenities in a big city like Chicago. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

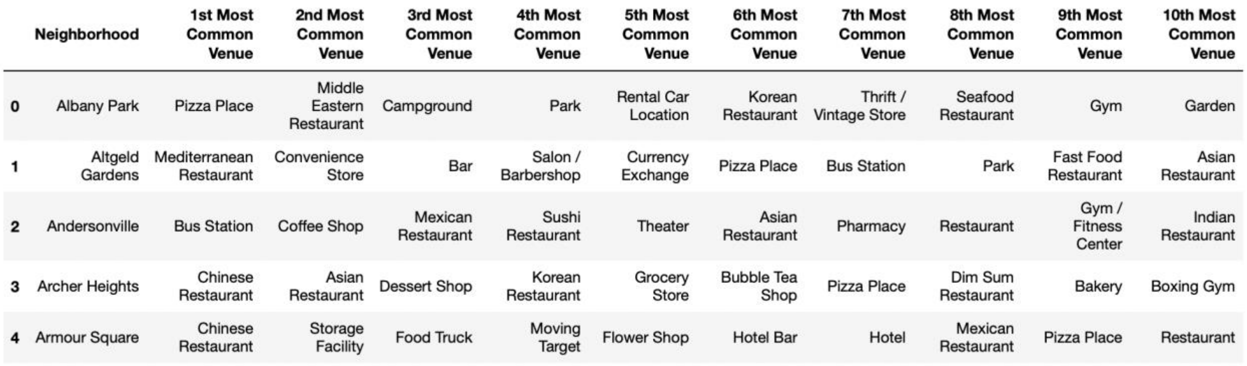
Using K-Means Clustering Approach



```
# Using K-Means to cluster neighborhood into 3 clusters
chicago_grouped_clustering = chicago_grouped.drop('Neighborhood', 1)
kmeans = KMeans(n_clusters = 3, random_state = 0).fit(chicago_grouped_clustering)
kmeans.labels_
```

```
array([1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,
       0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
       0, 0, 1, 1], dtype=int32)
```

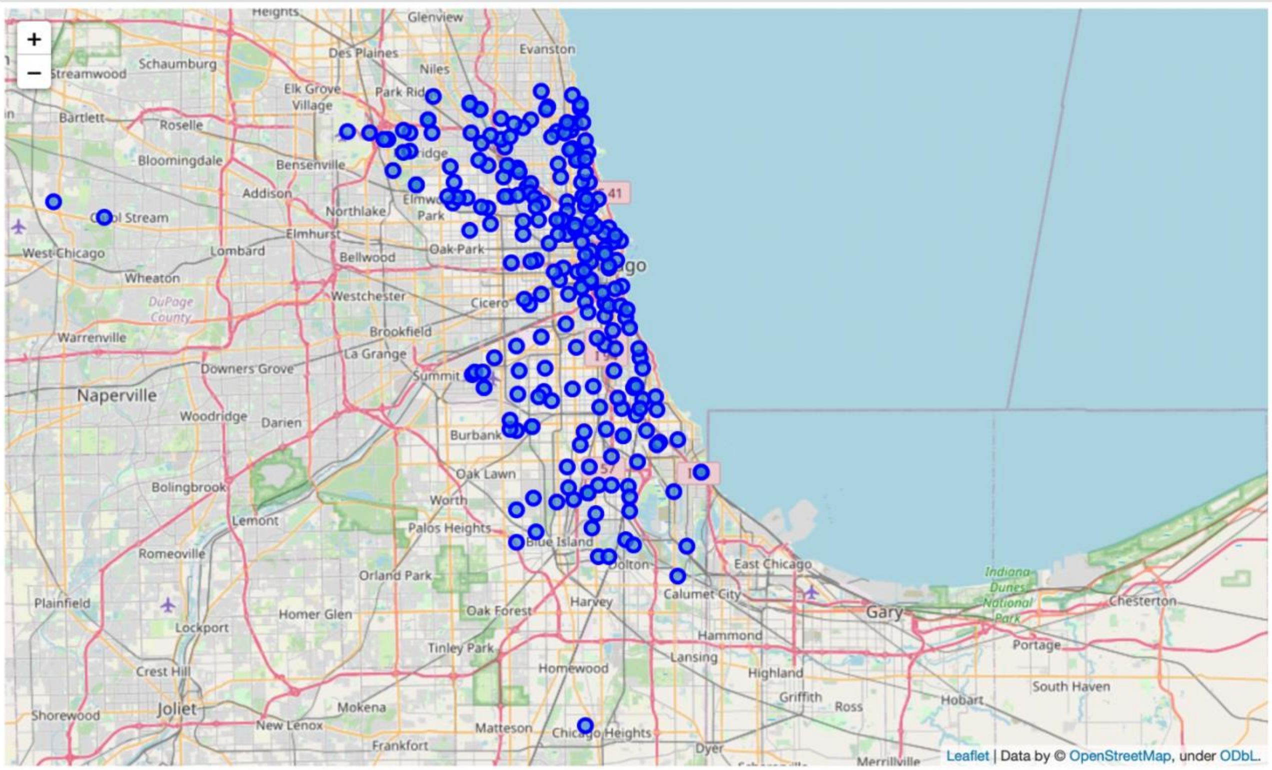
Most Common Venues near Neighborhood



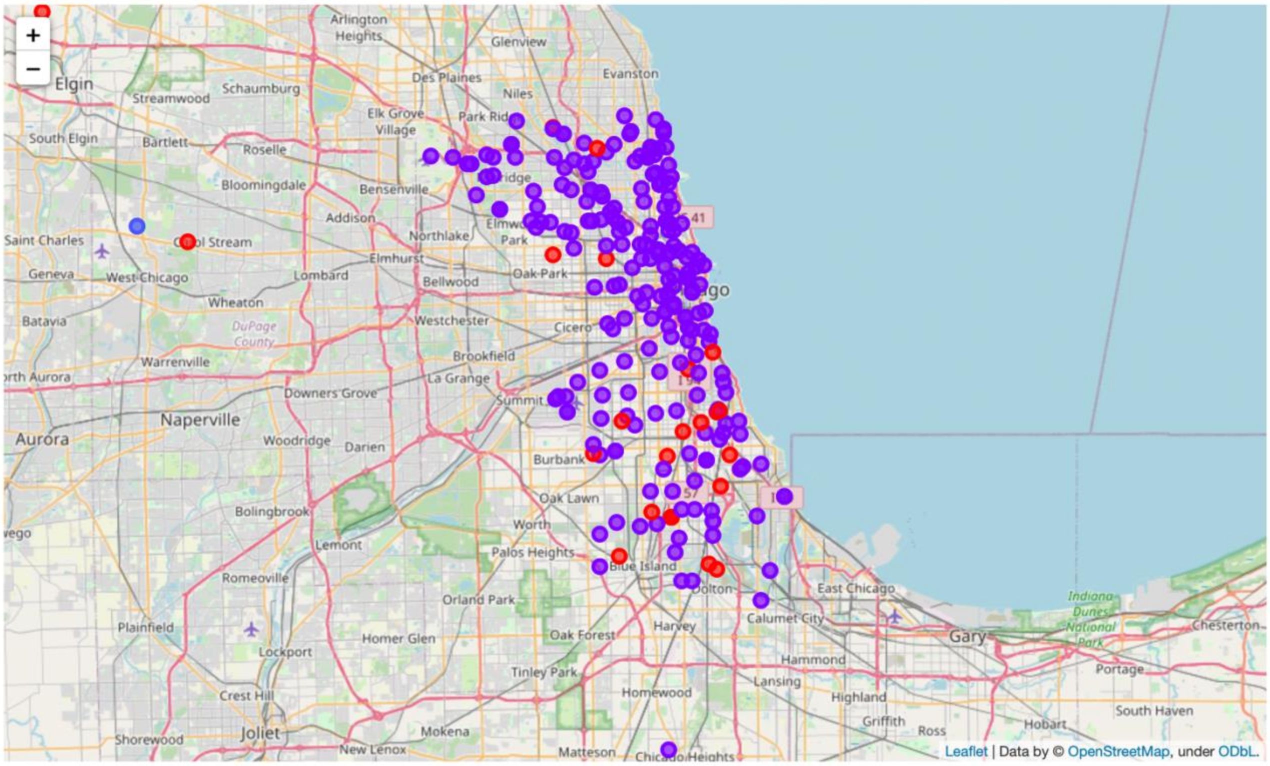
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park	Pizza Place	Middle Eastern Restaurant	Campground	Park	Rental Car Location	Korean Restaurant	Thrift / Vintage Store	Seafood Restaurant	Gym	Garden
1	Altgeld Gardens	Mediterranean Restaurant	Convenience Store	Bar	Salon / Barbershop	Currency Exchange	Pizza Place	Bus Station	Park	Fast Food Restaurant	Asian Restaurant
2	Andersonville	Bus Station	Coffee Shop	Mexican Restaurant	Sushi Restaurant	Theater	Asian Restaurant	Pharmacy	Restaurant	Gym / Fitness Center	Indian Restaurant
3	Archer Heights	Chinese Restaurant	Asian Restaurant	Dessert Shop	Korean Restaurant	Grocery Store	Bubble Tea Shop	Pizza Place	Dim Sum Restaurant	Bakery	Boxing Gym
4	Armour Square	Chinese Restaurant	Storage Facility	Food Truck	Moving Target	Flower Shop	Hotel Bar	Hotel	Mexican Restaurant	Pizza Place	Restaurant

4. Results Section

Map of Chicago Neighborhoods

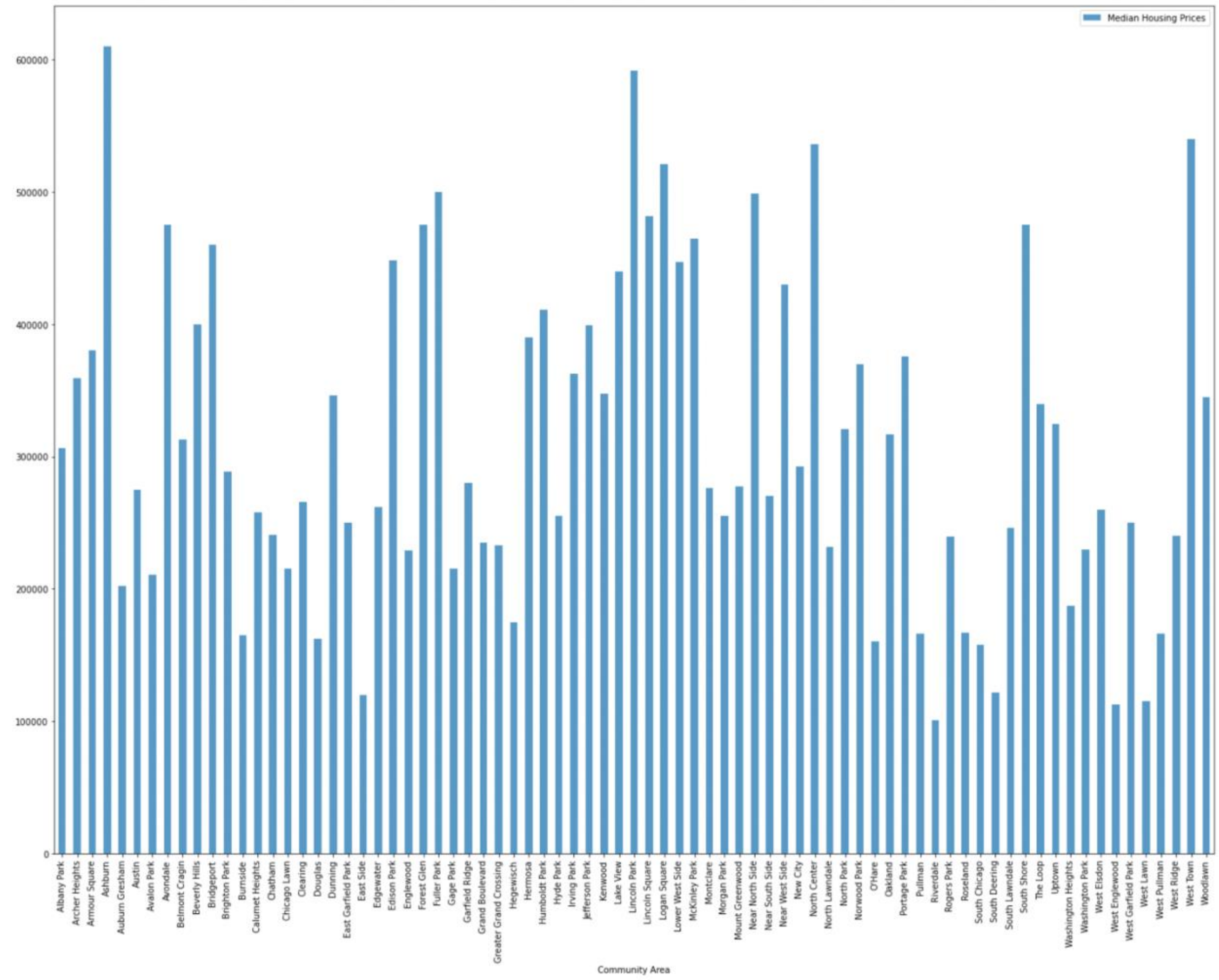


Map of Chicago after Clustering Venues Near Neighborhoods

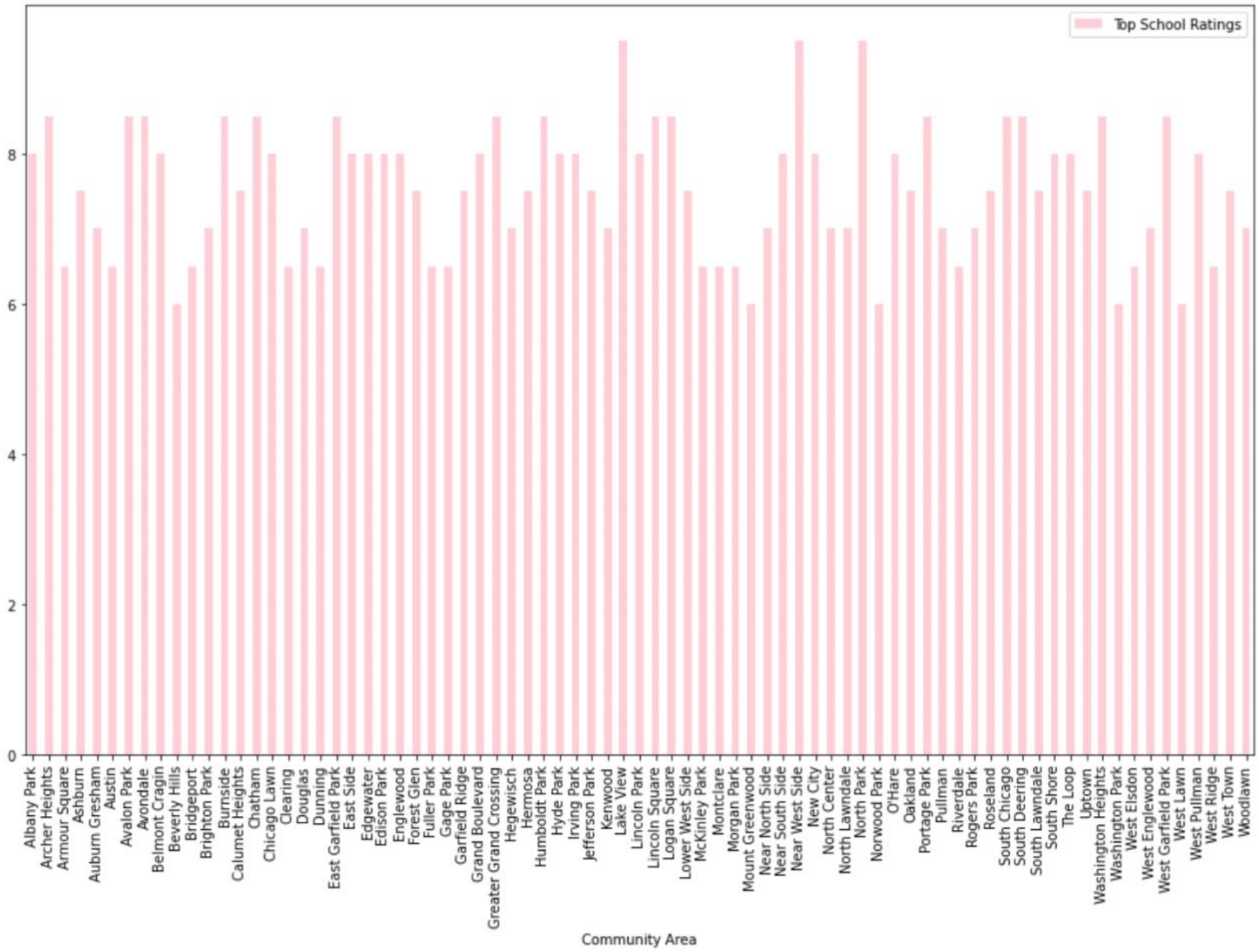


Leaflet | Data by © OpenStreetMap, under ODbL.

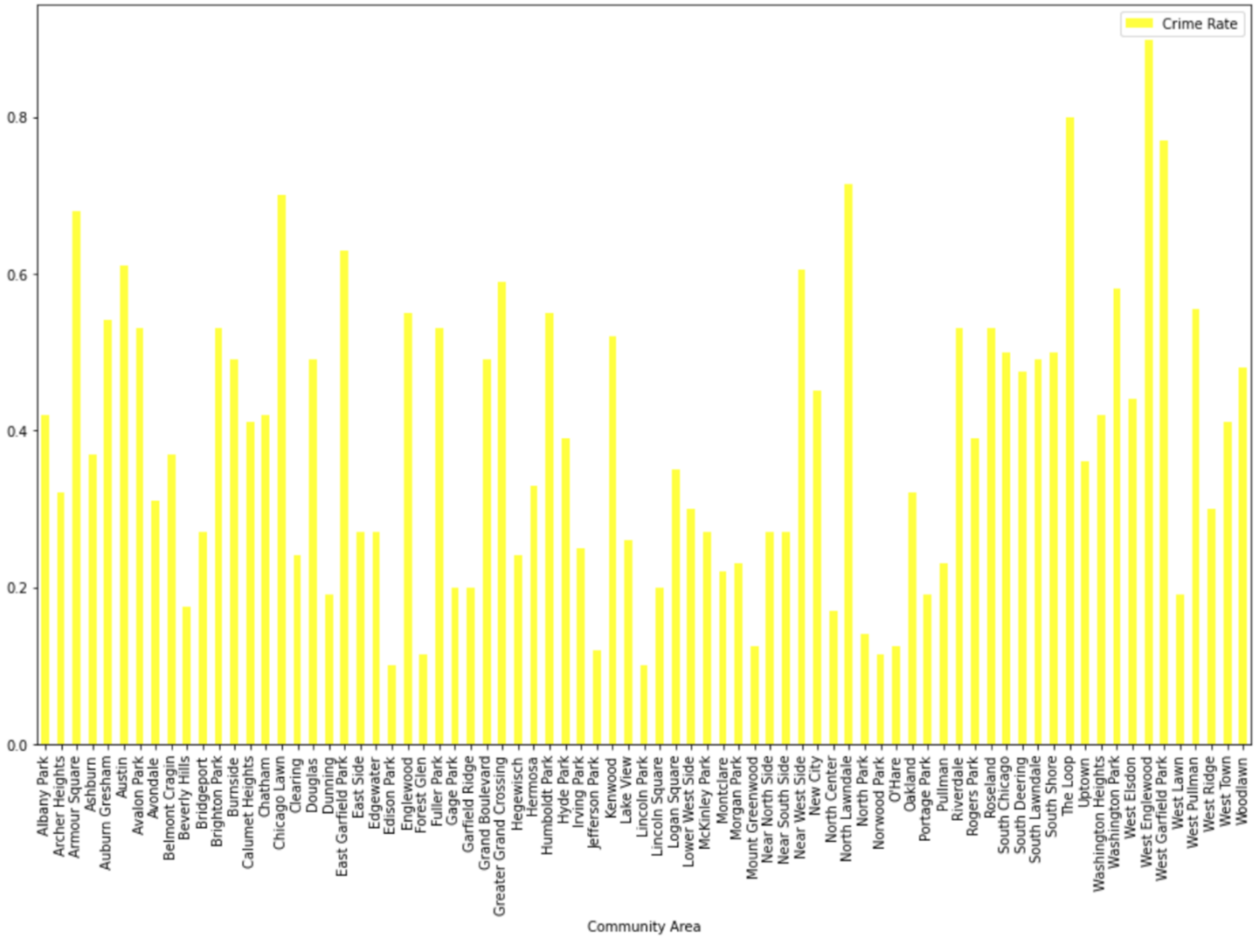
Average Housing Prices of Chicago Community Areas



Top School Ratings of Chicago Community Areas



Crime Rate of Chicago Community Areas



5. Discussion Section

• Insights

a. The top 10 list of venues which are in the vicinity of the neighborhood.

b. A sorted list of average housing prices in ascending or descending order.

c. A sorted list of schools in terms of location, fees, ratings and reviews.

d. A sorted list of community areas according to the crime rate of that place.

6. Conclusion Section

In this Capstone Project, using k-means cluster algorithm of Machine Learning I separated the neighborhoods into 10 different clusters and for 245 different latitude and longitude values obtained from the dataset, to find the similarities between all the neighborhoods. Using the charts above, results about a particular community area based on average house prices, school rating and crime rate have been made.

• Future Works

This Capstone project can be continued for making it more precise in terms to find the best place to reside in Chicago in terms of all the facilities available nearby and also in terms of cost effective.

- Libraries used in the Project

Pandas: For Dataframes.

Folium: Python Data Visualization Module.

Scikit Learn: For machine learning algorithms.

JSON: To handle JSON files.

XML: To handle XML data.

Geocoder: For retrieving Location Data.

Requests: For establishing connection with Foursquare API.

Matplotlib: Python Plotting Module.