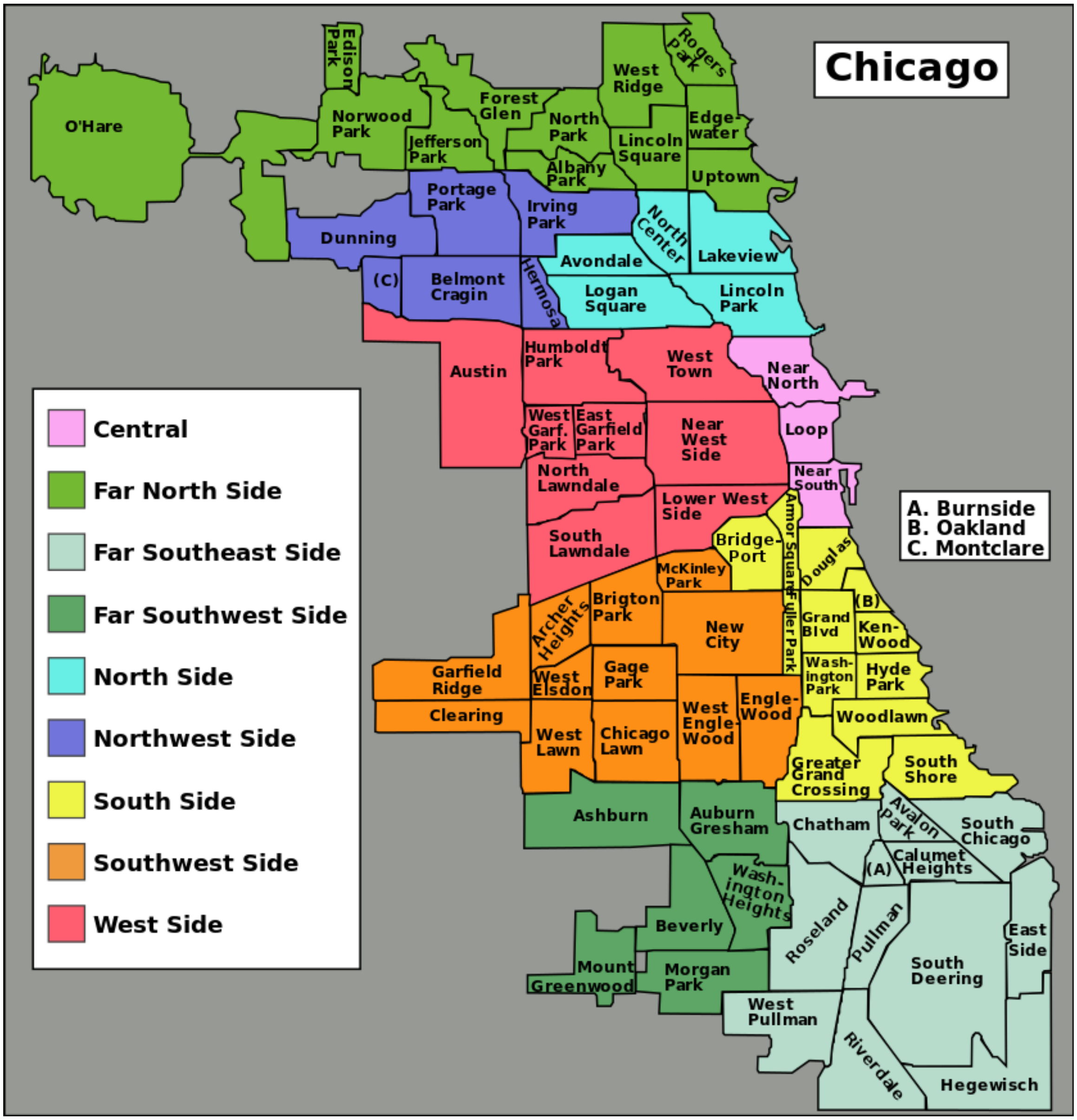


Figure 1 provided by Peter Fitzgerald, Chicago community areas map, CC BY-SA 3.0

Chicago Community Area Project

APPLIED DATA SCIENCE CAPSTONE

Ezra Witt | IBM Data Science Professional Certificate | 6/10/2020

1. Introduction to the Business Problem

1.1 BACKGROUND

Many people hate moving and even find it stressful. This is unfortunate because **moving seems to be inevitable**. In fact, the US Census Bureau states ***the average American moves 11.7 times in their lifetime***[1]. *That is a lot of moving!* The annual percentage of Americans who move is 11.2 percent.[2]

One reason people don't enjoy moving is likely due to the cost associated with it.

How much does moving cost?[3]

1. Average cost for an interstate move is ***4,100 dollars***
2. Average cost for an in-state move is ***980 dollars***
3. Each local mover hired charges, on average, ***25 dollars per hour***

1.2 PROBLEM

While we may not be able to completely avoid moving, we may be able to limit the times we move by **moving smarter**. Many movers stay in the same state and tend to not go very far at all, 40.2% of all movers relocated less than 50 miles from their old home to their new one.[4] It is my assumption that many of these people moving are realizing the neighborhood/area they were in was not ideal and they desire some change in location and/or access to local venues.

The problem I intend to address through this data analysis is to eliminate some of the reasons people tend to relocate to help people **move smarter**. The US Census question from 2015-2016 surveyed Americans as to why they moved[1], I will build upon their research to provide a model to help prospective moving families choose a neighborhood/area that will meet their desires/needs to ensure greater longevity in their new house.

For this model these are the concerns listed in the US Census Questionnaire I will attempt to curve/address:

- Wanted new or better home/apartment — 17.4%
- New job or transfer — 10.8%
- Other family reason — 10.5%
- Wanted to own a home, not rent — 5.9%
- Wanted better area/less crime — 3.1%

In order to give the model more purpose I will create a family profile with desires for their new home to better tailor the results to address the above listed concerns. The family I am profiling is a family of four with two young children. There preferences in a new home are listed in the table below.

Family Profile	**Desires/Wants**
1st Preference	Family-Friendly
2nd Preference	Safe / Low Crime
3rd Preference	Good Schools
Desired City	Chicago

1.3 INTEREST

I think that this model will have a lot of interest from the masses, with such a high number of people often moving it could address a common problem people face. My goal is to address several of the concerns above by creating a model that allows for **moving smarter**. This will help to limit those who have chosen to move because they *wanted new or better home/apartment 17.4%*, *other family reason 10.5%*, *wanted to own a home, not rent*, and/or *wanted a better area/less crime 3.1%*. While none of those reasons can be completely eliminated, especially with change over time, a better way of finding a great place to live may alleviate the need to move so often.

According to the survey 10.8% of people said they **moved due to a new job or transfer**, however, the American Moving & Storage Association[2] states that number is closer to 38% of those who move. This leads me to believe there may be interest from these corporations because if they can help in the relocation process and better ensure their relocated employees settle in, it allows the employee to focus with higher morale on the profession.

Lastly, I believe this could be **a good tool for real estate agents and companies** because the ability to give more personalized suggestions to prospective buyers may allow for quicker closings and greater customer relationships.

2. Data Overview

Based on definition of our problem, *factors that will influence our decision are*:

- The number of family friendly venues in the community (defined later)
- The number of crime incidents in the community

- The top rated schools in the community

2.1 DATA REQUIREMENTS

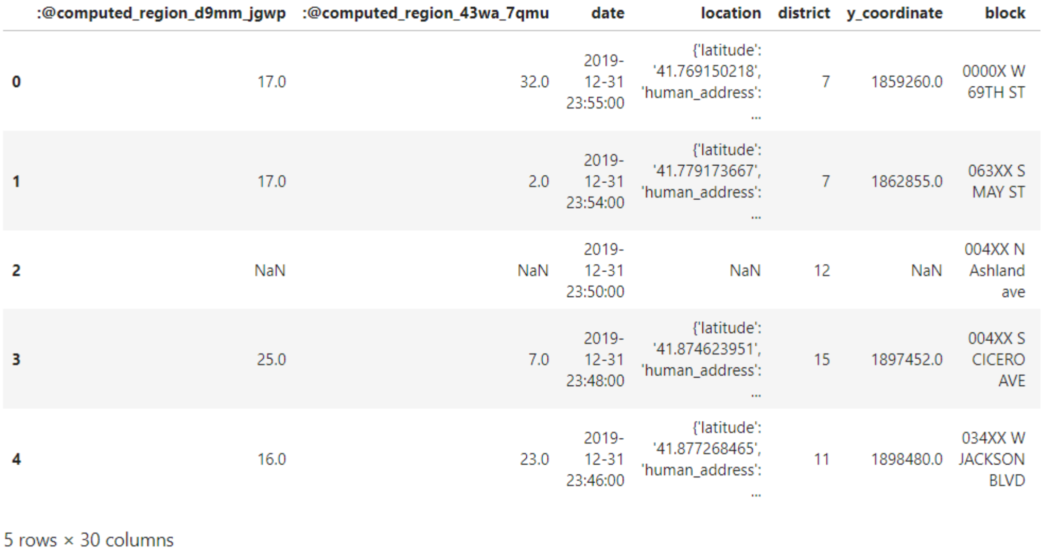
To consider the problem and address the concerns the family gave in their profile I used the following data.

- Community Areas in Chicago from the Chicago Data Portal in order to get the preliminary information on the areas of Chicago.
- Spatial Data of each community area from the Chicago Data Portal.
- The Foursquare API to get family-friendly venues of each neighborhood.
- The 2019 crime report from the Chicago Data Portal.
- School Progress Reports and Profiles from the Chicago Data Portal.

2.2 DATA COLLECTIONS

Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.

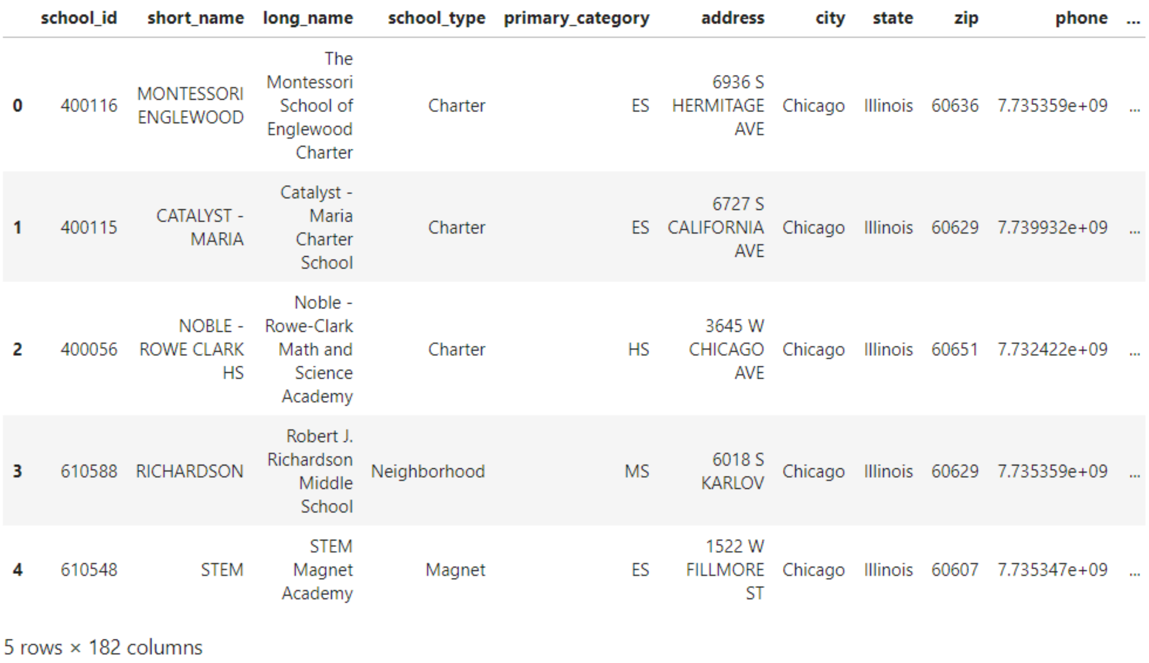
- The first data I pulled was the Chicago Crime Incidents from 2019. The dataset was available on the Chicago Data Portal. Upon reviewing the dataset, I found it contains 259,115 rows and 30 columns, I will use this data to group by neighborhood to determine what neighborhoods have large numbers of incidents and what neighborhoods have a low number of incidents.



	:@computed_region_d9mm_jgwp	:@computed_region_43wa_7qmu	date	location	district	y_coordinate	block
0	17.0	32.0	2019-12-31 23:55:00	{'latitude': '41.769150218', 'human_address': ...	7	1859260.0	0000X W 69TH ST
1	17.0	2.0	2019-12-31 23:54:00	{'latitude': '41.779173667', 'human_address': ...	7	1862855.0	063XX S MAY ST
2	NaN	NaN	2019-12-31 23:50:00	NaN	12	NaN	004XX N Ashland ave
3	25.0	7.0	2019-12-31 23:48:00	{'latitude': '41.874623951', 'human_address': ...	15	1897452.0	004XX S CICERO AVE
4	16.0	23.0	2019-12-31 23:46:00	{'latitude': '41.877268465', 'human_address': ...	11	1898480.0	034XX W JACKSON BLVD

5 rows × 30 columns

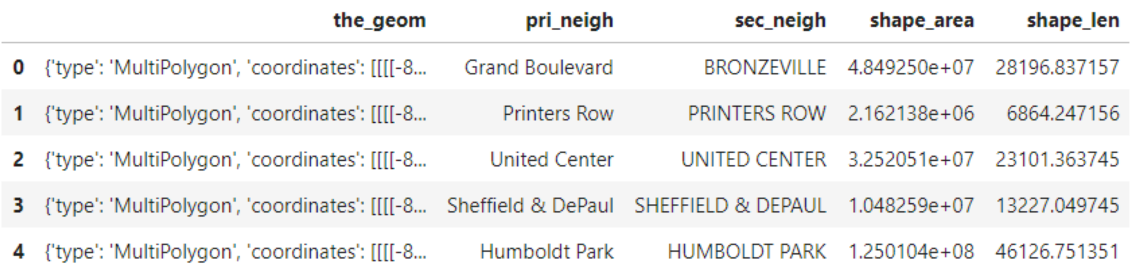
- The next data I pulled was the Chicago School Progress Reports from 2019. The dataset was available on the Chicago Data Portal. Upon reviewing the dataset I found it contains 654 rows and 182 columns, I will use this data to group by neighborhood to determine how many highly rated schools each neighborhood contains.



	school_id	short_name	long_name	school_type	primary_category	address	city	state	zip	phone	...
0	400116	MONTESSORI ENGLEWOOD	The Montessori School of Englewood Charter	Charter	ES	6936 S HERMITAGE AVE	Chicago	Illinois	60636	7.735359e+09	...
1	400115	CATALYST - MARIA	Catalyst - Maria Charter School	Charter	ES	6727 S CALIFORNIA AVE	Chicago	Illinois	60629	7.739932e+09	...
2	400056	NOBLE - ROWE CLARK HS	Noble - Rowe-Clark Math and Science Academy	Charter	HS	3645 W CHICAGO AVE	Chicago	Illinois	60651	7.732422e+09	...
3	610588	RICHARDSON	Robert J. Richardson Middle School	Neighborhood	MS	6018 S KARLOV	Chicago	Illinois	60629	7.735359e+09	...
4	610548	STEM	STEM Magnet Academy	Magnet	ES	1522 W FILLMORE ST	Chicago	Illinois	60607	7.735347e+09	...

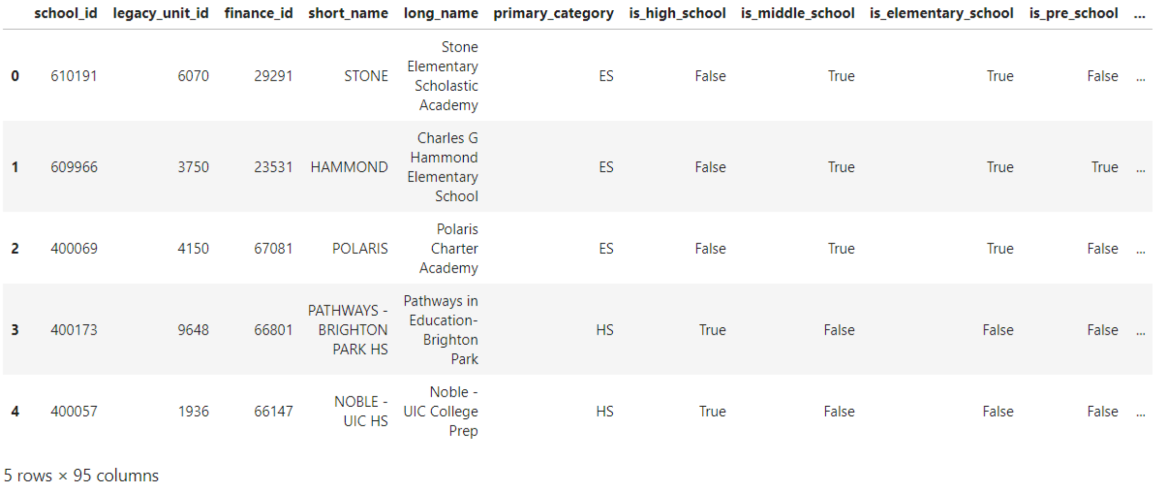
5 rows × 182 columns

- The next data I pulled was the Chicago neighborhoods geospatial data. The data was available on the Chicago Data Portal. I will use this data to create the boundaries and determine the best neighborhoods for our prospective family.



	the_geom	pri_neigh	sec_neigh	shape_area	shape_len
0	{'type': 'MultiPolygon', 'coordinates': [[[[-8...	Grand Boulevard	BRONZEVILLE	4.849250e+07	28196.837157
1	{'type': 'MultiPolygon', 'coordinates': [[[[-8...	Printers Row	PRINTERS ROW	2.162138e+06	6864.247156
2	{'type': 'MultiPolygon', 'coordinates': [[[[-8...	United Center	UNITED CENTER	3.252051e+07	23101.363745
3	{'type': 'MultiPolygon', 'coordinates': [[[[-8...	Sheffield & DePaul	SHEFFIELD & DEPAUL	1.048259e+07	13227.049745
4	{'type': 'MultiPolygon', 'coordinates': [[[[-8...	Humboldt Park	HUMBOLDT PARK	1.250104e+08	46126.751351

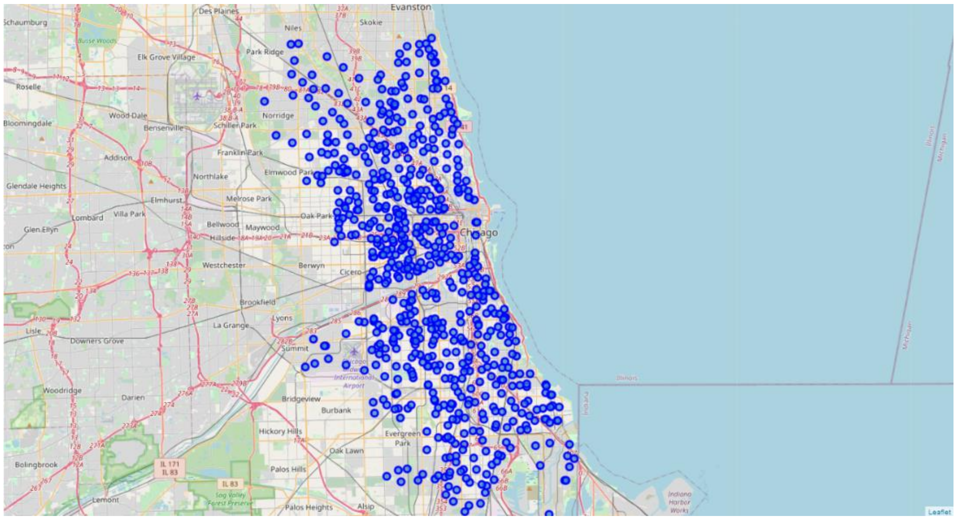
- The next data I pulled was the Chicago School Profiles Data. The dataset was available on the Chicago Data Portal. Upon reviewing the dataset I found it contains 654 rows and 95 columns, I will use this data to gather information on each school that is not included in the school progress reports.



	school_id	legacy_unit_id	finance_id	short_name	long_name	primary_category	is_high_school	is_middle_school	is_elementary_school	is_pre_school	...
0	610191	6070	29291	STONE	Stone Elementary Scholastic Academy	ES	False	True	True	False	...
1	609966	3750	23531	HAMMOND	Charles G Hammond Elementary School	ES	False	True	True	True	...
2	400069	4150	67081	POLARIS	Polaris Charter Academy	ES	False	True	True	False	...
3	400173	9648	66801	PATHWAYS - BRIGHTON PARK HS	Pathways in Education-Brighton Park	HS	True	False	False	False	...
4	400057	1936	66147	NOBLE - UIC HS	Noble - UIC College Prep	HS	True	False	False	False	...

5 rows × 95 columns

- The next data I pulled was using the Foursquare location data. I started by finding the venues with a limit of 100 within a mile radius (1609 meters), next I will define family-friendly venues and cluster based on neighborhoods.



2.3 WORKING WITH THE DATA

After I began working with the data I noticed that there wasn't one common connection between all of the data frames. Some of the data frames had zip codes, others had area numbers, while others contained the community area names. In order to create a working data set I was able to pull another data frame that included the zip codes paired with the community ID numbers. Using this new data allowed for consistency between the sets.

After finding consistency between the data frames, next I began cleaning and narrowing the data frames to only relevant rows and columns since some of the data frames were quite large. For example, one of the school data frames had 654 rows and 182 columns while the crime data from 2019 contained 259,143 rows and 31 columns.

Since our first goal was to find great schools, we were able to limit the school data frame to the school name, latitude and longitude, overall rating, rating status and zip codes. I then added the community name and ID number to the data frame for consistency across our data.

Our second goal was to find a community area with low crime, while this dataset was quite large we needed to keep the incidents in order to determine the overall crime of a given area. However, I was able to limit the columns to block, primary type, description, location-description, community area, arrest, latitude and longitude, location.

3. Methodology

3.1 INITIAL STEPS

In this project my goal is to find **the best possible community areas for a profiled family**. More specifically, ***a community area with low crime, top rated schools, and family friendly venues*** for our family to enjoy.

My initial steps were to find the data frames or create them. Once that was completed, I worked to narrow the information to those necessary to our project.

3.2 NEXT STEPS

Next, I will work to **analyze and narrow down the community areas** to the ideal locations for our family based on the preferences of what our family is looking for in a new community.

Since the family wanted to find a community area that **has low crime, top rated schools, and family friendly venues** we will seek to identify community areas which have highly rated schools and above/below average in crime and family friendly venues.

We will start by exploring our community boundaries. After we have found the defining borders we can begin working through each of our requirements and eliminate communities that do not match that criteria. i.e. ***those with high crimes, low rated schools, and a lack of family friendly venues***.

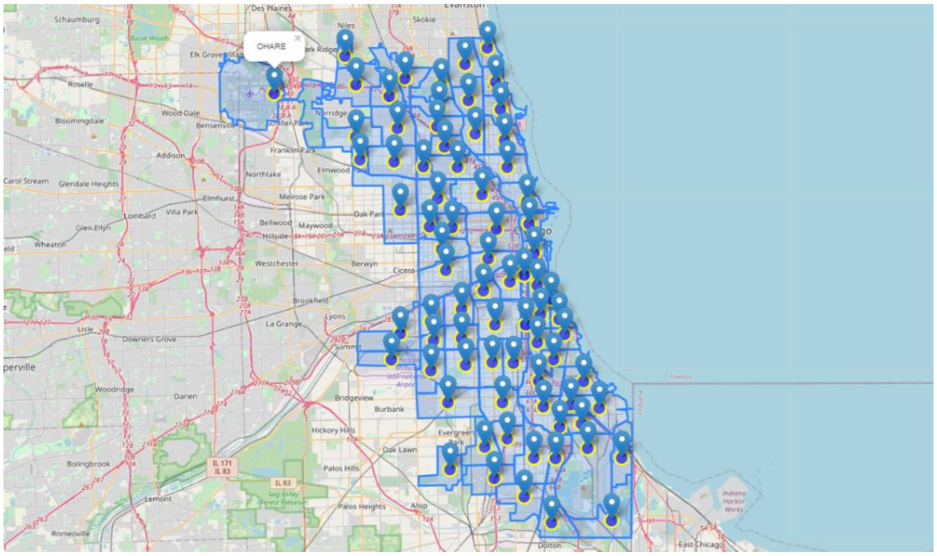
In the **final step** we will focus on narrowing down the communities to the ones that have highly rated schools and above/below average in crime and family friendly venues matching the preferences laid out by our profile family.

4 Analysis

4.1 CHICAGO COMMUNITY AREAS

The first thing I did was to define and locate each community area. Using geolocator I was able to request the longitude and latitude of Chicago. I then used the geojson provided by the Chicago Data Portal to create an overlay of each community area on a folium map.

Additionally, I took the Chicago Crime data frame and created an index of each community area and found the average longitude and latitude of each crime incident (these are the markers on the map). This will also allow for better clustering and mapping later. Altogether there are 77 community areas in Chicago, we will now begin working to narrow the 77 down to our recommended community areas for our family profile.



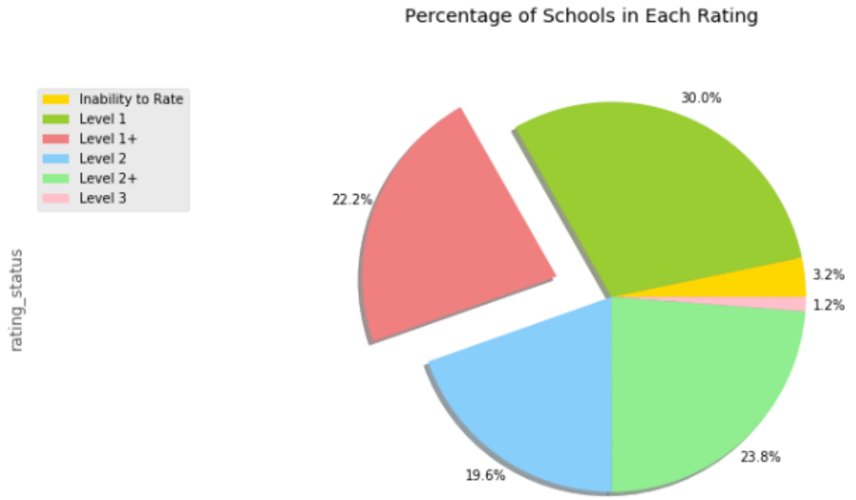
4.2 CHICAGO PUBLIC SCHOOLS

Now that we have our community areas it is now time to begin comparing each community area and find the best ones to recommend. We will start with the Chicago Public Schools. Each Chicago Public School is given an overall rating each year based on a variety of indicators. Schools are rated between Level 1+, being the best, to Level 3, being the worst. When exploring this data frame here is the breakdown of ratings:

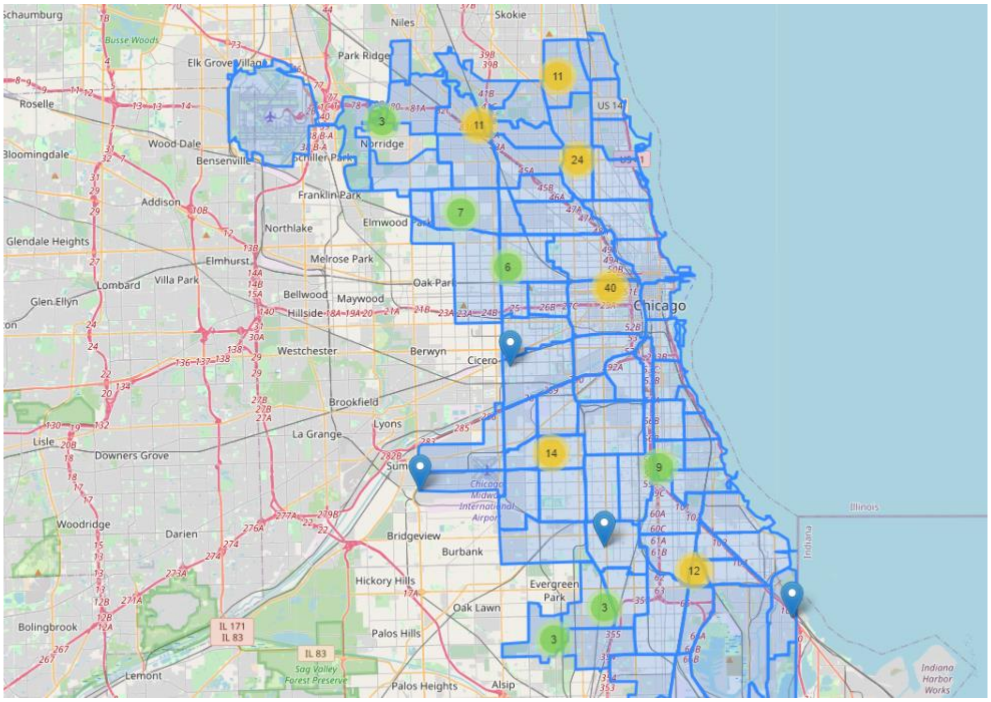
- Level 1+ (Best): 150 schools
- Level 1: 203 schools
- Level 2+: 161 schools
- Level 2: 133 schools

- Level 3: 8 schools
- Inability to Rate: 22 schools

In order to better compare these schools I created a pie chart of the percentage of schools in each rating:



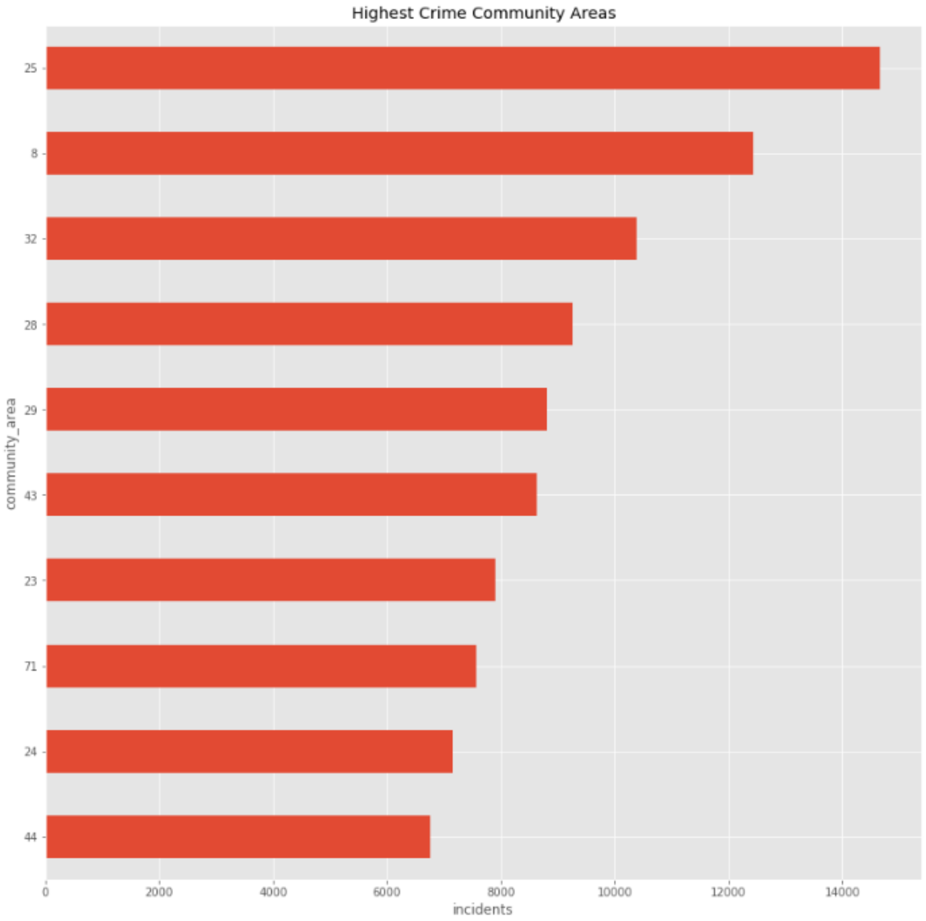
Next, I added these ratings to our data frame and keep track of how many Level 1+ schools are in each community area. Here is a clustering of the Level 1+ schools in Chicago:



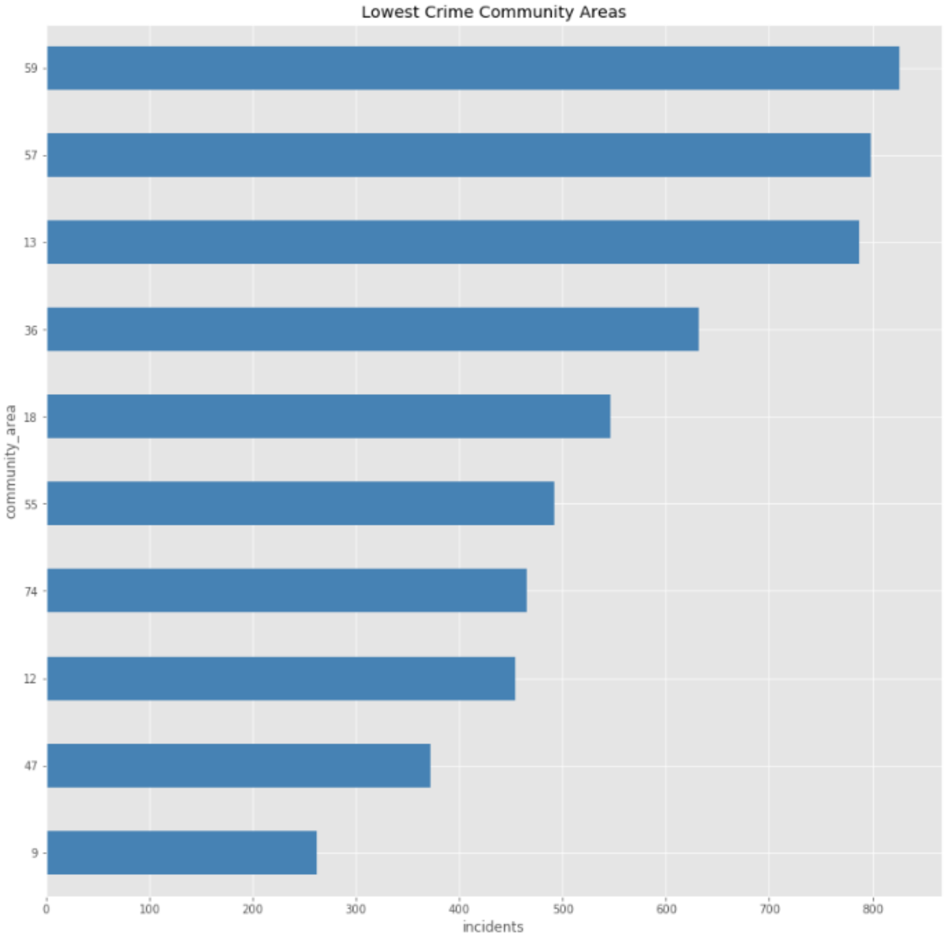
3.3 CHICAGO CRIMES

Now that we have explored the top-rated schools by community area, we can now explore the Chicago Crimes from 2019. In the data frame there were 257,990 incidents from 2019 in the city of Chicago. Since we want to find a community area with low crime, I started by finding how many crimes occurred. The crime data ranged widely with the most common crimes being: Theft (61,632 incidents), Battery (49,469 incidents), Criminal Damage (26,603 incidents), and Assault (20,597 incidents).

Next, I grouped the incidents by community area to determine the total number of incidents in each community area. The top ten highest crime communities were then visualized by community ID number on a horizontal bar graph:

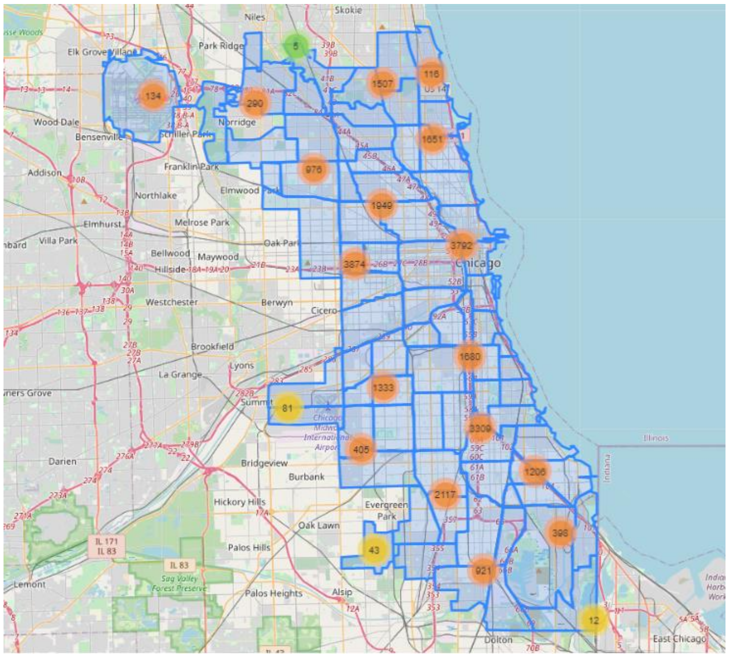


The top crime community area being Austin (25) with 14,661 crime incidents in 2019. Next, I created a horizontal bar graph of the lowest crime community areas:



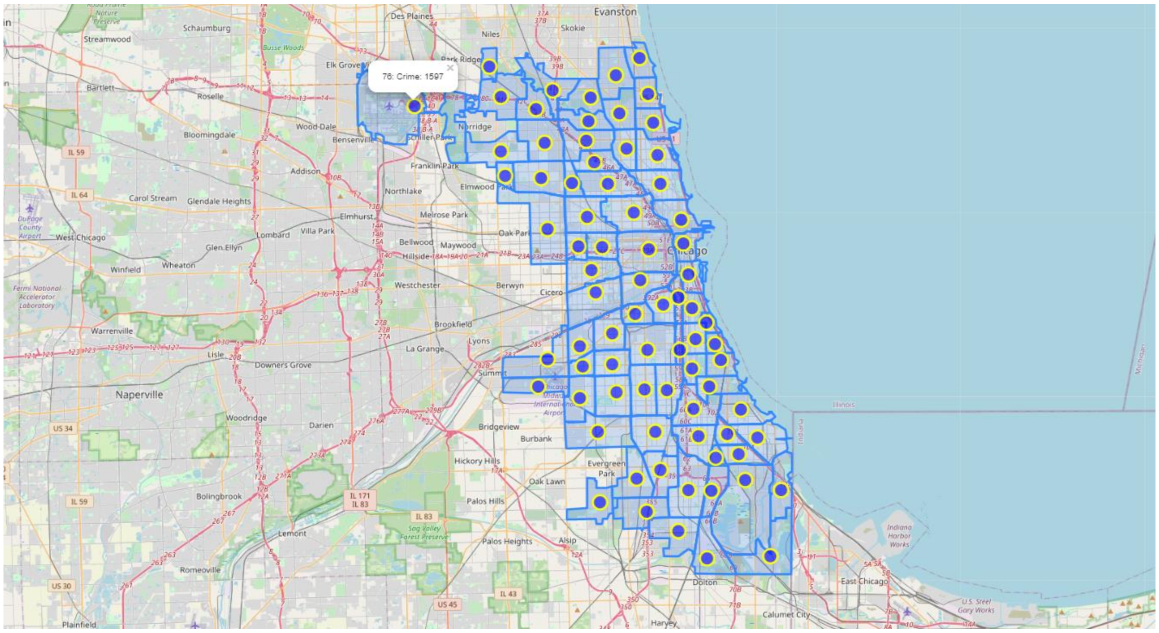
The lowest crime community area being Edison Park (9) with 262 crime incidents in 2019.

Next, I wanted to cluster the crimes to get an idea of where most were being committed but since the data frame was quite large, I created the visualization with 10% of the data so I could process and visualize it. This is the cluster map showing the 10% samples or 25,799 crime incidents:



The cluster map is helpful because we can clearly see the clusters of crimes and particular areas.

However, it does not give us a clear overview of how many crimes are in each community area. Let's visualize it again but instead let's see the counts for each community area.

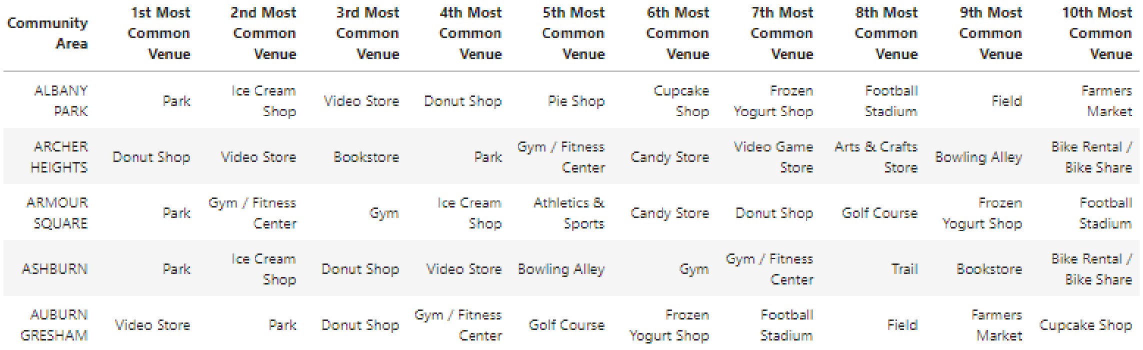


4.4 FAMILY FRIENDLY VENUES

The final criteria of the recommended community area will be based on family friendly venues. This will be defined using the Foursquare API and then listing the number of venues by community area. The first thing I did was to create a function that allowed me to run each community area and their average latitude and longitude (created earlier from crimes) through the function to get the nearby venues of each community area. For the sake of our search we set the radius to 1609 meters which is approximately 1 mile.

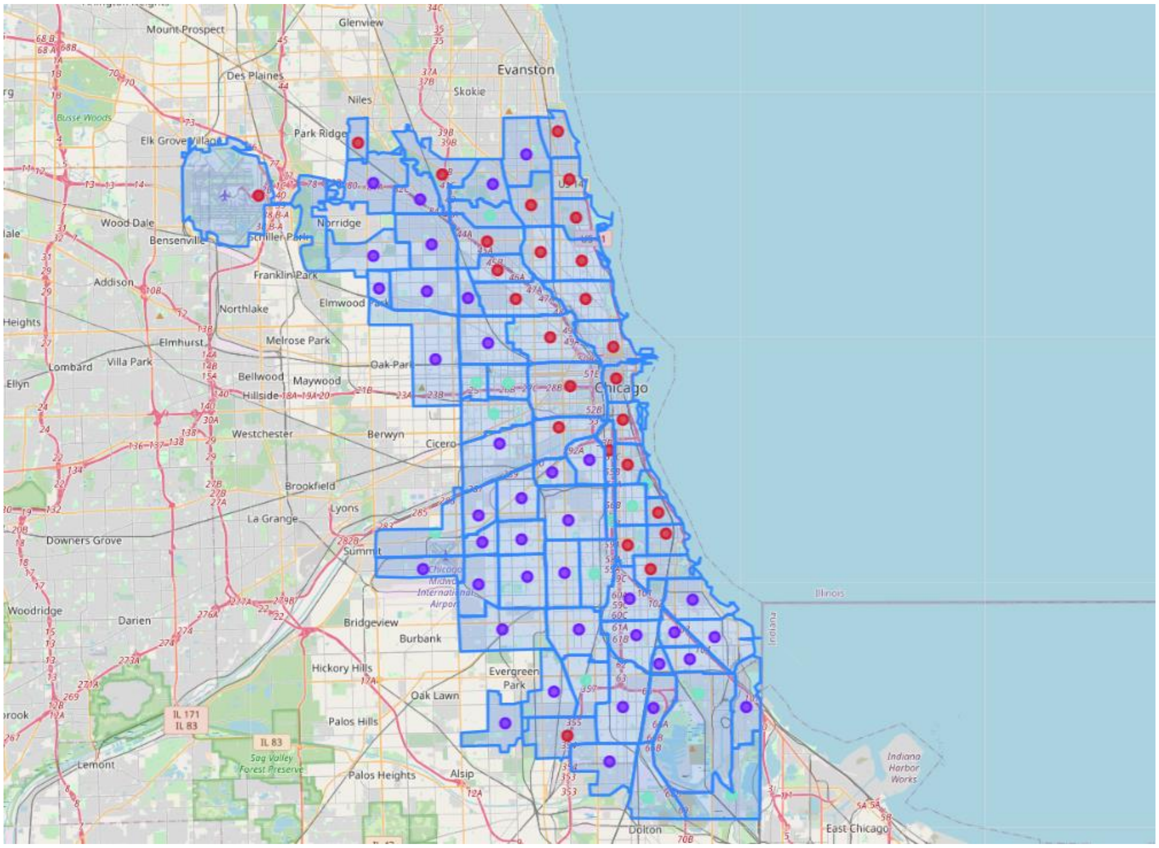
After running each venue through the function, I created a data frame and defined 41 unique venues categories that were considered family friendly. These venues were categorized into four major groups: Sports and Outdoors, Sweets & Treats, Entertainment, and Shopping. We did not include food/restaurants in our results since food categories can be difficult to determine family friendliness. For example, a restaurant that is “Italian” could be family friendly or fine dining and would be difficult to determine from the venue category.

Since we are working with categorical data, I used the One Hot Encoding process to allow for the categorical data to be easier to use and manipulate. Next, I created a function to run each community area through and determine the 10 most common venues for each community. You can view the first five rows in the next image. This data would be beneficial when presenting the recommended community areas to the family.



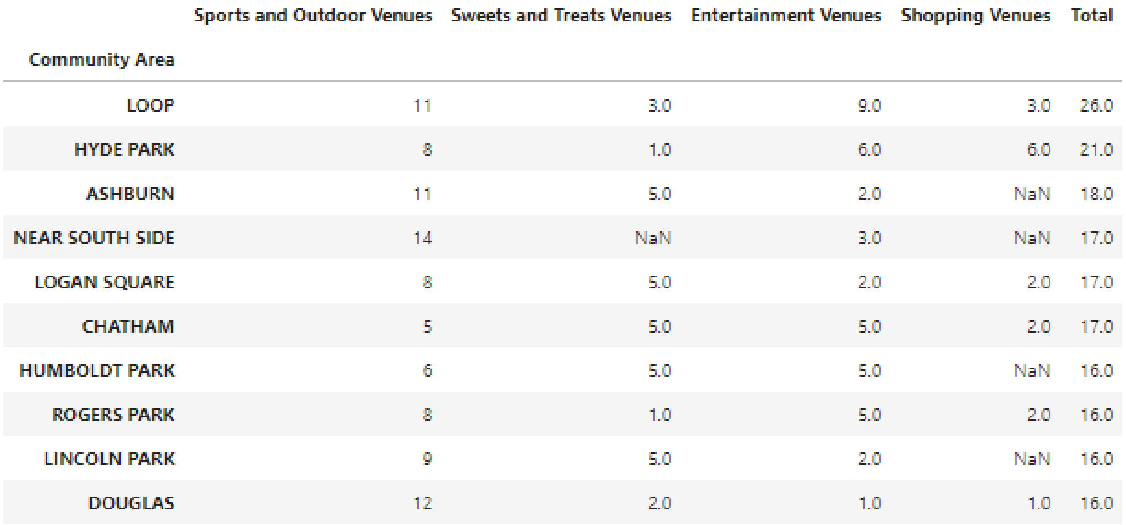
Community Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
ALBANY PARK	Park	Ice Cream Shop	Video Store	Donut Shop	Pie Shop	Cupcake Shop	Frozen Yogurt Shop	Football Stadium	Field	Farmers Market
ARCHER HEIGHTS	Donut Shop	Video Store	Bookstore	Park	Gym / Fitness Center	Candy Store	Video Game Store	Arts & Crafts Store	Bowling Alley	Bike Rental / Bike Share
ARMOUR SQUARE	Park	Gym / Fitness Center	Gym	Ice Cream Shop	Athletics & Sports	Candy Store	Donut Shop	Golf Course	Frozen Yogurt Shop	Football Stadium
ASHBURN	Park	Ice Cream Shop	Donut Shop	Video Store	Bowling Alley	Gym	Gym / Fitness Center	Trail	Bookstore	Bike Rental / Bike Share
AUBURN GRESHAM	Video Store	Park	Donut Shop	Gym / Fitness Center	Golf Course	Frozen Yogurt Shop	Football Stadium	Field	Farmers Market	Cupcake Shop

Next, I wanted to create a grouping of community areas based on the similarity and differences between them. To do this I used a machine learning algorithm called K-Means, this algorithm allows the machine to make observations and create partitions placing each community into a different set. After running the communities through, each community was placed into one of three different groups (see figure below).

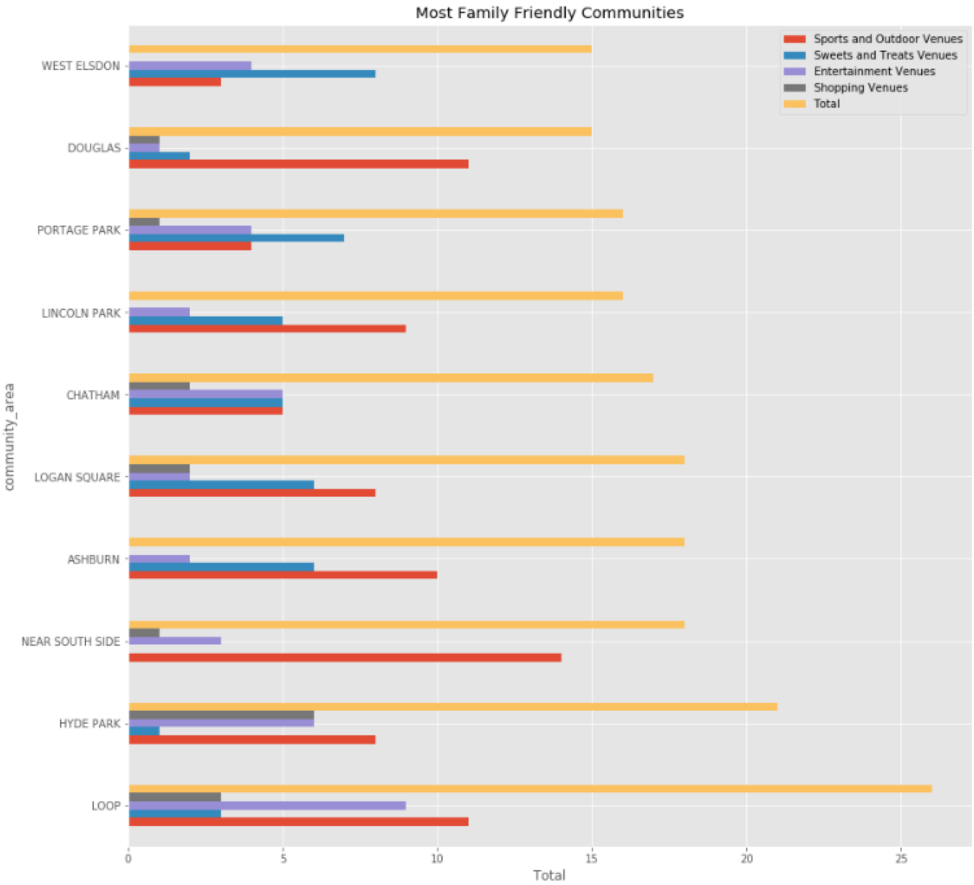


The ability to view and examine each cluster begins to allow us to see some of the differences between the groups and could potentially be used to help select a family friendly community area with a level of personality type assignment (more on this later).

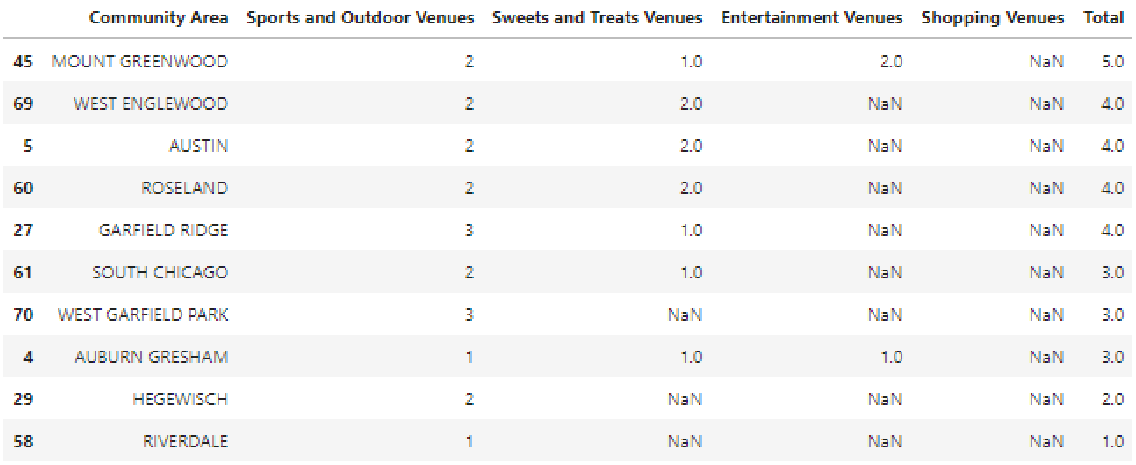
The next step was to begin to create a data frame with the number of each family friendly category. This allowed us to find the total family friendly venues while also preserving the breakdown of each type of venue. Then by sorting this dataframe we can determine the most and least family friendly communities based on the total family friendly venues in that area. The result was the following data frame and visualization:



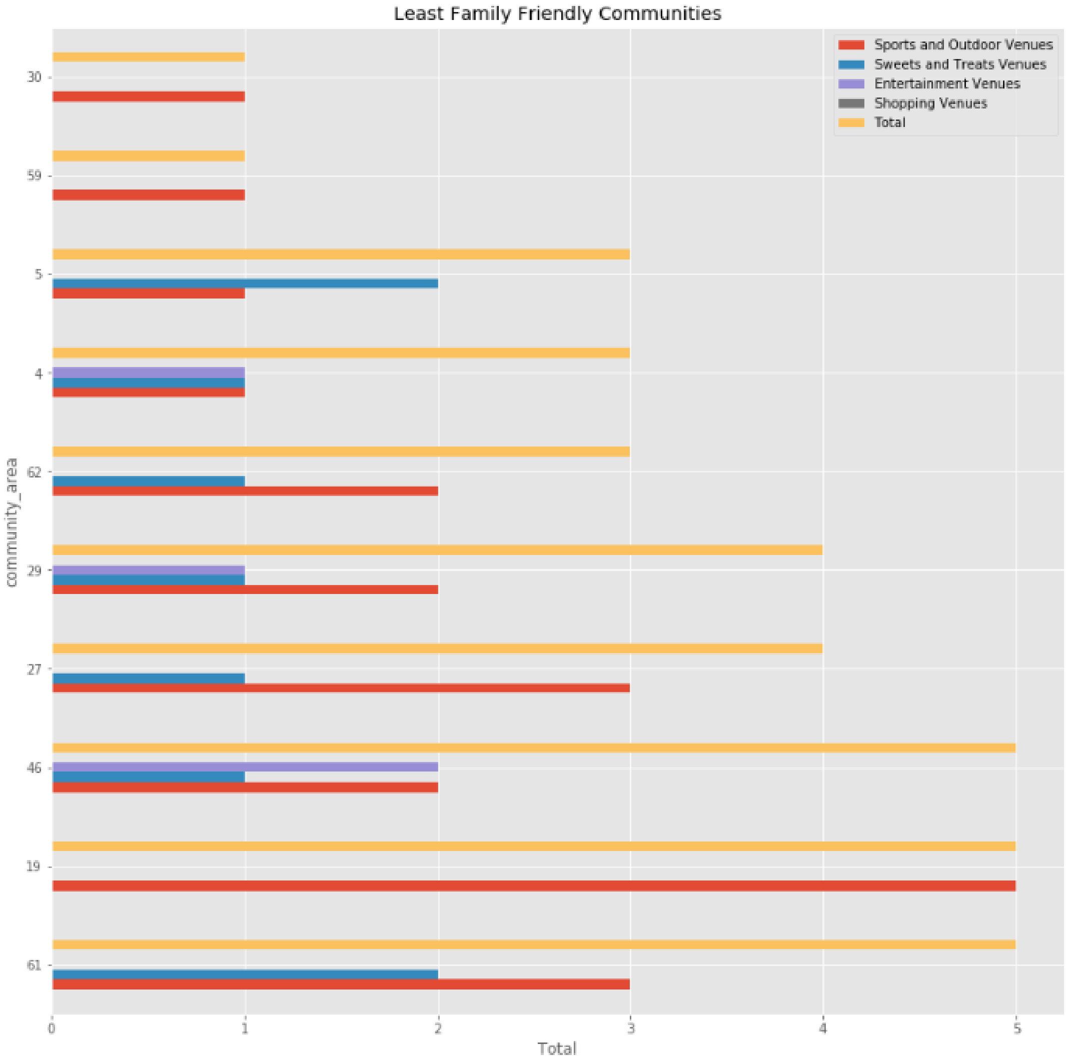
	Sports and Outdoor Venues	Sweets and Treats Venues	Entertainment Venues	Shopping Venues	Total
Community Area					
LOOP	11	3.0	9.0	3.0	26.0
HYDE PARK	8	1.0	6.0	6.0	21.0
ASHBURN	11	5.0	2.0	NaN	18.0
NEAR SOUTH SIDE	14	NaN	3.0	NaN	17.0
LOGAN SQUARE	8	5.0	2.0	2.0	17.0
CHATHAM	5	5.0	5.0	2.0	17.0
HUMBOLDT PARK	6	5.0	5.0	NaN	16.0
ROGERS PARK	8	1.0	5.0	2.0	16.0
LINCOLN PARK	9	5.0	2.0	NaN	16.0
DOUGLAS	12	2.0	1.0	1.0	16.0



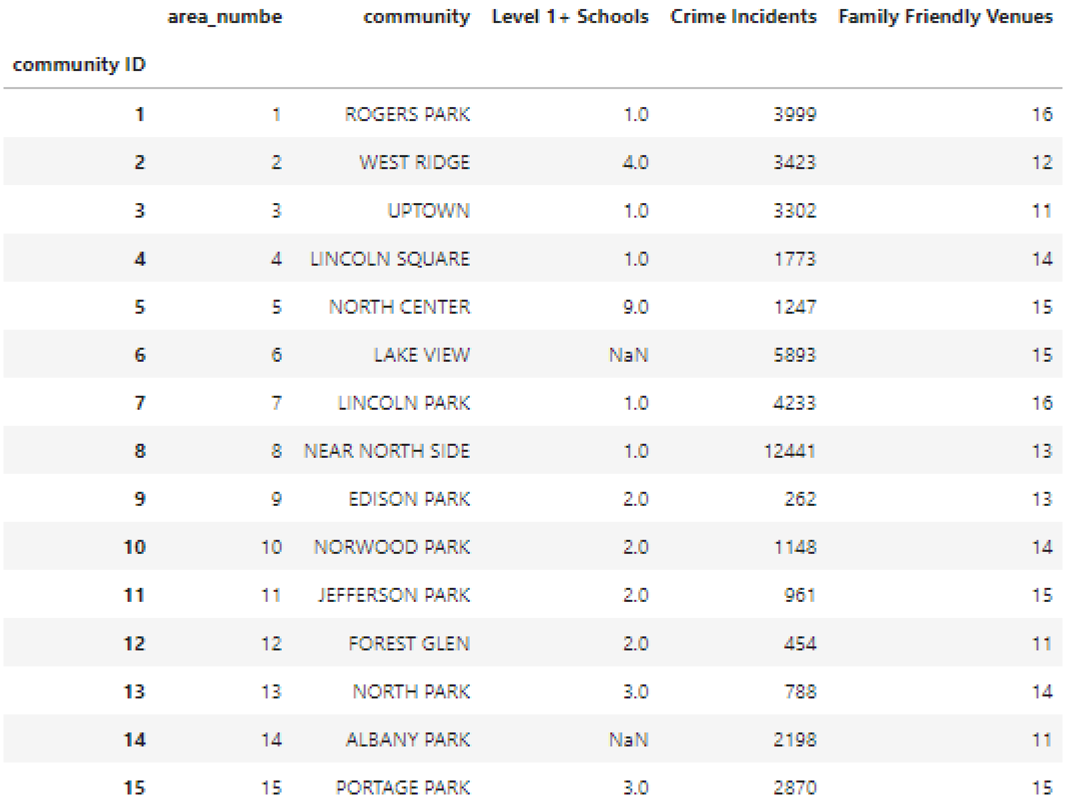
We can also sort the data frame to see the least family friendly communities and visualize the breakdown of those communities as well:



	Community Area	Sports and Outdoor Venues	Sweets and Treats Venues	Entertainment Venues	Shopping Venues	Total
45	MOUNT GREENWOOD	2	1.0	2.0	NaN	5.0
69	WEST ENGLEWOOD	2	2.0	NaN	NaN	4.0
5	AUSTIN	2	2.0	NaN	NaN	4.0
60	ROSELAND	2	2.0	NaN	NaN	4.0
27	GARFIELD RIDGE	3	1.0	NaN	NaN	4.0
61	SOUTH CHICAGO	2	1.0	NaN	NaN	3.0
70	WEST GARFIELD PARK	3	NaN	NaN	NaN	3.0
4	AUBURN GRESHAM	1	1.0	1.0	NaN	3.0
29	HEGEWISCH	2	NaN	NaN	NaN	2.0
58	RIVERDALE	1	NaN	NaN	NaN	1.0



To apply a final analysis, I then created a *final data frame* with the number of ***Top Schools, Crime,*** and ***family friendly venues*** for each community area.



	area_numbe	community	Level 1+ Schools	Crime Incidents	Family Friendly Venues
community ID					
1	1	ROGERS PARK	1.0	3999	16
2	2	WEST RIDGE	4.0	3423	12
3	3	UPTOWN	1.0	3302	11
4	4	LINCOLN SQUARE	1.0	1773	14
5	5	NORTH CENTER	9.0	1247	15
6	6	LAKE VIEW	NaN	5893	15
7	7	LINCOLN PARK	1.0	4233	16
8	8	NEAR NORTH SIDE	1.0	12441	13
9	9	EDISON PARK	2.0	262	13
10	10	NORWOOD PARK	2.0	1148	14
11	11	JEFFERSON PARK	2.0	961	15
12	12	FOREST GLEN	2.0	454	11
13	13	NORTH PARK	3.0	788	14
14	14	ALBANY PARK	NaN	2198	11
15	15	PORTAGE PARK	3.0	2870	15

Now that the data frame was created, I then used the describe attribute to determine the statistical data for each category. Through this discovery I learned that average community had ***3.2* Level 1+ schools**, ***3,350.5* crime incidents**, and ***10.61* family friendly venues**. I then used these averages to narrow down our 77 community areas to those we will recommend to our family.

4.7 FINAL ANALYSIS

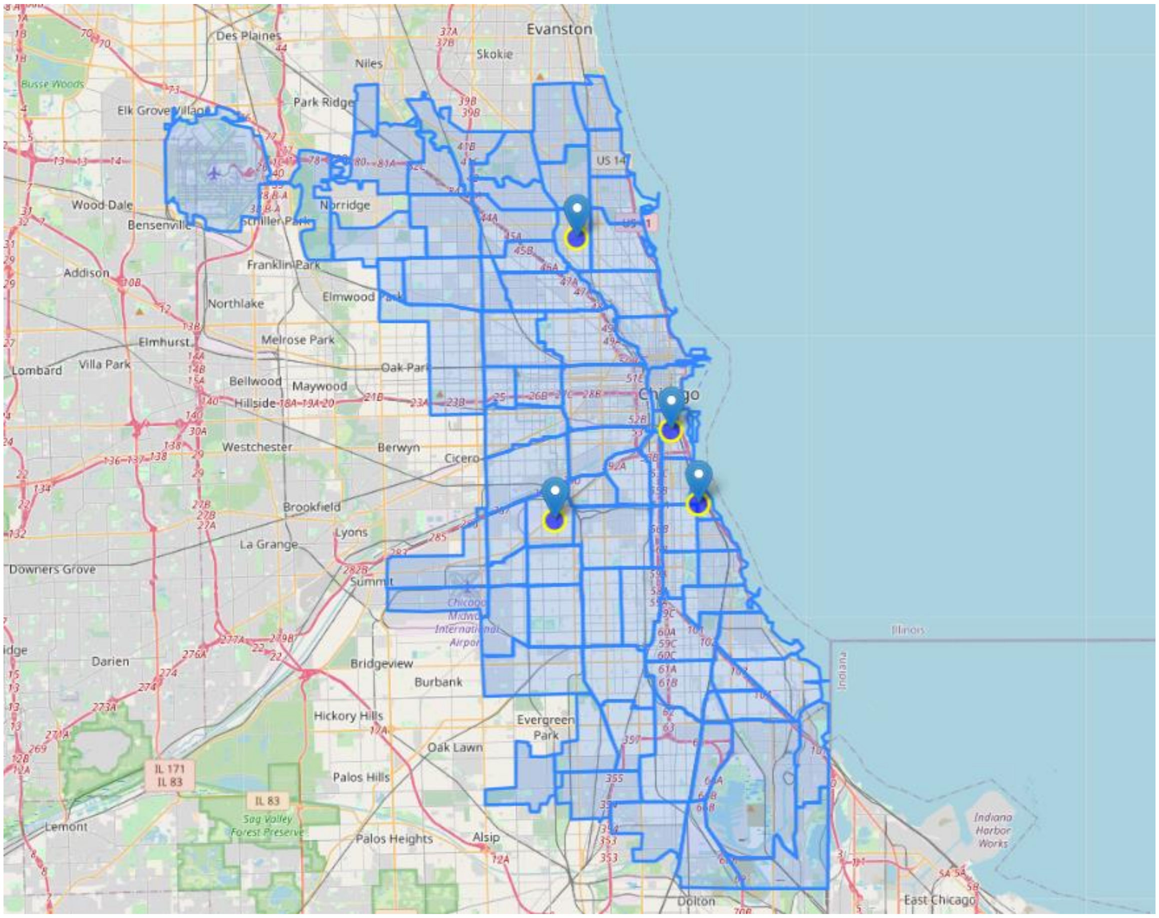
Now that we have our final data frame with our three indicators, we can now begin eliminating community areas that do not meet our ideal criteria. We will start dropping any community areas that do not have any Level 1+ rated schools (there are 31 community areas that have no Level 1+ rated schools). This leaves us with 46 potential communities to examine and compare.

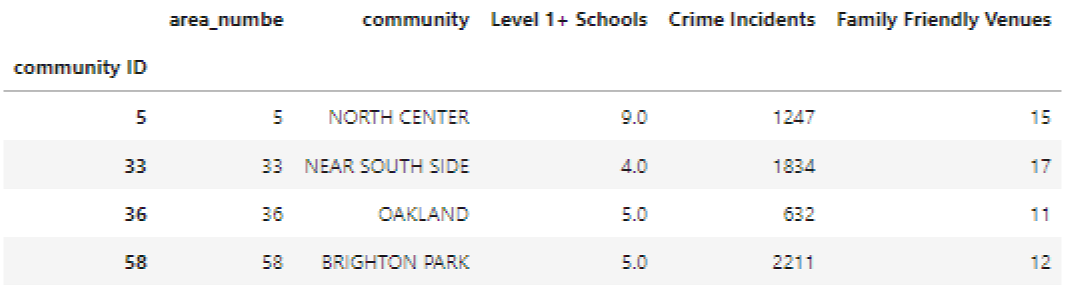
Using the remaining 46 potential communities we will now seek to find the community areas that have an above average number of Level 1+ schools. The average community had 3.2 Level 1+ schools so we will eliminate any community areas that do not have four or more Level 1+ schools. There are 30 community areas that have less than 4 Level 1+ rated schools leaving us with 16 potential community areas that have 4 or more Level 1+ rated schools.

Using the remaining 16 potential communities we will now explore which communities have below average crime. The average crime for Chicago communities was 3,350.5 incidents in 2019. For our recommended communities we will only consider those with crime incidents lower than the average. Of our 16 remaining communities there were 7 that had above average crime, leaving us with 9 potential communities.

Next, we will take the 9 remaining communities and only keep those that have an above average number of family friendly venues. The average community had 10.36 family friendly venues so we will only keep those that have 11 or more family friendly venues within a one-mile radius. After applying this new filter, we are now 4 remaining communities to recommend that have ***below average crime***, ***above average Level 1+ rated schools***, and ***above average family friendly venues.***

We can recommend the 4 community areas by location, preference (K-means clustering), or by the final counts of each of our indicators (if family places more emphasis on one over another). Here is the final data frame and map of our recommended communities.





	area_numbe	community	Level 1+ Schools	Crime Incidents	Family Friendly Venues
community ID					
5	5	NORTH CENTER	9.0	1247	15
33	33	NEAR SOUTH SIDE	4.0	1834	17
36	36	OAKLAND	5.0	632	11
58	58	BRIGHTON PARK	5.0	2211	12

5. Results and Discussion

Through our analysis we are now able to suggest four different community areas to our prospective family. Having four results allows for a little more personalization for our family and the opportunity to choose the location that is ideal for them. Presenting the information below would also allow our family to prioritize which of the measures we took is most important. For example, if crime is most important than they may lean towards Oakland who had the lowest crime counts of our final four results; or if schools are most important than the family may choose North Center who has by far the most Level 1+ rated schools in our final four.

Using K-Means we were able to sort the communities into three clusters based on the types of venues returned in the sets. Oakland is located in the first cluster, Brighton Park is located in the second cluster, and Near South Side and North Center are located in the third cluster. Since these clusters are based on the types of venues, it allows for the family to have more input on which type of community they may be most interested in.

Through our analysis of the Foursquare API we were able to identify 41 categories we deemed ***Family Friendly***. We were not able to identify specific restaurants based on categories because the Foursquare API does not distinguish through the standard search which might be family friendly and which may be fine dining for each category.

Overall, being able to find four community areas that are better than the average in all three categories is ideal because it allows for some flexibility and control from the prospective family to have an element of choice in the process.

6. Conclusion and Further Study

The results of the study are beneficial but not all encompassing. Further personalization would likely be needed to determine the best fit. For example, if our family does not consider some of the venues we choose to be family friendly it may have resulted in all things not being equal. However, the idea and exploratory of this project could still be found useful to some. Without the ability to truly test the results it leaves something to be desired, additionally the small scope of this project leaves room for expansion. These are a few of the areas that could be improved upon before using this as a full scale model.

- While these four community areas fit the three criteria we laid out, other factors, such as housing costs, types of housing in those areas, and family budgets could alter the results.

- Extending the crime scope to beyond one year of data would be necessary for a project of this nature. While the 2019 data gives us an idea, using a larger data frame with a range of years could allow us to predict up and coming family friendly community areas. This would also eliminate any possible irregularities in the data from an unusual year in a trend.

- Expanding the data to use Level 1 schools in addition to the Level 1+ schools could be necessary. I am not clear on all of the differences between a level 1 and level 1+ school but some families may prefer we leave those schools in the model.

I enjoyed working on this project and hope you find it interesting in the least. If you have experience with or live in Chicago I would be interested to hear your feedback on these community areas (while they all meet better than average on our three conditions public perception and other factors may give an alternative point of view).

References:

1. *US Census Bureau*, Calculating Migration Expectancy Using ACS Data

2. *American Moving & Storage Association*, About Our Industry

3. *moveBuddha*, Moving Cost Calculator

4. *Porch*, Mover's Remorse