Finding Safe Neighborhoods in Chicago,IL

Jennifer Rajkumar

A. Introduction:

A friend of mine, a Software Developer, recently been offered a job at Chicago, IL. As this young family, with two small children and a dog, prepares their way to Chicago, they are struck with two important questions as with anyone who plans to move to a new location.
The First question is to find the safe Neighborhoods where the crime rates are lower and the Second most important question is to find the safest Schools in that neighborhoods for their children.
Another important consideration for this young family is, they want to be close to parks, cafes and Restaurants.

About Chicago,IL:
Chicago, IL is a Vibrant, Multicultural city that thrives on the harmony and diversity of its Neighborhoods. It's the Third largest city in the United States with a population of nearly three million people. Chicago is home to about 100 Neighborhoods, 600 parks, more than 7,300 restaurants, etc.. NO wonder it will be one of the best places on the planet to raise a family.

The Important Questions to Answer:

1. ***Can we be able to find a safest Neighborhood(s) in Chicago,IL where the Crime rates are low?.***
2. ***Can we find the safest School(s) in that Neighborhood?***
3. ***Can we find the Neighborhoods where the above two conditions are met, pact with venues which this Family could enjoy?.***

B. Data acquisition and description:

1. Data Sources:

1. The **datasets** used in this project are.,
   . Crimes data from **Kaggle** which will be used to find the Neighborhoods with No major crimes committed i.e No arrests were made in the neighbourhoods of Chicago,IL.
   . Chicago Public Schools Report which I scrapped from Chicago District data will be used to find the safest schools in the Neighbourhoods of Chicago,IL.

2. **Folium** - a Python visualization library will be used in this project to view the neighborhoods before and after clustering using an **interactive leaflet map**.

3. **Foursquare API** : Foursquare is a location data provider where you get all sorts of information about the venues, events,etc.. around the area of interest. I used Foursquare API to get the details of nearby venues around the neighborhood.

2. Feature selection and Data description :

The main objective of the family is to live and raise a family in a safe Neighborhood. The District of Chicago School Quality Rating Policy measures each school by their performance and the safety protocols implemented by them in case of an emergency situation. In this project, I used the **Chicago Public Schools dataset**, to get the Name of the schools, and their safety score, the Neighborhoods where the schools are and their geospatial Coordinates. Similarly, I used the **Crimes data** from Kaggle, to find the Neighborhoods where there are No arrest made(signaling it is safe to live there), and their geospatial coordinates. Once the desired features are selected and datasets are cleaned, I merged them Together, to display the final dataset with the names of the safe schools and the neighborhoods and their geospatial coordinates.

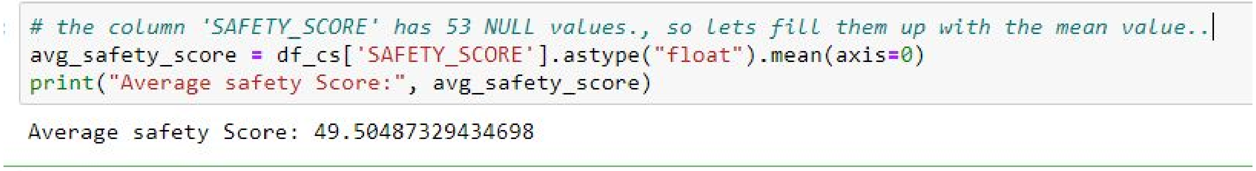
I will then use the **Foursquare API** to get the names of the venues and their categories near those neighborhoods so the family could find not only safe but a fun place to live in.

C. Methodology:

1. Data Analysis:

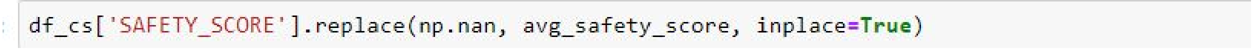
a. Analysis of Chicago Public school dataset:

Since the family's goal is to find the safe school for their children, I focused on the highest Safety scores that the schools got in a neighborhood. The Redundant features like the schools policy information, per grade information, college enrollment information will be ignored as these information are not pertaining to this project. **Pandas library** will be used extensively for the analysis of this dataset. The missing values from the Safety scores feature are dealt with by taking the average of that feature.



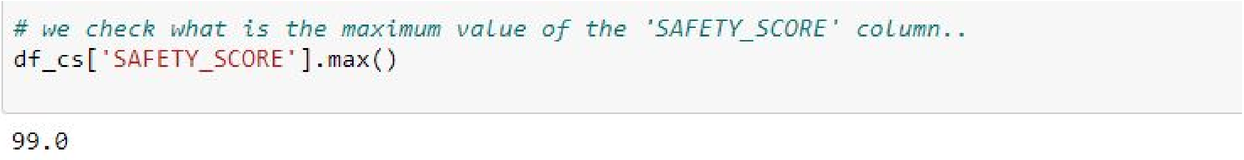
```
# the column 'SAFETY_SCORE' has 53 NULL values., so lets fill them up with the mean value..
avg_safety_score = df_cs['SAFETY_SCORE'].astype("float").mean(axis=0)
print("Average safety Score:", avg_safety_score)
```

```
Average safety Score: 49.50487329434698
```



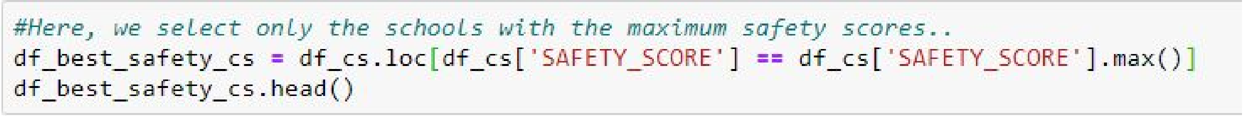
```
df_cs['SAFETY_SCORE'].replace(np.nan, avg_safety_score, inplace=True)
```

The maximum value of the safety score feature is chosen to find the safest schools.



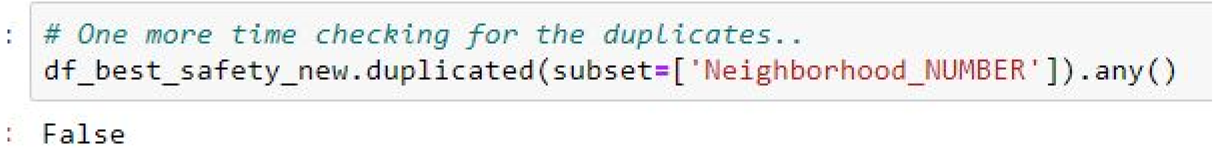
```
# we check what is the maximum value of the 'SAFETY_SCORE' column..
df_cs['SAFETY_SCORE'].max()
```

```
99.0
```



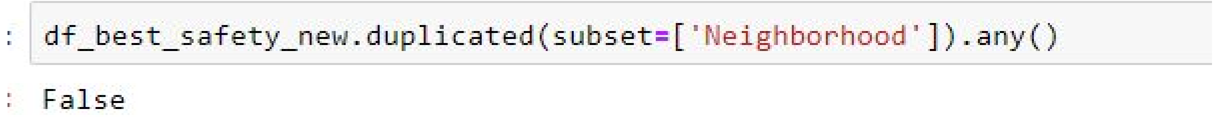
```
#Here, we select only the schools with the maximum safety scores..
df_best_safety_cs = df_cs.loc[df_cs['SAFETY_SCORE'] == df_cs['SAFETY_SCORE'].max()]
df_best_safety_cs.head()
```

The duplicates from the dataset features are checked and removed.



```
# One more time checking for the duplicates..
df_best_safety_new.duplicated(subset=['Neighborhood_NUMBER']).any()
```

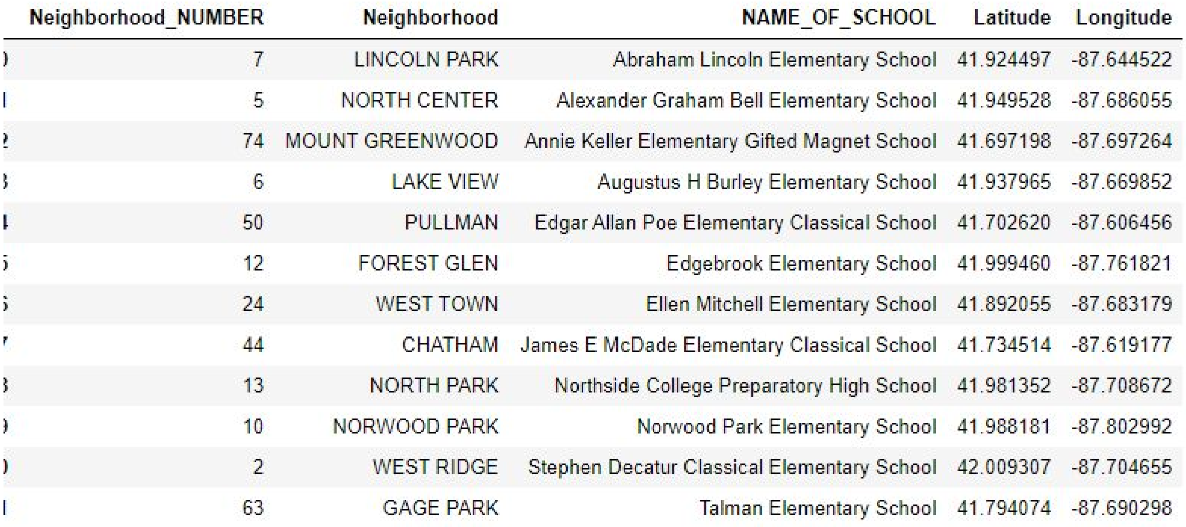
```
False
```



```
df_best_safety_new.duplicated(subset=['Neighborhood']).any()
```

```
False
```

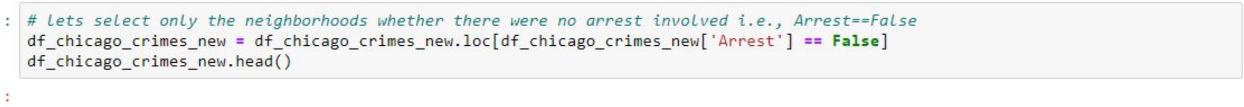
And the dataset is checked (more than one time) for any Null values in the selected features which might affect the result. The Final dataset which represents the Schools with highest Safety score and the Neighborhood where they are at and their latitude and Longitude



Neighborhood_NUMBER	Neighborhood	NAME_OF_SCHOOL	Latitude	Longitude
7	LINCOLN PARK	Abraham Lincoln Elementary School	41.924497	-87.644522
5	NORTH CENTER	Alexander Graham Bell Elementary School	41.949528	-87.686055
74	MOUNT GREENWOOD	Annie Keller Elementary Gifted Magnet School	41.697198	-87.697264
6	LAKE VIEW	Augustus H Burley Elementary School	41.937965	-87.669852
50	PULLMAN	Edgar Allan Poe Elementary Classical School	41.702620	-87.606456
12	FOREST GLEN	Edgebrook Elementary School	41.999460	-87.761821
24	WEST TOWN	Ellen Mitchell Elementary School	41.892055	-87.683179
44	CHATHAM	James E McDade Elementary Classical School	41.734514	-87.619177
13	NORTH PARK	Northside College Preparatory High School	41.981352	-87.708672
10	NORWOOD PARK	Norwood Park Elementary School	41.988181	-87.802992
2	WEST RIDGE	Stephen Decatur Classical Elementary School	42.009307	-87.704655
63	GAGE PARK	Talman Elementary School	41.794074	-87.690298

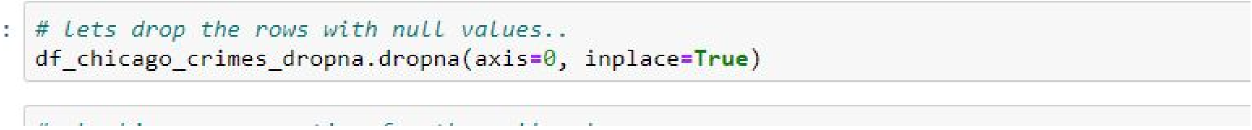
b. Analysis of Crimes dataset:

Another important criteria for the family is to live in safe neighborhood, so I focused on the **Crimes** dataset to find the neighborhoods where there are **No arrests** involved. Again I used **Pandas Library** extensively for the analysis of this dataset. The redundant features like Location and Description of the crime, Beat, District,Ward etc., are omitted since they are not pertaining to this project.
The dataset is selected based on **No Arrest** value



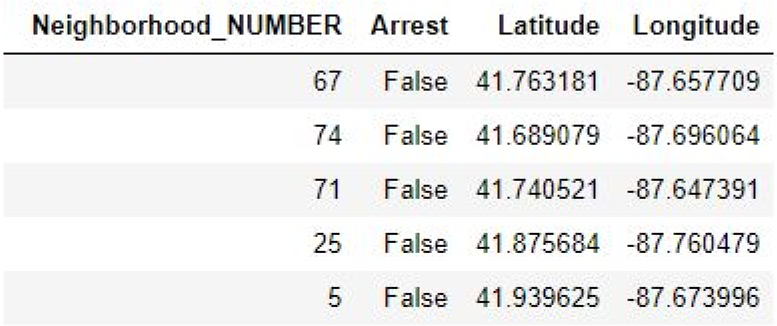
```
# lets select only the neighborhoods whether there were no arrest involved i.e., Arrest==False
df_chicago_crimes_new = df_chicago_crimes_new.loc[df_chicago_crimes_new['Arrest'] == False]
df_chicago_crimes_new.head()
```

The Null values in the dataset were removed as they contribute less than 5% of the entire dataset



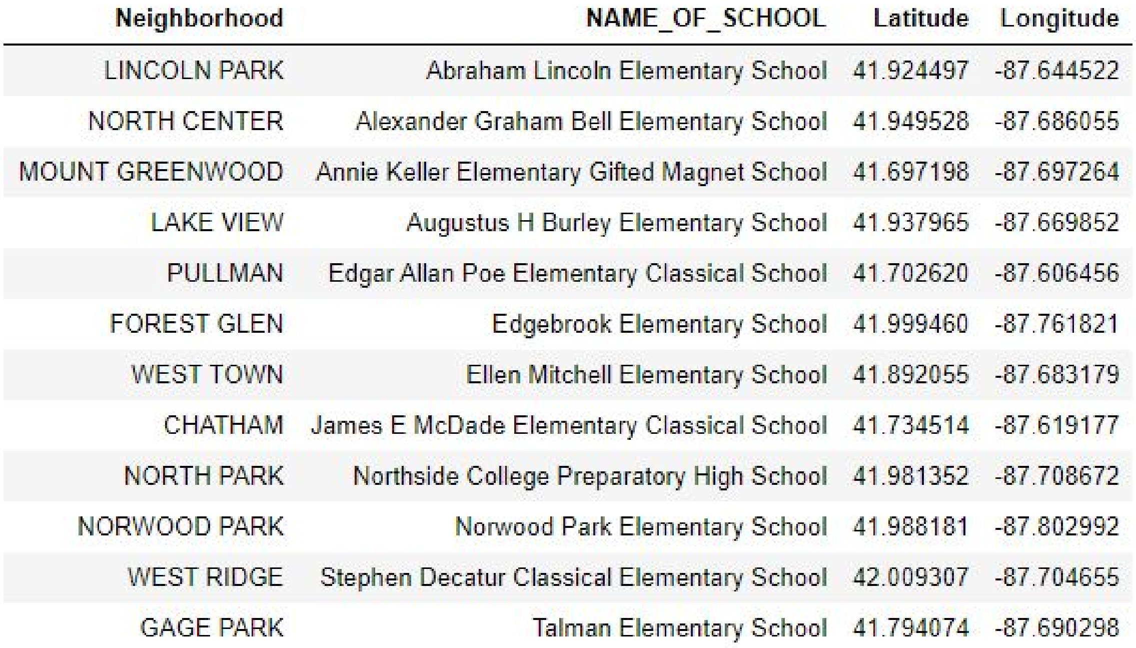
```
# lets drop the rows with null values..
df_chicago_crimes_dropna.dropna(axis=0, inplace=True)
```

After the dataset is checked for duplicates and removed, it represents the Safe Neighbourhoods with their Latitude and Longitude values.



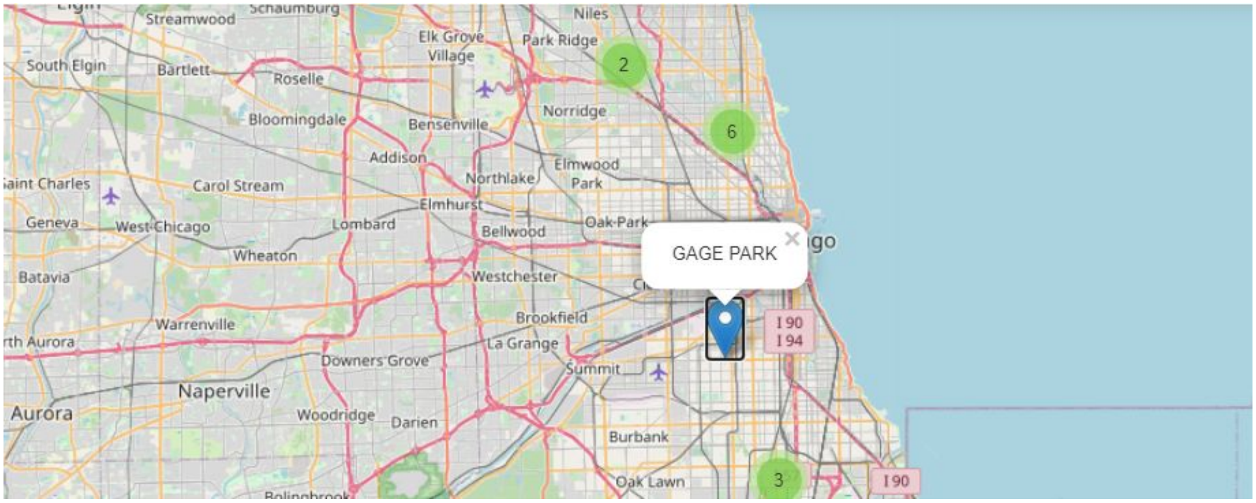
Neighborhood_NUMBER	Arrest	Latitude	Longitude
67	False	41.763181	-87.657709
74	False	41.689079	-87.696064
71	False	41.740521	-87.647391
25	False	41.875684	-87.760479
5	False	41.939625	-87.673996

Finally, I merged the above cleaned Chicago Public School dataset and the Crimes Dataset (using **Pandas pd.merge()**) to get the Neighborhoods and their geospatial coordinates that represents Not only the safe Places for the family to live in but also the Safe schools that the children can attend to.



Neighborhood	NAME_OF_SCHOOL	Latitude	Longitude
LINCOLN PARK	Abraham Lincoln Elementary School	41.924497	-87.644522
NORTH CENTER	Alexander Graham Bell Elementary School	41.949528	-87.686055
MOUNT GREENWOOD	Annie Keller Elementary Gifted Magnet School	41.697198	-87.697264
LAKE VIEW	Augustus H Burley Elementary School	41.937965	-87.669852
PULLMAN	Edgar Allan Poe Elementary Classical School	41.702620	-87.606456
FOREST GLEN	Edgebrook Elementary School	41.999460	-87.761821
WEST TOWN	Ellen Mitchell Elementary School	41.892055	-87.683179
CHATHAM	James E McDade Elementary Classical School	41.734514	-87.619177
NORTH PARK	Northside College Preparatory High School	41.981352	-87.708672
NORWOOD PARK	Norwood Park Elementary School	41.988181	-87.802992
WEST RIDGE	Stephen Decatur Classical Elementary School	42.009307	-87.704655
GAGE PARK	Talman Elementary School	41.794074	-87.690298

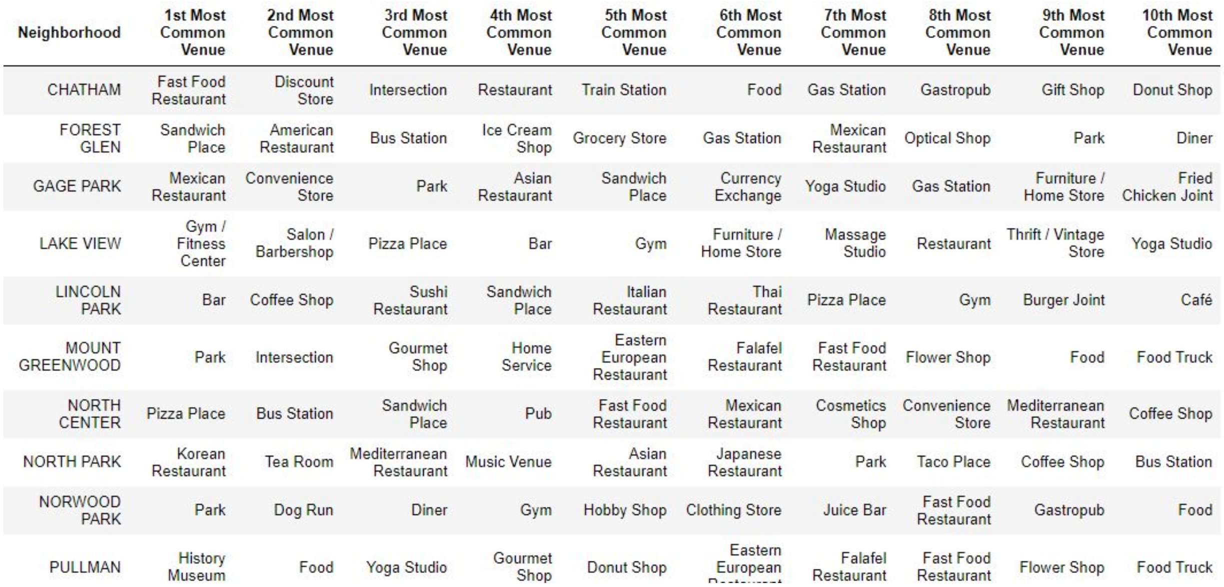
C. Initial Visualization of the Neighborhoods using Folium:



d. Foursquare API application:

Since the Family has small children and a dog they want to live in a Neighborhood that is pact with Parks, Restaurants, Cafes, Museums,etc., So I used my Foursquare credentials, to connect with **Foursquare API** to find out venues around each Neighborhood. I set the radius to 500m around each neighborhood. **Foursquare** returned **117** unique venue categories around these neighborhoods. Then I used the **Foursquare** to get the top 10 venues around each neighborhood which shows in the below table.

.



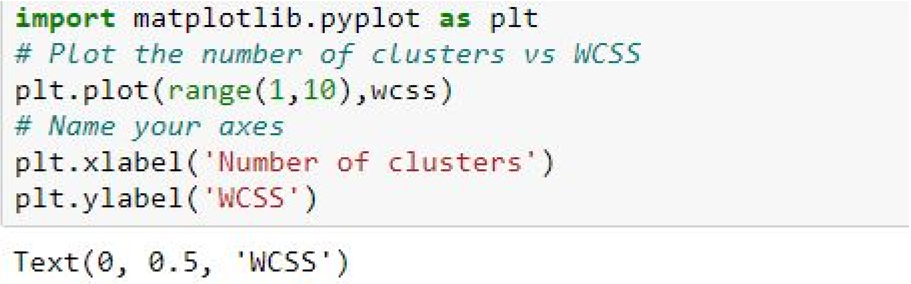
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
CHATHAM	Fast Food Restaurant	Discount Store	Intersection	Restaurant	Train Station	Food	Gas Station	Gastropub	Gift Shop	Donut Shop
FOREST GLEN	Sandwich Place	American Restaurant	Bus Station	Ice Cream Shop	Grocery Store	Gas Station	Mexican Restaurant	Optical Shop	Park	Diner
GAGE PARK	Mexican Restaurant	Convenience Store	Park	Asian Restaurant	Sandwich Place	Currency Exchange	Yoga Studio	Gas Station	Furniture / Home Store	Fried Chicken Joint
LAKE VIEW	Gym / Fitness Center	Salon / Barbershop	Pizza Place	Bar	Gym	Furniture / Home Store	Massage Studio	Restaurant	Thrift / Vintage Store	Yoga Studio
LINCOLN PARK	Bar	Coffee Shop	Sushi Restaurant	Sandwich Place	Italian Restaurant	Thai Restaurant	Pizza Place	Gym	Burger Joint	Café
MOUNT GREENWOOD	Park	Intersection	Gourmet Shop	Home Service	Eastern European Restaurant	Falafel Restaurant	Fast Food Restaurant	Flower Shop	Food	Food Truck
NORTH CENTER	Pizza Place	Bus Station	Sandwich Place	Pub	Fast Food Restaurant	Mexican Restaurant	Cosmetics Shop	Convenience Store	Mediterranean Restaurant	Coffee Shop
NORTH PARK	Korean Restaurant	Tea Room	Mediterranean Restaurant	Music Venue	Asian Restaurant	Japanese Restaurant	Park	Taco Place	Coffee Shop	Bus Station
NORWOOD PARK	Park	Dog Run	Diner	Gym	Hobby Shop	Clothing Store	Juice Bar	Fast Food Restaurant	Gastropub	Food
PULLMAN	History Museum	Food	Yoga Studio	Gourmet Shop	Donut Shop	Eastern European	Falafel Restaurant	Fast Food Restaurant	Flower Shop	Food Truck

2. Modeling:

When I comb through the table returned by the Foursquare, i see a lot of similar venues exists between the neighborhoods, that is why I choose **KMeans Clustering** algorithm to cluster those neighborhoods based on their similarities. KMeans can arrange data only **unsupervised**, and can group data based on their similarities.

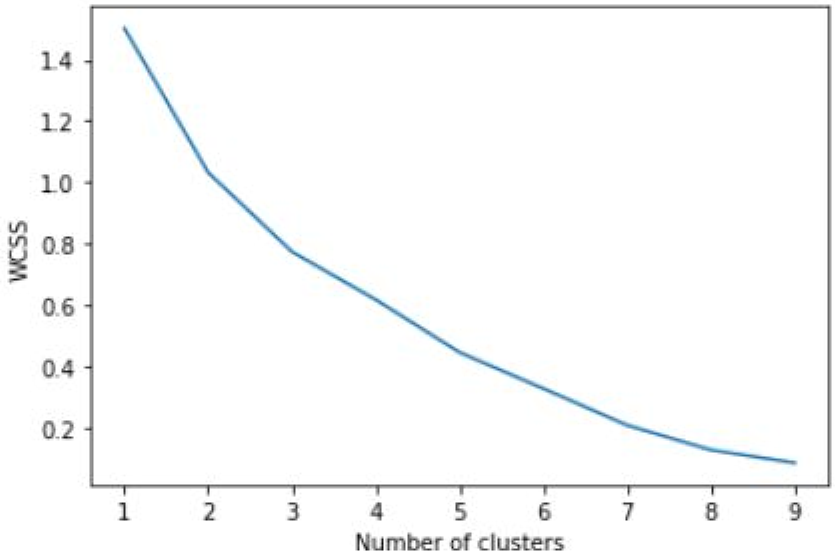
Instead of randomly choosing the Number of cluster I want to use **Elbow Method** to see what is the optimal number of clusters that the KMeans algorithm can create. Selecting the optimal number of clusters involved calculating the **WCSS** for each number of clusters. **WCSS (Within Cluster Sum Of Squares)** is a measure developed within the **ANOVA** Framework and it is calculated using **KMeans.inertia_** method. This will give us the number of clusters where WCSS is drastically reduced (The ELbow point) and remains almost unchanged as the number of clusters increase.

Using the Matplotlib.Pyplot Library we can plot the **WCSS** against the number of clusters



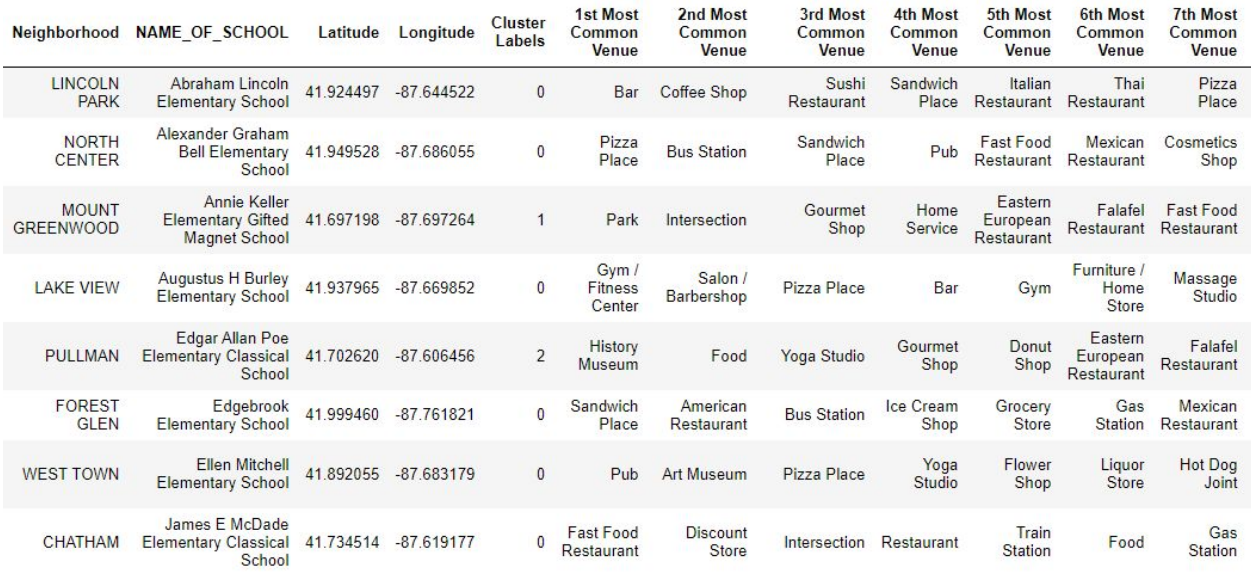
```
import matplotlib.pyplot as plt
# Plot the number of clusters vs WCSS
plt.plot(range(1,10),wcss)
# Name your axes
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
```

```
Text(0, 0.5, 'WCSS')
```



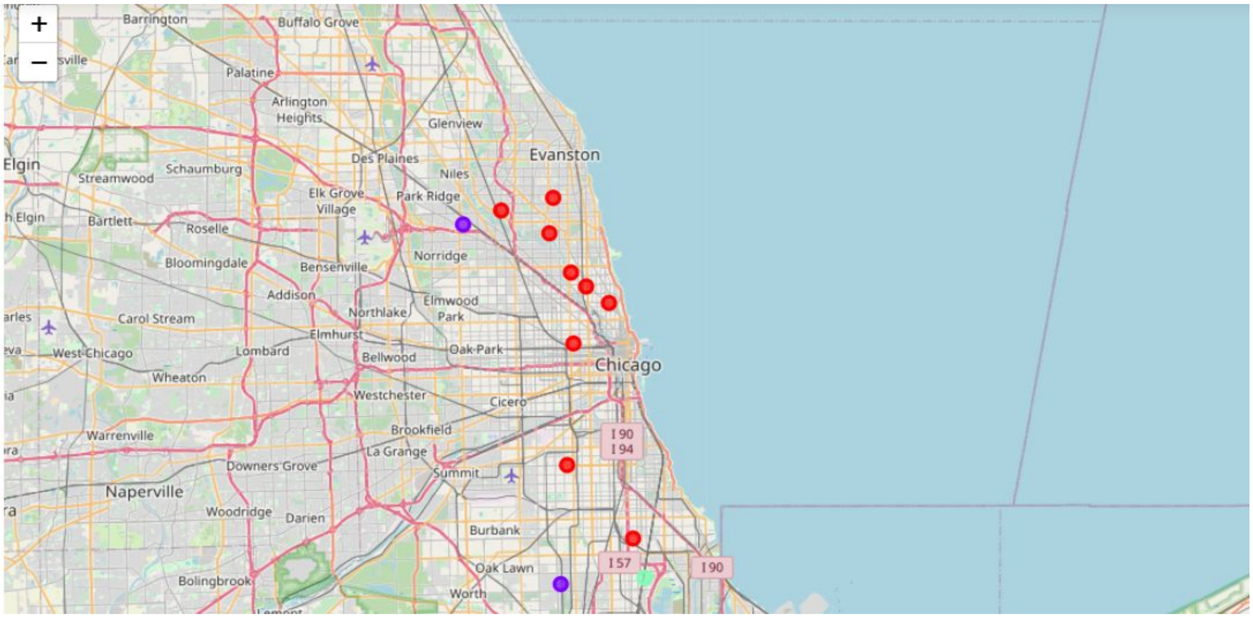
Based on the above plot, the WCSS drastically decreases at 2 and 3 number of Clusters. So i choose 3 as the optimal number of clusters to cluster the neighborhoods using KMeans algorithm.

When applied the KMeans algorithm to the Neighborhoods and their top 10 venues returned by the Foursquare API, it grouped the neighborhoods into 3 clusters based on their similarities. Below is the merged this dataset with cluster labels and top 10 venues for each neighborhood to my original dataset which contains highest Safety Score schools in these neighborhoods with their Latitude and Longitude.



Neighborhood	NAME_OF_SCHOOL	Latitude	Longitude	Cluster Labels	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
LINCOLN PARK	Abraham Lincoln Elementary School	41.924497	-87.644522	0	Bar	Coffee Shop	Sushi Restaurant	Sandwich Place	Italian Restaurant	Thai Restaurant	Pizza Place
NORTH CENTER	Alexander Graham Bell Elementary School	41.949528	-87.686055	0	Pizza Place	Bus Station	Sandwich Place	Pub	Fast Food Restaurant	Mexican Restaurant	Cosmetics Shop
MOUNT GREENWOOD	Annie Keller Elementary Gifted Magnet School	41.697198	-87.697264	1	Park	Intersection	Gourmet Shop	Home Service	Eastern European Restaurant	Falafel Restaurant	Fast Food Restaurant
LAKE VIEW	Augustus H Burley Elementary School	41.937965	-87.669852	0	Gym / Fitness Center	Salon / Barbershop	Pizza Place	Bar	Gym	Furniture / Home Store	Massage Studio
PULLMAN	Edgar Allan Poe Elementary Classical School	41.702620	-87.606456	2	History Museum	Food	Yoga Studio	Gourmet Shop	Donut Shop	Eastern European Restaurant	Falafel Restaurant
FOREST GLEN	Edgebrook Elementary School	41.999460	-87.761821	0	Sandwich Place	American Restaurant	Bus Station	Ice Cream Shop	Grocery Store	Gas Station	Mexican Restaurant
WEST TOWN	Ellen Mitchell Elementary School	41.892055	-87.683179	0	Pub	Art Museum	Pizza Place	Yoga Studio	Flower Shop	Liquor Store	Hot Dog Joint
CHATHAM	James E McDade Elementary Classical School	41.734514	-87.619177	0	Fast Food Restaurant	Discount Store	Intersection	Restaurant	Train Station	Food	Gas Station

Using Folium to view the results of KMeans clustering:

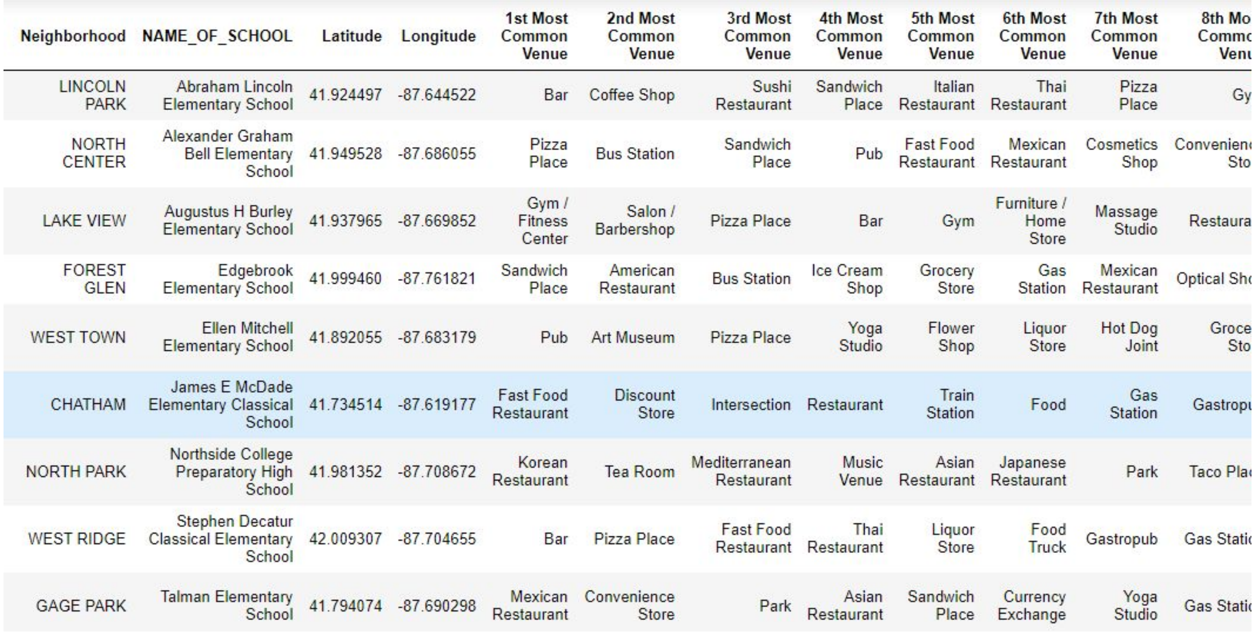


3. Results Session:

Results of KMeans Clustering:

Let's examine the results done by the Kmeans clustering algorithm.

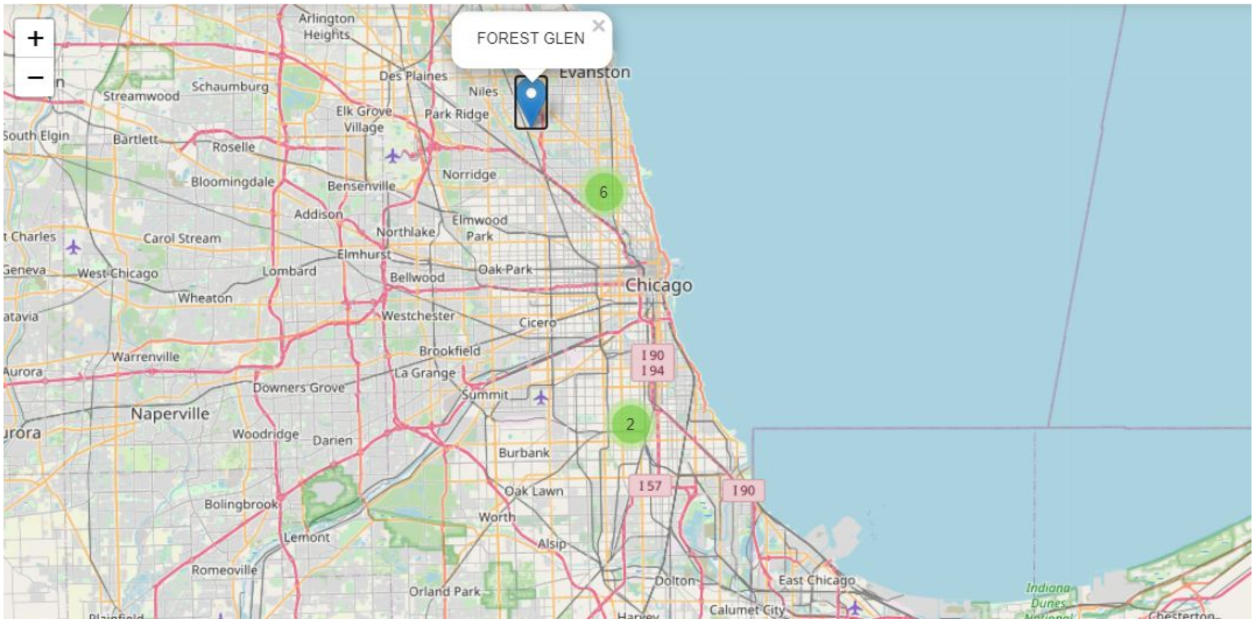
Cluster 1 neighborhoods:



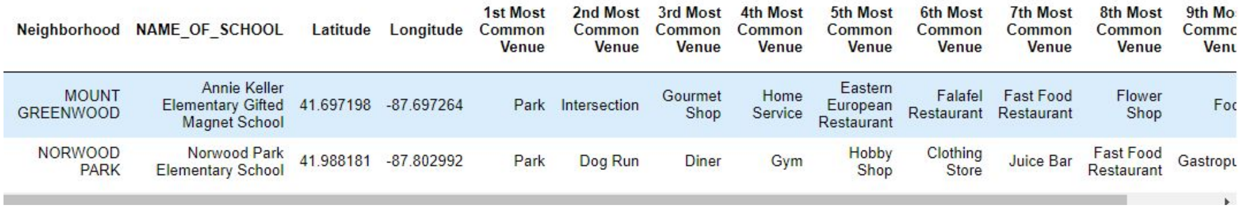
Neighborhood	NAME_OF_SCHOOL	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mo Commo Venu
LINCOLN PARK	Abraham Lincoln Elementary School	41.924497	-87.644522	Bar	Coffee Shop	Sushi Restaurant	Sandwich Place	Italian Restaurant	Thai Restaurant	Pizza Place	Gy
NORTH CENTER	Alexander Graham Bell Elementary School	41.949528	-87.686055	Pizza Place	Bus Station	Sandwich Place	Pub	Fast Food Restaurant	Mexican Restaurant	Cosmetics Shop	Convenien Sto
LAKE VIEW	Augustus H Burley Elementary School	41.937965	-87.669852	Gym / Fitness Center	Salon / Barbershop	Pizza Place	Bar	Gym	Furniture / Home Store	Massage Studio	Restaura
FOREST GLEN	Edgebrook Elementary School	41.999460	-87.761821	Sandwich Place	American Restaurant	Bus Station	Ice Cream Shop	Grocery Store	Gas Station	Mexican Restaurant	Optical Sho
WEST TOWN	Ellen Mitchell Elementary School	41.892055	-87.683179	Pub	Art Museum	Pizza Place	Yoga Studio	Flower Shop	Liquor Store	Hot Dog Joint	Groce Sto
CHATHAM	James E McDade Elementary Classical School	41.734514	-87.619177	Fast Food Restaurant	Discount Store	Intersection	Restaurant	Train Station	Food	Gas Station	Gastropu
NORTH PARK	Northside College Preparatory High School	41.981352	-87.708672	Korean Restaurant	Tea Room	Mediterranean Restaurant	Music Venue	Asian Restaurant	Japanese Restaurant	Park	Taco Plac
WEST RIDGE	Stephen Decatur Classical Elementary School	42.009307	-87.704655	Bar	Pizza Place	Fast Food Restaurant	Thai Restaurant	Liquor Store	Food Truck	Gastropub	Gas Static
GAGE PARK	Talman Elementary School	41.794074	-87.690298	Mexican Restaurant	Convenience Store	Park	Asian Restaurant	Sandwich Place	Currency Exchange	Yoga Studio	Gas Static

Cluster 1 consists of 9 neighborhoods (Lincoln park, North Center, Lake View, Forest Glen, West Town, Chatham, North park, West Ridge, Gage Park) along with their highest safety score schools, packed with a variety of venues like Restaurants, Pizza places, Bar, Gym, etc..

Using Folium to visualize the Cluster 1 neighborhoods:



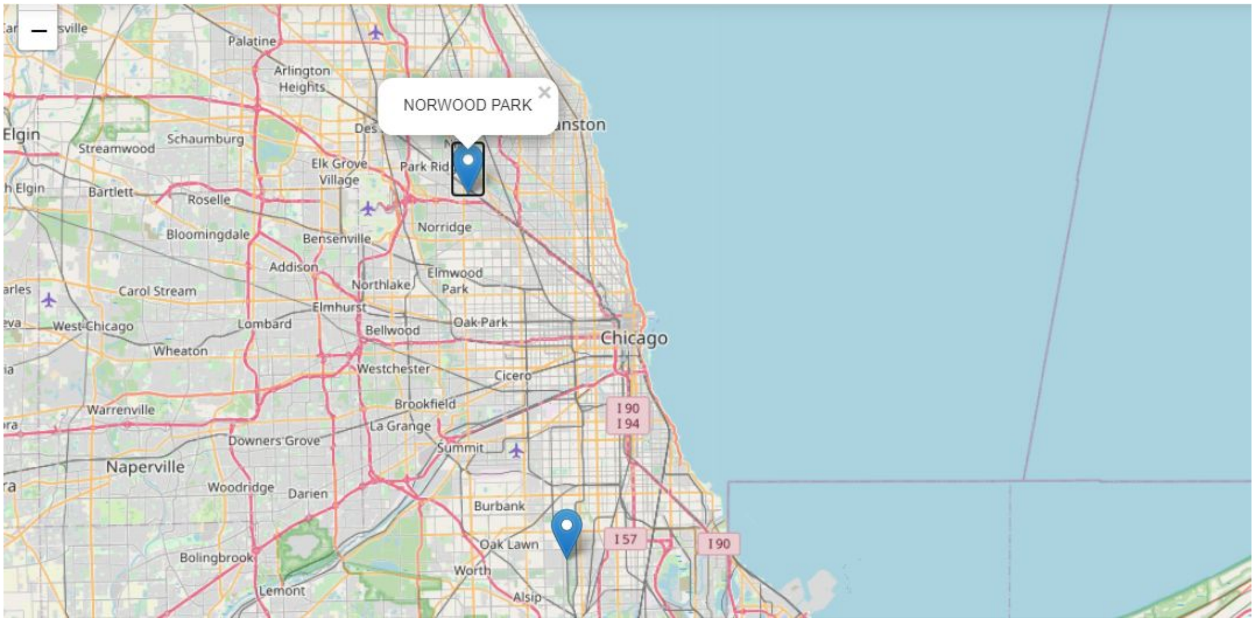
Cluster 2 Neighborhoods:



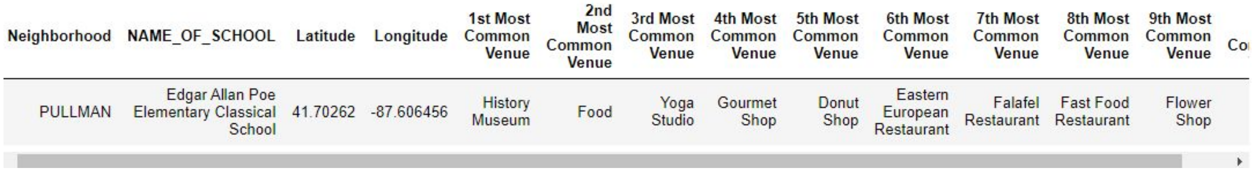
Neighborhood	NAME_OF_SCHOOL	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Mo Commc Venu
MOUNT GREENWOOD	Annie Keller Elementary Gifted Magnet School	41.697198	-87.697264	Park	Intersection	Gourmet Shop	Home Service	Eastern European Restaurant	Falafel Restaurant	Fast Food Restaurant	Flower Shop	Foc
NORWOOD PARK	Norwood Park Elementary School	41.988181	-87.802992	Park	Dog Run	Diner	Gym	Hobby Shop	Clothing Store	Juice Bar	Fast Food Restaurant	Gastropu

Cluster 2 consists of 2 neighborhoods, Mount Greenwood and Northwood Park along with their schools and top 10 venues information. Parks comes as 1st most common venues is a welcoming sight.

Using Folium to visualize the Cluster 2 neighborhoods:



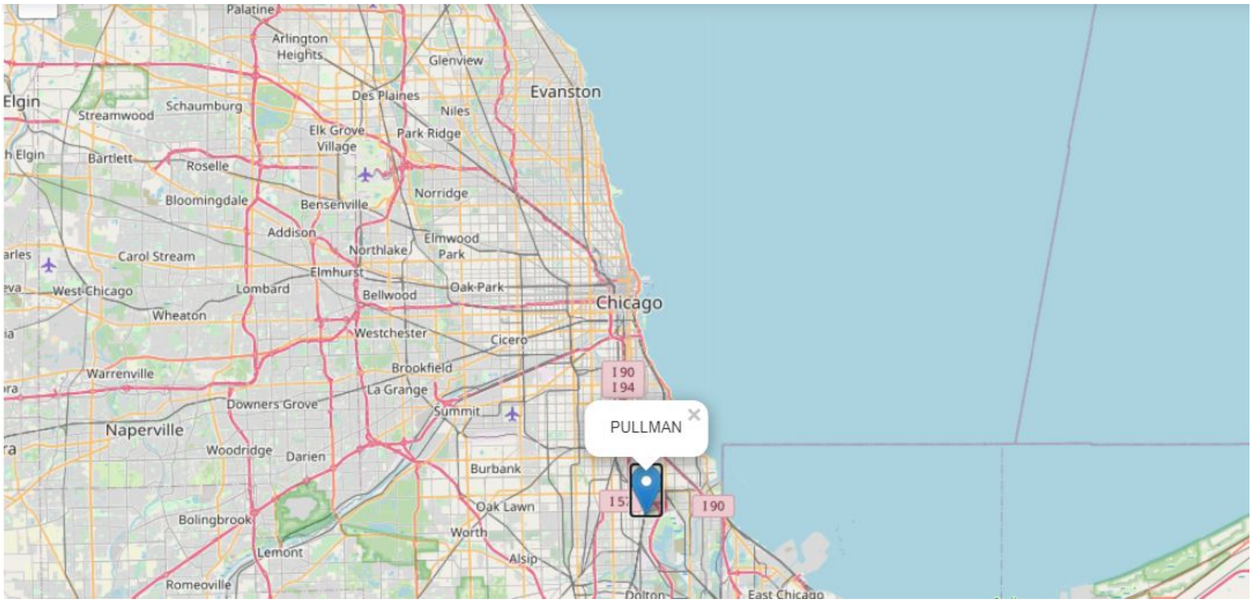
Cluster 3 Neighborhoods:



Neighborhood	NAME_OF_SCHOOL	Latitude	Longitude	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	Co
PULLMAN	Edgar Allan Poe Elementary Classical School	41.70262	-87.606456	History Museum	Food	Yoga Studio	Gourmet Shop	Donut Shop	Eastern European Restaurant	Falafel Restaurant	Fast Food Restaurant	Flower Shop	

Cluster 3 consists of just 1 neighborhood, Pullman. It also has variety of venues like Museum, Restaurants, etc..

Using Folium to visualize the Cluster 3 neighborhood:



4. Discussion Section:

I chose **KMeans clustering algorithm** for this project because KMeans is an unsupervised algorithm, which can group the neighborhoods based on their similarities.

Upon close look at the clusters that the Kmeans produced, Cluster 1 contains 9 neighborhoods with a wide variety of venues like Restaurants, Bar, Gym, Pizza Place,etc., but Parks and Museums rarely appear. Cluster 2 contains 2 neighborhoods, but Parks appear as the 1st most common venue in those neighborhoods. The neighborhood ‘Norwood Park’ even contains a Dog Run as the 2nd most common venue for the Family’s dog to enjoy. Cluster 3 contains 1 neighborhood but much of venues look similar to Cluster 2 except the Museum.

Based on my study, I see the Cluster 2 i.e, the Neighborhoods ‘Mount Greenwood’ and ‘Norwood Park’ will be a good fit for my Friend and his family to move in as it contains Schools as per their request but also comes with venues like Parks (as The most common), variety of Restaurants even a Dog Run for the children and their pet could Enjoy!.

However, in the future, as the family’s needs and wants change, the data regarding Neighborhoods, Schools and the Venues can be expanded and various data analysis data visualization skills can be applied to get more insight of the data, and Machine Learning techniques can be altered to get more accurate results.

5. Conclusion:

Finding a safe place to live and raise a family is a daunting task for parents. In this project I have to managed to address the three important concerns that the family has in their process of moving to Chicago,IL. A list of Neighborhoods which are not only contains highest safety score schools but also comes with venues like Parks, Restaurant, Dog run and a variety of shops like the Family desired.