Applied Data Science Capstone Project Report

Identifying the Best Neighborhoods to Move to in Chicago

By Sera K. Anderoglu

TABLE OF CONTENTS

I. INTRODUCTION/BUSINESS PROBLEM

Moving to a new city is exciting but at the same time can be a little bit of scary. Depending on the family size, age, style of living, etc. considerations and priorities differ when choosing the best neighborhood to Moving is expensive, thus you want to minimize the likelihood of having to move to another neighborhood after your first move to the city.

There are websites that list best neighborhoods in cities, however these are not sufficient in making decisions and don't allow one to weigh each neighborhood based on specific preferences.

In this project, we'll use data science to help a family moving to Chicago to locate the best neighborhoods that match their priorities/preferences. Here are the considerations of this family:

- Neighborhood safety (crime rates)
- School (K-12) rankings
- Proximity to venues such as parks, martial art schools, and music schools
- Housing prices
- Proximity to new job location (which is Downtown Chicago)

The considerations listed here can be considered quite common for families with young kids. In fact, any person moving to a place would have safety, proximity to work, and budget considerations. There would be differences in terms of venue preferences.

We'll provide this family a list of community areas that match their preferences. The explatory data analysis and visualizations will further provide valuable information about the city.

II. DATA COLLECTION AND CLEANING

Chicago is one of the largest metropolises in the United States with a population of 2.706 million as of 2018. It has more than 200 neighborhoods and 77 community areas. Community areas are used for city-wide statistical and planning purposes. As most data we need is available at the community area level, we perform our analysis at the community area level. We collect data on crime, school rankings, house prices, as well as driving distances and venues of interest. Here's a list of data collected:

1. Geographic boundaries of Chicago community areas
2. Wikipedia data on population and population density of community areas
3. Wikipedia data for community area - neighborhood relation
4. Latitude and longitude information for community areas
5. Crime Data
6. Zillow Housing Data
7. School Data
8. Venue data (Foursquare API)
9. Driving distance to Downtown Chicago

There are six main sources of data used in this project:

- Wikipedia
- Chicago Data Portal
- Zillow Housing Data
- Foursquare API
- Microsoft Bing Maps Distance API

- Geocoder from the geopy library

A. Geographic boundaries of Chicago community areas

This data is available in Chicago Data Portal as a geojson file. We use this data to create maps.
Source:https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

B. Wikipedia data on population and population density of community areas

We scraped this data from Wikipedia using Pandas. We use this data to match community area names to community area numbers as well as population density visualization and crime rate computation.
Source: https://en.wikipedia.org/wiki/Community_areas_in_Chicago#cite_note-City_basics-9

Community_Number	Community_Name	2017_Population	2017_Population_Density
1	ROGERS PARK	55062	29925.00
2	WEST RIDGE	76215	21590.65
3	UPTOWN	57973	24988.36
4	LINCOLN SQUARE	41715	16294.92
5	NORTH CENTER	35789	17458.05

C. Wikipedia data for community area - neighborhood relation

The housing prices are available at the neighborhood level. There is no dataset in the Chicago Data Portal showing the community area-neighborhood relation. We'll get this data from Wikipedia. Wikipedia list has 246 Chicago neighborhoods. Some neighborhoods are within boundaries of more than one community area. Different sources list different number of neighborhoods. For instance, Zillow Housing Dataset has 200 Chicago neighborhoods.
We create a dictionary that maps community areas to neighborhoods using this data.
Source: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago#cite_note-1

```
neighbor_dict=comm_neigh.groupby('Community_Name')['Neighborhood'].apply(list).to_dict()
neighbor_dict
```

```
{'ALBANY PARK': ['Albany Park',
  'Mayfair',
  'North Mayfair',
  'Ravenswood Manor'],
 'ARCHER HEIGHTS': ['Archer Heights'],
 'ARMOUR SQUARE': ['Armour Square', 'Chinatown', 'Wentworth Gardens'],
 'ASHBURN': ['Ashburn',
```

D. Latitude and longitude information for community areas

We used the **geopy.geocoder** library to extract the latitude and longitude information for community areas. We provided the address as “community area name + Chicago”. We use this information when using Foursquare API to search for venues, as well as for the Microsoft Bing Maps Distance API to

find driving distances from community areas to Downtown. Latitude-longitude data is also used for creating labels on the maps.

E. Crime Data

This data set displays the incidents of crime reported in Chicago from 2001 to present (minus the most recent 7 days). In addition to details on the nature of the crime, this data set also reports the community area that the crime took place.

Since this is a huge dataset, we'll use SODA API to filter results to focus on certain columns and years (2019 and 2020), and only download the relevant data.
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

To do so, we need to pass a query (SQL-like) to the Socrata client. We'll get the *date*, *primary_type*, *community_area* and *year* columns for the years 2019 and 2020.

query='Select date, primary_type,community_area, year where year > 2018 limit 500000'

In this query, it is important to explicitly set limit to a large value. The default value is 1000, and if left as is, your query will only return 1000 of the rows that match your query.

We clean the data and group by community area to determine the total number of crimes reported in each community area since the beginning of 2019.

date	crime	community_area	year
2019-12-03T13:08:00.000	BATTERY	66	2019
2019-07-22T00:01:00.000	DECEPTIVE PRACTICE	2	2019
2019-12-07T12:33:00.000	SEX OFFENSE	42	2019
2019-12-04T06:00:00.000	SEX OFFENSE	35	2019
2019-07-21T18:35:00.000	BATTERY	30	2019

Community_Number	Number_of_Crimes
1	6952
2	6187
3	5963
4	3458
5	2259

F. Zillow Housing Data

Zillow has time series data available for download. Source: https://www.zillow.com/research/data/
Among several datasets listed here, we will use **ZHVI Single-Family Homes Time Series data** at the neighborhood level.

ZVHI stands for Zillow Home Value Index. This data set displays typical value in dollars for all single-family homes in a given region. After downloading this data set, we extract the entries for

Chicago Neighborhoods. In Zillow dataset, there are 200 Chicago Neighborhoods.

	RegionID	SizeRank	RegionName	RegionType	StateName	State	City	Metro	CountyName	1996-01-31	...	2020-01-31	2020-02-29	2020-03-31	2020-0
53	269592	53	Logan Square	Neighborhood		IL	Chicago	Chicago-Naperville-Elgin	Cook County	154086.0	...	499367.0	503470.0	508541.0	51350
38	403117	88	Little Village	Neighborhood		IL	Chicago	Chicago-Naperville-Elgin	Cook County	NaN	...	173503.0	173270.0	173594.0	17347
05	403169	106	West Rogers Park	Neighborhood		IL	Chicago	Chicago-Naperville-Elgin	Cook County	157304.0	...	364883.0	367482.0	370122.0	37211
42	403120	144	South Austin	Neighborhood		IL	Chicago	Chicago-Naperville-Elgin	Cook County	NaN	...	178077.0	179646.0	181244.0	18263
								Chicago-							

For the purposes of this project, we use the ZVHI estimate of the very last month, October 2020. We create a dictionary that maps neighborhoods with median house prices. This dictionary is used to estimate the median house prices of community areas.

```
price_dict= zillow_data.set_index('Neighborhood').to_dict()['House_Price_$']
price_dict
```

```
{'Logan Square': 542723.0,
 'Little Village': 184240.0,
 'West Rogers Park': 389635.0,
 'South Austin': 205639.0,
 'Albany Park': 395389.0,
 'Uptown': 738394.0,
 'Lake View': 978092.0,
 'Rogers Park': 410840.0,
 'Gresham': 159790.0,
 'Brighton Park': 227531.0,
 'Jefferson Park': 356928.0,
 'Portage Park': 353911.0,
 'South Loop': 642945.0,
```

G. School Data

Chicago Data Portal has several datasets available in the education category. Among these "Chicago Public Schools - School Profile Information SY2021" lists schools including their school quality rating.

Source: https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/83yd-jxxw

The relevant columns to our project are: *Overall_Rating, Rating_Status, Rating_Statement, School_Latitude, School_Longitude*

	Overall_Rating	Rating_Status	Rating_Statement	School_Latitude	School_Longitude
0	Inability to Rate	GOOD STANDING	This school did not have enough data to receiv...	41.842533	-87.695261
1	Level 2+	NOT APPLICABLE	This school received a Level 2+ rating, which ...	41.801762	-87.711025
2	Level 1	GOOD STANDING	This school received a Level 1 rating, which i...	41.928841	-87.669528
3	Level 2+	NOT APPLICABLE	This school received a Level 2+ rating, which ...	41.778130	-87.598114
4	Level 2	PROVISIONAL SUPPORT	This school received a Level 2 rating, which i...	41.935634	-87.783515

We want to gather data at the community area level. We need to determine which community areas the schools belong to. We write a function that returns the community area number a latitude-longitude pair belongs to.

```
def find_community_area(latitude, longitude):
    point= Point(longitude,latitude)
    for feature in c_data['features']:
        polygon = shape(feature['geometry'])
        if polygon.contains(point):
            return feature['properties']['area_numbe']
    return 0
```

We assign numerical values to Overall_Rating categories. The table below summarizes the meaning of school quality rankings as well as the numerical scores we'll use.

Overall Rating	Overall Score
Level 1+ (good standing, above 90th percentile)	5
Level 1 (good standing, above 70th percentile)	4
Level 2+ (good standing, above 50th percentile)	3
Level 2 (provisional support, minimum at 40th percentile)	2
Level 3 (intensive support)	1

There are 21 schools which were not rated (*'inability to rate'*) because of not having enough data. Looking at the Rating_Status of these schools, we decided to impute these missing values with an overall score of 3. Almost all of these schools had a 'good standing' rating status.
Finally, we calculate the median school score for each community area.

Community_Number	Median_School_Score
1	4.0
10	4.0
11	4.5
12	5.0
13	4.5

H. Venue data (Foursquare API)

For each community area, we'll determine:

- Number of parks within 1000m
- Number of music schools within 2000m
- Number of martial art schools within 2000m

We find the category id's for these venues, and use the search endpoint of the Foursquare API to get the total number of venues of these types in each community area.

Source: https://developer.foursquare.com/docs/build-with-foursquare/categories/

Venue category	Foursquare Category ID
Martial Arts Dojo	4bf58dd8d48988d101941735
Music School	4f04b10d2fb6e1c99f3db0be
Park	4bf58dd8d48988d163941735

Here's part of the data we collected:

Community_Number	Num_of_Parks	Num_of_Martial_Arts	Num_of_Music_School
1	19	4	1
2	9	1	1
3	33	6	4
4	11	14	4
5	10	23	6
...	...	...	...

I. Driving distance to Downtown Chicago

We use **Microsoft Bing Maps Distance Matrix API** to calculate the driving distance from each community area to Downtown Chicago since this is the job location for the family.

Source: https://docs.microsoft.com/en-us/bingmaps/rest-services/routes/distance-matrix-data
After creating a Developer account, we create a url with information on origins, destinations, API key and specify **travelMode=driving**.

```
url='https://dev.virtualearth.net/REST/v1/Routes/DistanceMatrix?origins={}&destinations={}&travelMode=driving&key={}
    origins,
    destination,
    bing_key)
```

We gather the data in a pandas dataframe.

	Community_Number	Travel_Distance_km
0	1	17.959
1	2	19.432
2	3	12.446
3	4	16.141
4	5	13.518

After collecting all these data, we combined them in a pandas dataframe called chicago_data:

Community_Number	Community_Name	2017_Population	2017_Population_Density	Latitude	Longitude	Number_of_Crimes	Median_House_Price	Median_Sch
1	ROGERS PARK	55062	29925.00	42.010531	-87.670748	6952	410840.0	
2	WEST RIDGE	76215	21590.65	42.003548	-87.696243	6187	429336.0	
3	UPTOWN	57973	24988.36	41.966630	-87.655546	5963	738394.0	
4	LINCOLN SQUARE	41715	16294.92	41.975990	-87.689616	3458	719352.0	

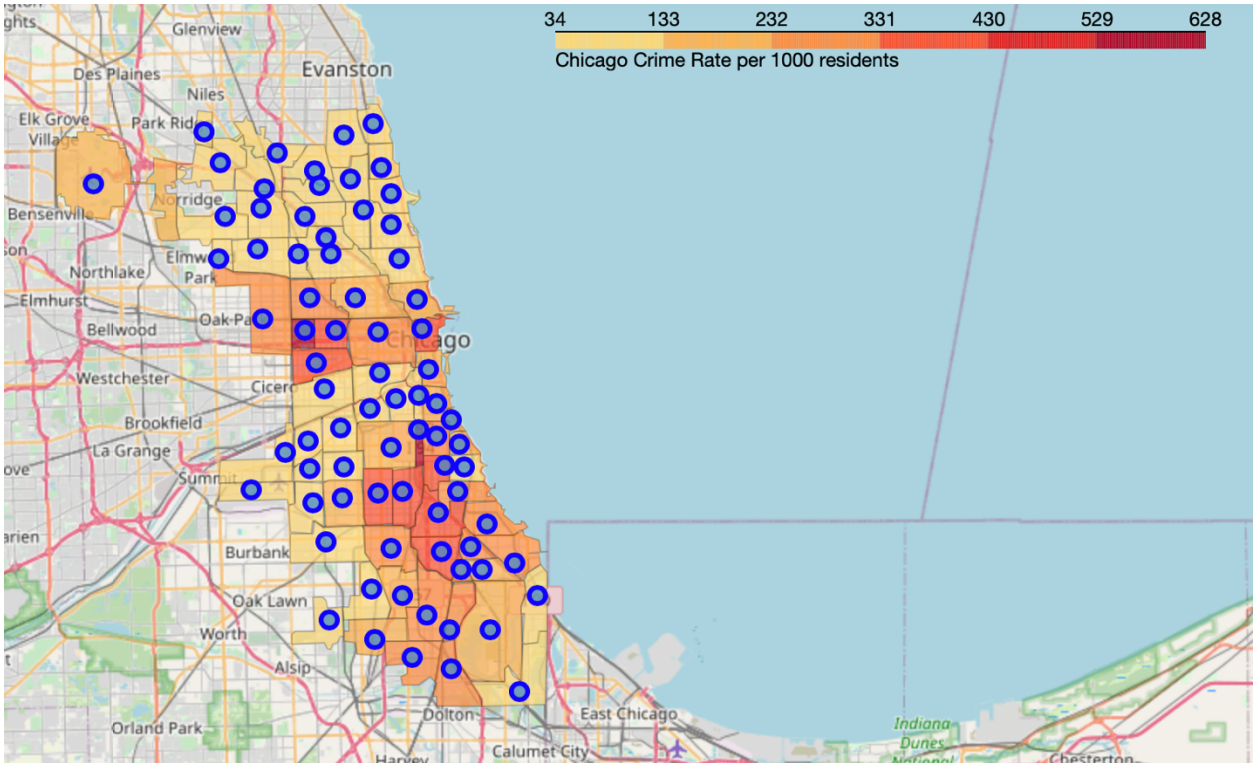
III. METHODOLOGY

We performed explatory data analysis as well as clustering on our data set. Before then we redefined the feature relevant to crimes. Instead of using number of crimes, we defined a new feature called crime rate per 1000 residents.

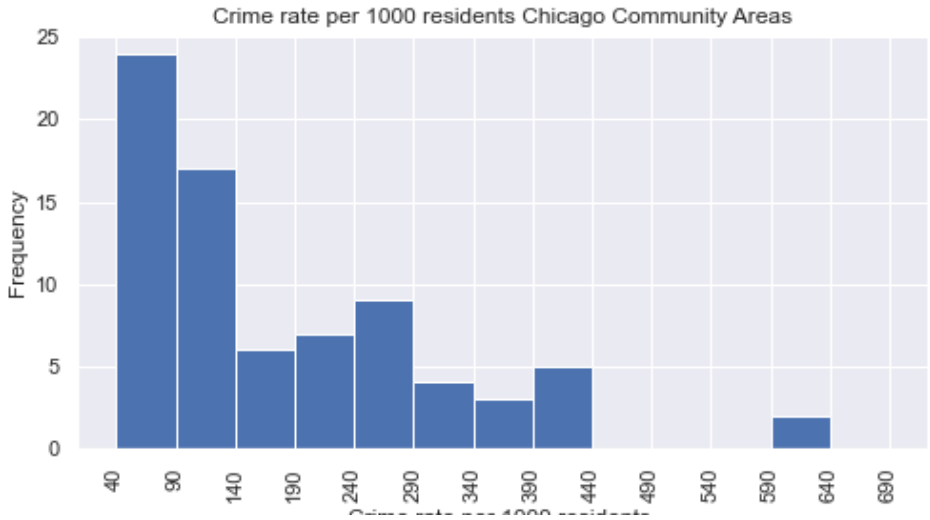
A. Explatory Data Analysis

We used folium choropleth maps to create visualizations of each feature we have in our data frame. We also created histograms. These visualizations provided a lot of insight about community areas. In the choropleth maps, clicking on the blue markers displays the community name and the relevant feature value.

CRIME RATES:

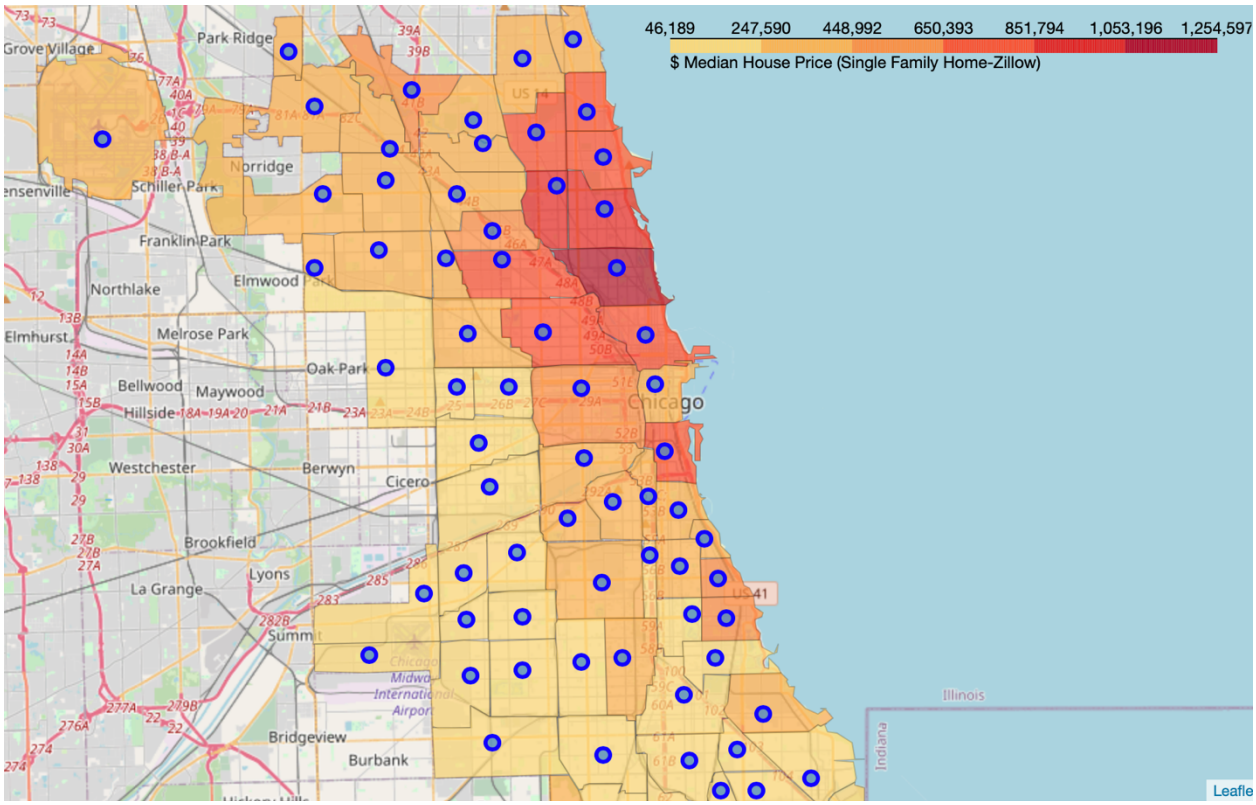


Based on this map, it appears that Northern Chicago and Southwest Chicago are safer than other regions. The crime rate is highest in West Garfield Park and Fuller Park followed by East Garfield Park, North Lawndale, Englewood, West Englewood, Greater Grand Crossing and Chatham.

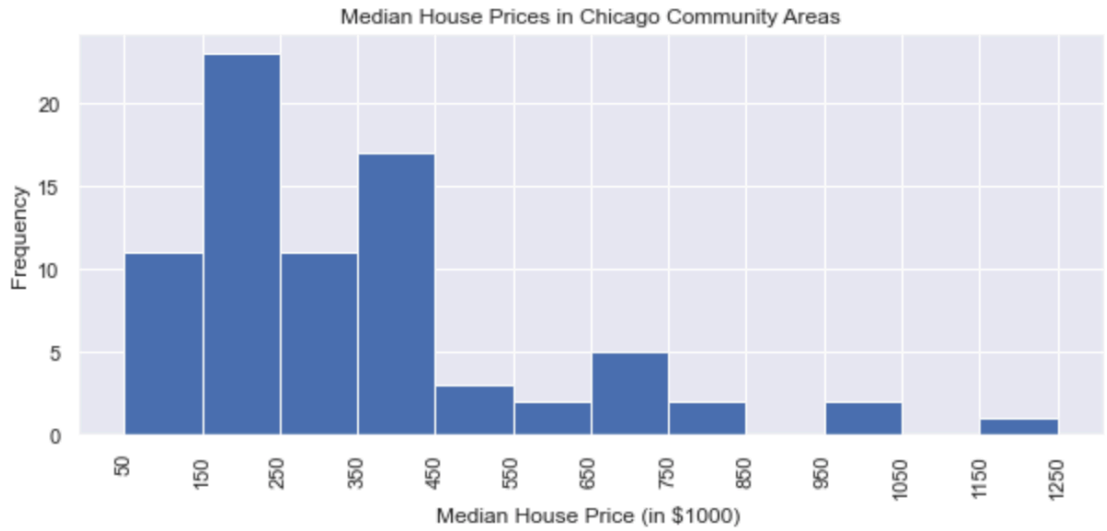


The distribution of crime rates is skewed to the right meaning there are fewer number of community areas with higher crime rates. This is an expected distribution.

MEDIAN HOUSE PRICES:

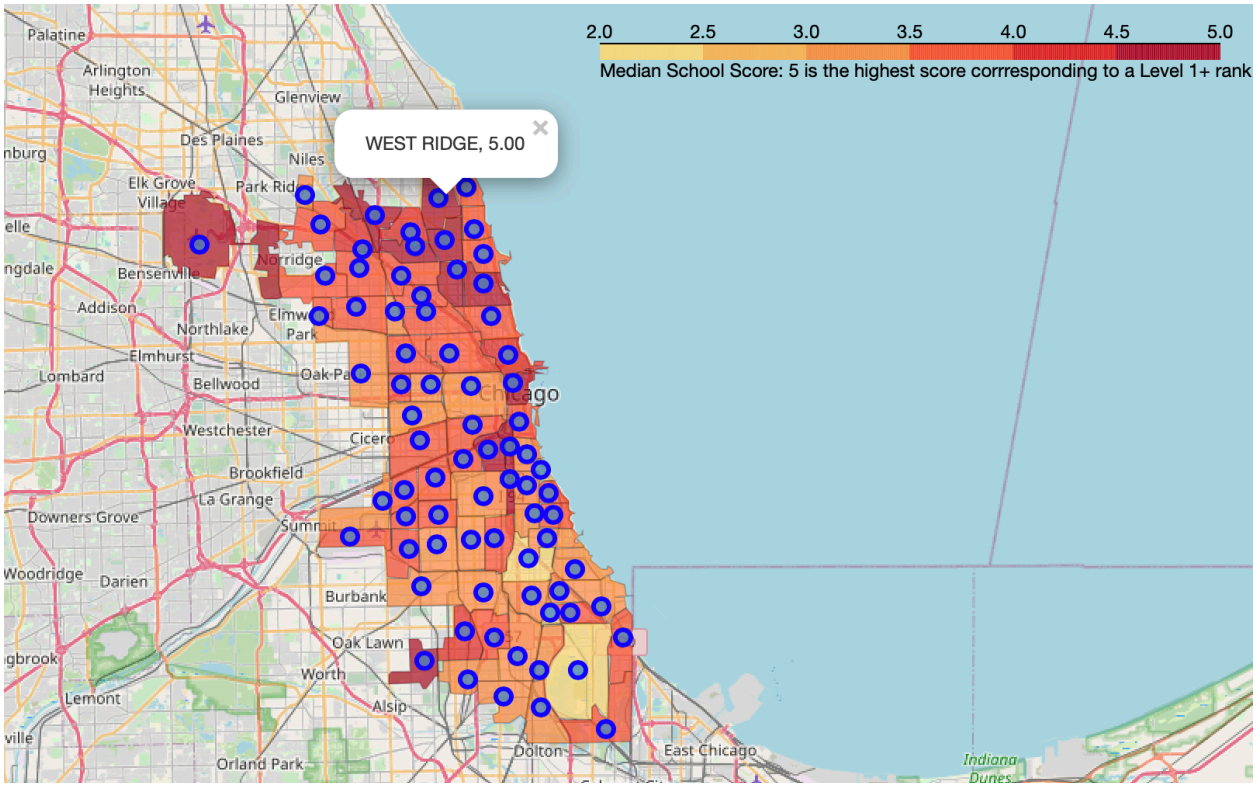


Northeast Chicago has the highest population density. It appears that the median house prices are pretty high in this region. The community areas in the west and south have lower median house prices.



The distribution of median house prices is also skewed to the right meaning that there are very few community areas with very high median house prices. The highest prices are in Lincoln Park followed by North Center and Lake View.

MEDIAN SCHOOL SCORE:



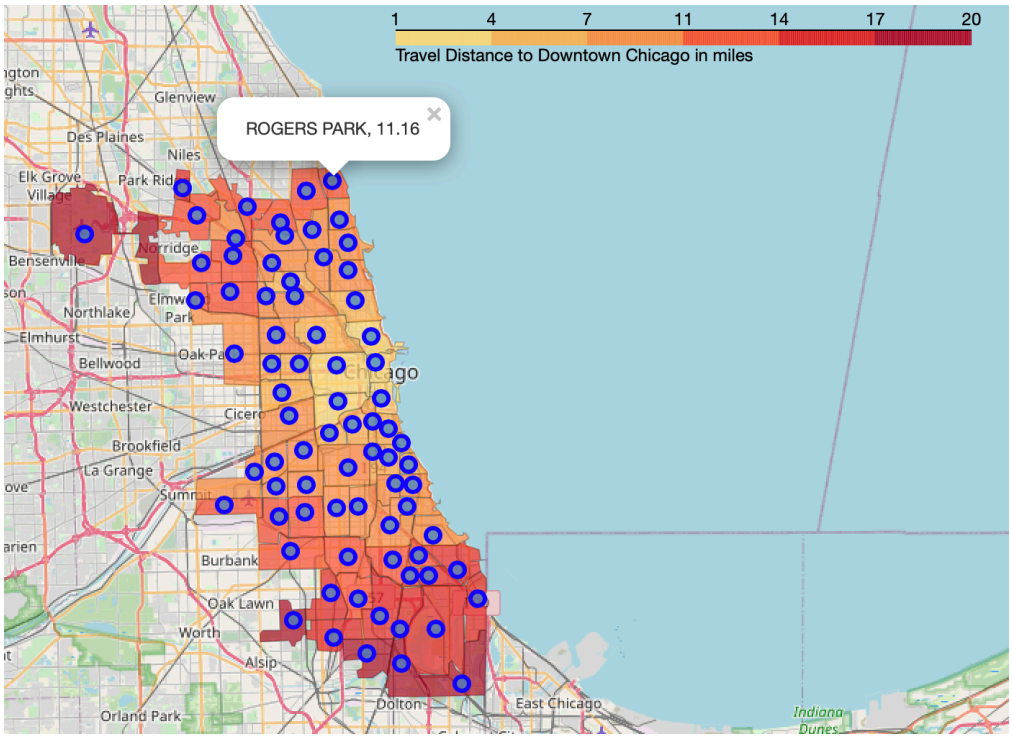
Northern Chicago and South West Chicago seem to have higher median school scores although there are some other highly ranked community areas as well.

Here are the community areas with the lowest and highest median school scores:

Community_Name	Median_School_Score
GREATER GRAND CROSSING	2.0
SOUTH DEERING	2.0
BURNSIDE	2.0
OAKLAND	2.0
RIVERDALE	3.0
AUSTIN	3.0
EAST GARFIELD PARK	3.0
NEAR WEST SIDE	3.0
NORTH LAWNDALE	3.0

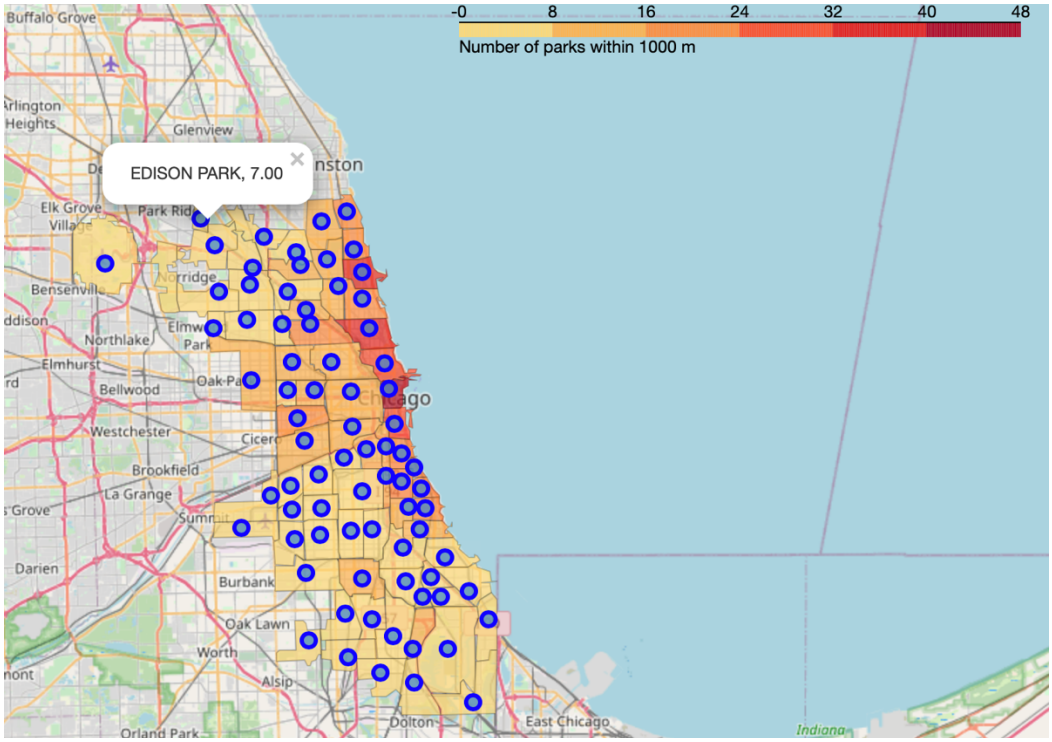
Community_Name	Median_School_Score
ALBANY PARK	5.0
ARMOUR SQUARE	5.0
OHARE	5.0
LINCOLN SQUARE	5.0
NORTH CENTER	5.0
LAKE VIEW	5.0
MOUNT GREENWOOD	5.0
WEST RIDGE	5.0
LOOP	5.0
BRIDGEPORT	5.0
FULLER PARK	5.0
FOREST GLEN	5.0

TRAVEL DISTANCE:



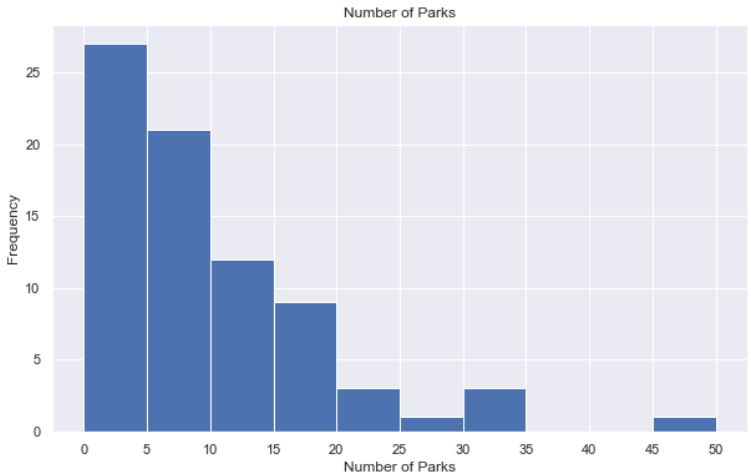
The pattern on this map was as expected. This visualization helped us to make some community name corrections in our dataframe to match with the geojson file.

NUMBER OF PARKS:

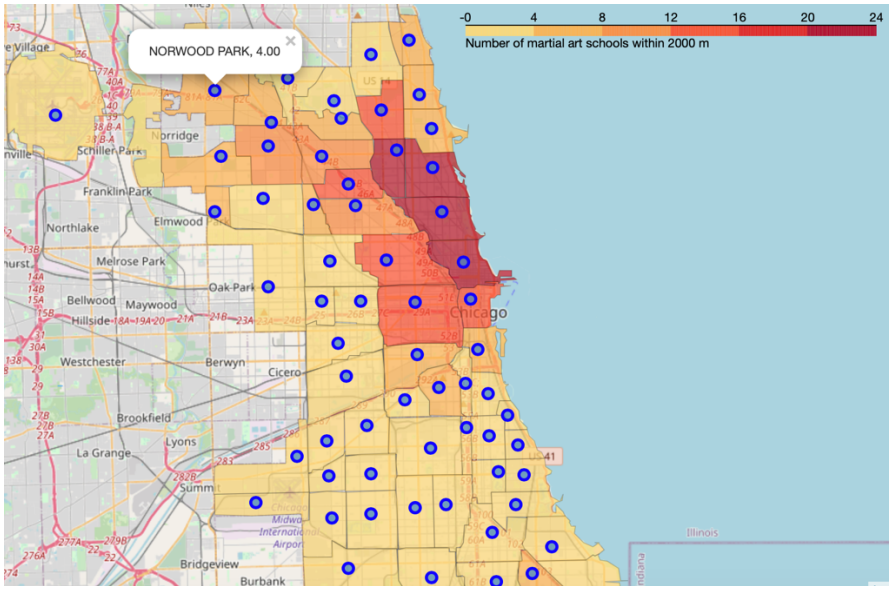


Looking at the visualization, the community area with the greatest number of parks (48) is the Loop, which is a portion of Downtown Chicago. Further observation led to the fact that Foursquare API considers a wide range of venues under the park category including botanical gardens, monuments, etc.

The distribution of number of parks is also skewed to the right.

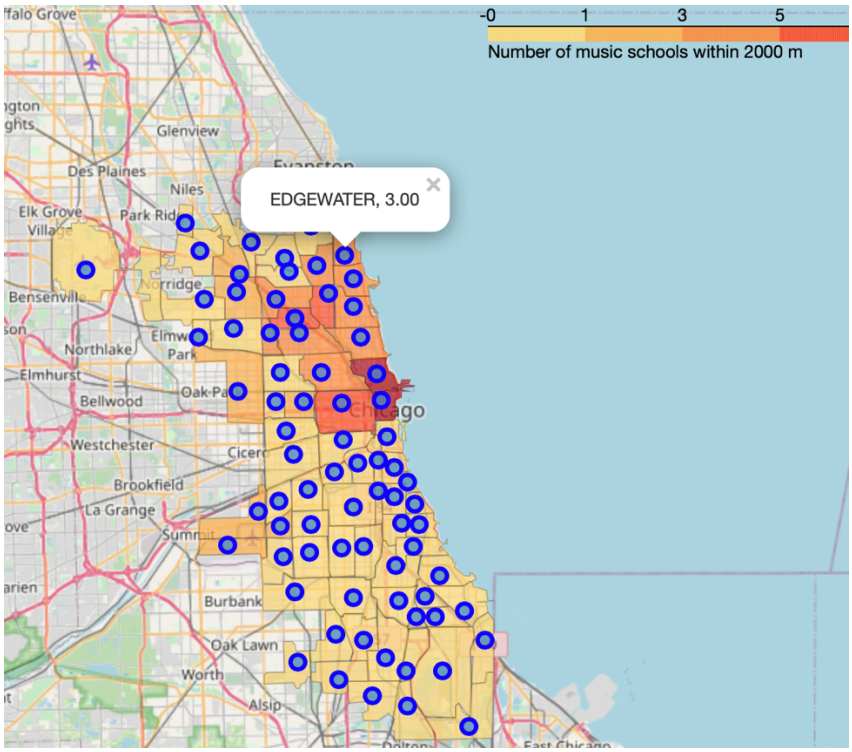


MARTIAL ARTS SCHOOLS:



The community areas in the Northern side have more martial arts schools. This is in fact not surprising as these areas are highly populated.

MUSIC SCHOOLS:



This pattern is quite similar to martial arts schools except that there are more martial arts schools in Chicago then music schools.

All these visualizations already give an idea about the location this family would want to be close to. The North side of Chicago is safer, has high median school scores, and many venues. The only downside is that the house prices are much higher in this region.

B. Clustering

With the help of folium maps, we get a pretty good understanding about how community areas are clustered based on different features. However, we'd like to form clusters of community areas based on all of these features.

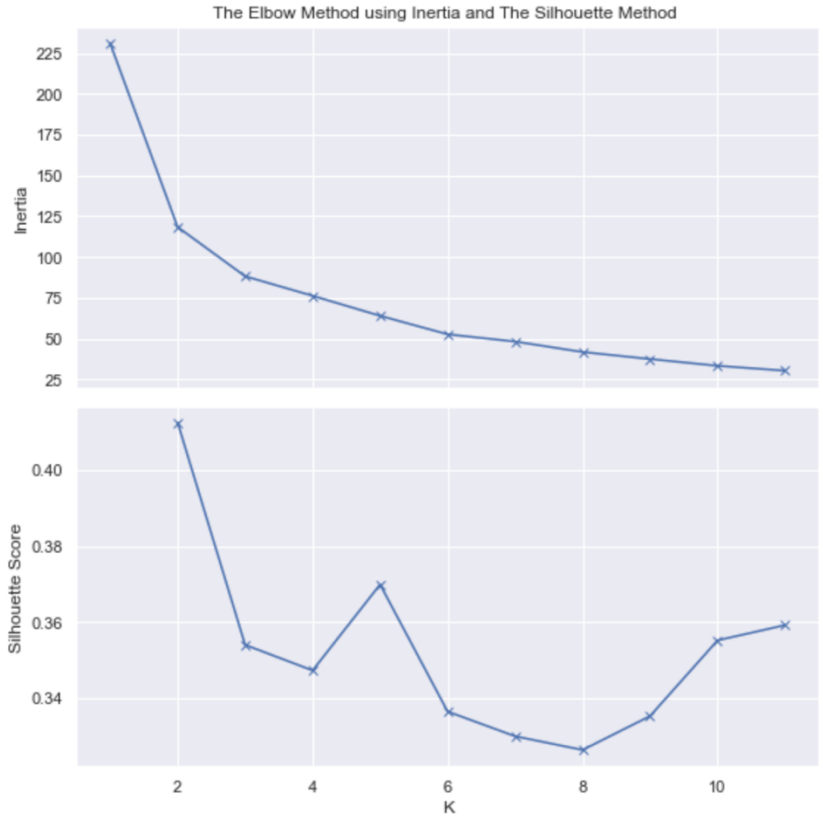
We use KMeans clustering algorithm for this purpose. The KMeans algorithm clusters data by trying to separate samples in K groups of equal variance, minimizing a criterion known as the *inertia* or within-cluster sum-of-squares. It is a heuristic algorithm which means optimal solution is not guaranteed. Because of the way distances are calculated, it is important to scale data.

We normalized and scaled our features (for the features with skewed distributions, we took the logarithm first to normalize and then applied standard scaler to come up with the z scores).

For clustering, we categorized our features as primary and secondary. Primary features are the ones that are more important such as crime rate, median house price, and school ranking. We first clustered the community areas into two based on primary features. The reason for this methodology is to account for the importance of features.

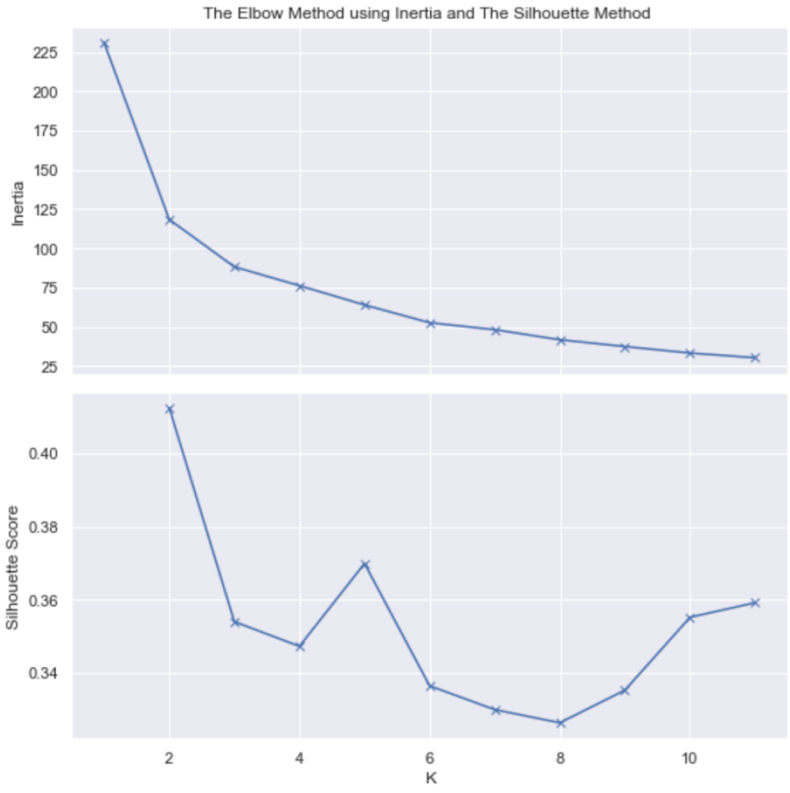
To determine K, we plot K values against inertia. Inertia is a measure of how internally coherent clusters are. We look for the K value after which the decrease in inertia diminishes. This is called the **Elbow method** and the point is called the elbow of the curve. Elbow method is also a heuristic method. Sometimes it is hard to locate the elbow of the curve. The other useful method to determine K is plotting silhouette scores and locating the K that gives the peak score. Silhoutte score is a measure of how close each point in one cluster is to points in the neighboring clusters.

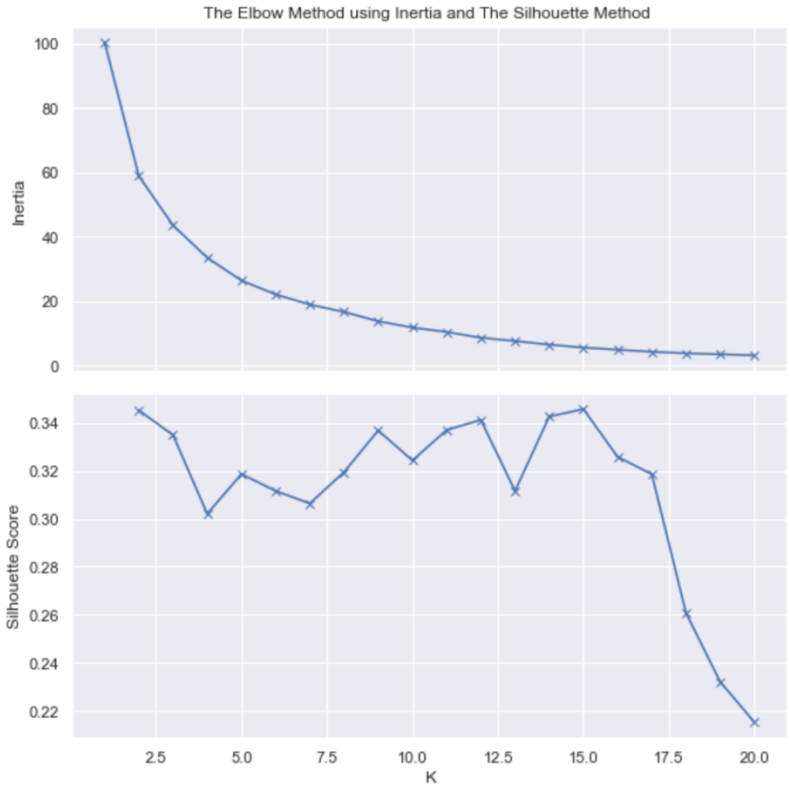
These are the inertia and silhouette score plots for KMeans using only the primary features. Based on these plots, we determine K to be 2.



Let's name our clusters group 1 and group 2. Next, we use the secondary features to further cluster the community areas within group 1 and group 2. Again, we plot inertia and silhouette scores to determine the values of K. We find both K values are equal to 2. We observe a peak at K=15 on the silhouette score plot of group 2. However, the score at K=2 is as high as that peal value.

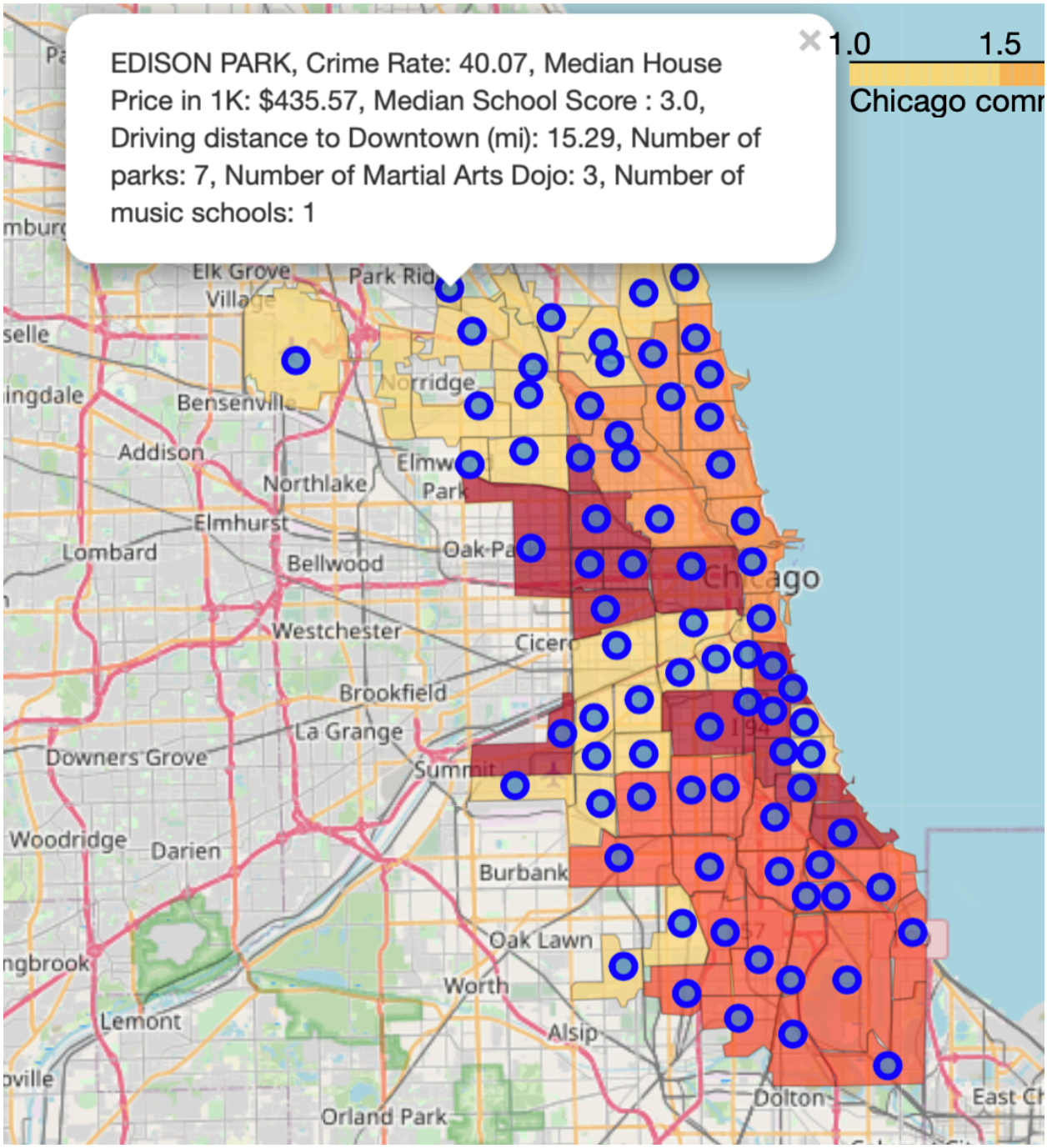
We generate a cohorpleth map of community areas colored by cluster labels and report our findings in the next section.


The Elbow Method using Inertia and The Silhouette Method
Inertia
225
200
175
150
125
100
75
50
25
Silhouette Score
0.40
0.38
0.36
0.34
2
4
6
8
10
K


The Elbow Method using Inertia and The Silhouette Method
Inertia
100
80
60
40
20
0
Silhouette Score
0.34
0.32
0.30
0.28
0.26
0.24
0.22
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
K

IV. RESULTS AND DISCUSSION

Our main result is a choropleth map where each community area is colored based on the cluster it belongs to. We also provide markers for each community area that lists the feature values when clicked on. The family can use the above map to compare community areas.



Please recall the following when interpreting these feature values:

- The crime rate is calculated by adding all crime incidents reported in the community area since the beginning of 2019 until November 28, 2020 and then dividing that number by the community area population in 1000 so that we get a rate per 1000 residents.

- The median school score is the median score of the schools within a community area. However, enrollment in public schools is address based and family should check which schools cover their candidate address. A score of 5 is highest and corresponds to **Level 1+** based on CPS school quality ranking. CPS uses several metrics when determining school rankings. We refer the reader to CPS (Chicago Public Schools) website for further information on the metrics.

- The median house prices are single family house prices as of October 31, 2020. Since this data is available in Zillow at the neighborhood area level, we took the median of the prices of the neighborhoods to come up with an estimate for the community area median house price. 77 community areas in Chicago are officially recognized. However, neighborhood area names are not official and are not consistent across different sources. For instance, 200 neighborhoods are listed in Zillow whereas Wikipedia lists 245 neighborhoods.

An analysis of the KMeans clusters reveal the following about Chicago community areas:

Yellow cluster (1):

- These community areas in general have lower crime rates (most are below 100 and all are below 145 per 1000 residents).
- The median house prices are below $450000 except for the two community areas in the shore with prices in the $640,000 (Kenwood and Hyde Park).
- The median school scores are mostly at or above 4 with a few exceptions of 3.5 (Gage Park, Clearing)
- Either the number of music schools or martial arts dojos within 2km is 0 or 1. (Jefferson Park and Portage Park are exceptions)

Orange cluster (2):

- These community areas in general have lower crime rates. Most are below 120 and the exceptions are the community areas very close to Downtown such as Loop (424.64) and Near North Side (214.99)
- The median house prices start at $400K and goes as high as $1200K. The community areas in the north west region of this cluster, Irving Park and Avondale, have the lowest house prices in this cluster.
- The median school scores are all at or above 4 with only one exception of 3.5 (Near South Side)
- Parks, music schools, and martial arts dojos are abundant in these community areas.

Dark orange cluster (3):

- The community areas in the northern part of this cluster have pretty high crime rates (Englewood: 425, Greater Grand Crossing: 378). The other community areas in this cluster have crime rates around 200 and above. The only exceptions are East Side (77.73) and Ashburn (90.95).
- The median house prices are the lowest in the city as most of them are under $150K. Only exception is Englewood with a median house price of $255K.
- The median school scores are all at or below 3 with only two exceptions of 3.5 (Englewood and Eastside)

- Although these community areas have some parks, none has music schools within 2km distance. Few of them have martial arts schools.

Darkest cluster (4):

- Crime rates are pretty high in this cluster as most community areas have rates above 200. Hermosa and Garfield Ridge are the safest community areas in this cluster with crime rates below 100.
- Median house prices are close to $200K and above going as high as $500K.
- The median school scores are 3 or below except for Humboldt Park (4) and West Garfield Park (3).
- The community areas in this cluster at the top side have in general good number of parks, music schools and martial art dojos. The southern parts of the cluster lack those venue types except for the parks.

Based on this analysis **the community areas in the Orange Cluster (2) are the best if the family can afford it.** Especially **Irving Park** and **Avondale** can be good choices as they have lower house prices. **The next best cluster is the Yellow Cluster (1).** The family can especially consider **the community areas in this cluster that are at the boundary of the Orange Cluster.**

We can also filter the values in our data frame to come up with a short list. Let's use the following criteria:

1. median house price is less than $700,000 dollars

2. crime rate less than 230 (per 1000 residents considering total number of crimes in since 2019)

3. median school score is greater than or equal to 4

4. more than 2 parks

5. at least one music school and one martial arts school

6. driving distance less than 11 miles

Here's a list of communities that match these criteria:

Community Area	Crime Rate	Median House Price	Median School Score	Driving Distance to Downtown (mi)	Number of Parks within 1km	Number of Martial Arts Schools within 2 km	Number of Music Schools within 2 km
Jefferson Park	66.73	$312,732.5	4.5	10.58	5	7	2
Irving Park	81.75	$427,607.0	4	8.38	4	9	3
Avondale	99.76	$486,231.0	4	7.19	15	12	6
West Town	141.62	$692,204.5	4	5.21	11	14	3

From this list Irving Park, Avondale, and West Town belong to the Orange Cluster. Jefferson Park belongs to the Yellow Cluster (1) and it is close to Irving Park in the Orange Cluster (2).

A possible improvement for this project is creating new features for the number of martial arts schools and music schools. Instead of exact numbers, we could look for whether there are any venues of this type within a distance. Having more of the venues presents more choices for the family. This may be desired for a family with multiple kids with different musical interests.

Possible future work involves adding more features such as most popular venue types, crime trends, and house price trends.

V. CONCLUSION

We collected and combined data from several sources. Exploring data sources, using web scraping and APIs to get our data formed a major part of this project.

In the Explatory data analysis part, we especially relied on Choropleth maps using Folium. Folium and Choropleth are amazing visualization tools that helped us make important observations about how crime rates, school rankings, house prices vary across Chicago.

Although we specifically looked into the preferences of one family, the results of this project are valuable to any person who is considering to move to Chicago and people/stakeholders who would want to know more about Chicago community areas for other purposes including city/business planning.

We shortlisted some community areas for the family that match their preferences. In addition, we are providing a map of community areas displaying clusters as well as community area specific features via labels on the map. This map is a great visual resource to compare community areas.

For clustering, we categorized our features as primary and secondary. Primary features are the ones that are more important such as crime rate, median house price, and school ranking. We first clustered the community areas into two based on primary features. Next, we further clustered the initial clusters using secondary features. We determined the value of K in KMeans clustering by plotting the silhouette score as well as the inertia. We normalized and scaled our features before clustering (for the features with skewed distributions, we took the logarithm first to normalize and then applied standard scaler to come up with the z scores).

We were able to interpret the meaning of the clusters we obtained by the KMeans algorithm. Often times, interpretation part is the hardest part of clustering. When you run any clustering algorithm with any set of compatible features, you will get clusters. However, those clusters may not have any meaning. Since there are no labels to compare to, it is really hard to know how good your clusters are. One may argue that we have metrics such as silhouette score, inertia, distortion to tell how good the clusters are but the point is in terms of how relevant the clusters are to the problem. Feature selection and scaling are important contributors to the success of the method as well.

This project had given me the opportunity to find out many great data sources and so many ideas for new projects.