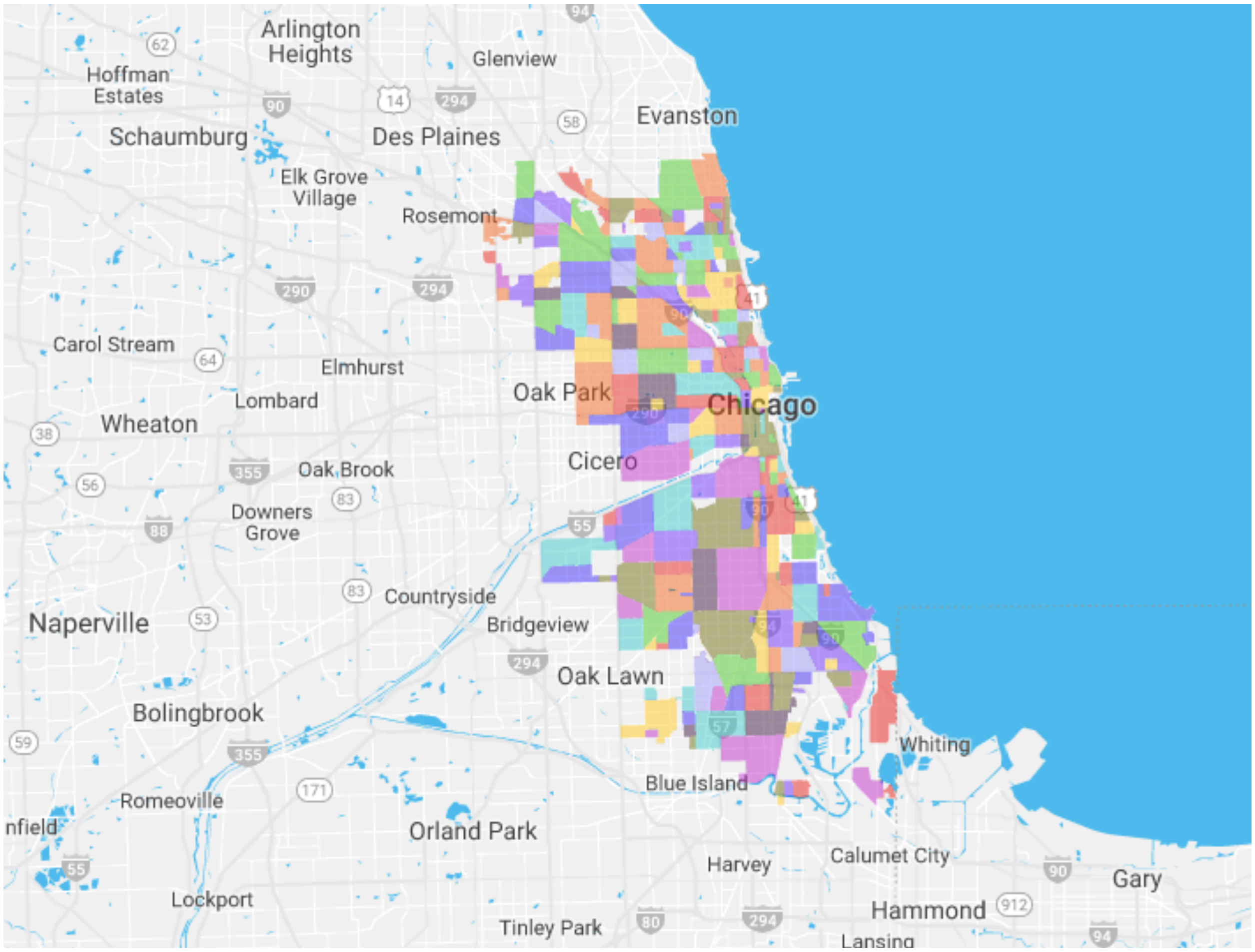
Capstone Project Final Assignment

Submitted by Rajaram Challa

Analysis of Chicago Neighborhoods

Chicago, Illinois Skyline



Chicago Neighborhoods

1. Introduction

So far, in my lifetime, I have moved over 20 times for education and for jobs. I now live in the USA after having lived in India and in the Middle East. Within the USA, I have lived in 10 different cities - not just for a casual visit, but actually established residence and lived - for a minimum of 6 months each.

When I was young and single the selection of the neighborhood to live in was straightforward. Now, with a family and school going kids the process is much more elaborate. While proximity to work / school is an important factor, we also relied on inputs from coworkers. In addition, the decision process relied on our family members' subjective opinions based on observations during a Saturday morning apartment hunting trip:

- My daughter saying "dad, that school building looks new"
- My spouse saying "this strip mall has decent shops and appears to be safe"
- Me thinking "the rent in this area is within our budget"
- etc.

So far, many times, our gut feeling turned out to be OK, but not always. In the past we did not have as much quantitative data available as we have now-a-days. These days we have access to publicly available statistics on socio economic factors, demographic data, crime statistics from municipal, state and research / university sources. We can now augment this public data with commercial data such as Foursquare.

In this assignment I would like use Data Science techniques to make the selection process – of a neighborhood to live in less subjective and more quantitative, by comparing different neighborhoods based on attributes from socio-economic factors, crime data and Foursquare venues data.

For this assignment, I will choose one city (Chicago). In the future we should be able to expand the solution to cover any city of the user's choice.

I am sure many other people move frequently and would benefit from such an analysis.

2. Data

I realize that the in-depth analysis, identifying correlations and clusters, drawing conclusions are heavily dependent on the available data. The availability of ***relevant*** and ***reliable*** data is of utmost importance to analyzing and solving any problem.

In a real-life business scenario, the organization will have inhouse historical data. To augment inhouse data, businesses routinely purchase data from external (syndication / industry) sources.

For my Capstone project, I will rely on data available from

- Foursquare venue data (as required by the assignment)
- Publicly made available demographic / crime / education data from municipal / state agencies; Examples such as
    - Chicago Crime Data (from a previous Coursera Course)
    - Chicago Public School Data (from a previous Coursera Course)
    - Data from websites such as apartments.com

In the Venues data from Foursquare I will look for apartments and train stations.

Additional information on the location of the data sets and the attributes in available in the **Reference Section** at the end of this document.

Since these data sources are likely to have disparate data, I will apply data preparation, data cleansing and data filtering techniques to make the data useful for my analysis.

3. Methodology

In this assignment we will focus our efforts on identifying 'good schools' in higher per-capita neighborhoods with low incidents of crimes. We will limit this analysis to the City of Chicago.

Here are the steps I have followed in this assignment:

- Locate and map all the schools in the Chicago area.
- Identify schools that meet certain requirements.
    - Adequate Yearly Progress Made?
    - High SAFETY_SCORE
    - High AVERAGE_STUDENT_ATTENDANCE
- Analyze socio economic data and identify neighborhoods with a high per-capita income
    - Choose a high quantile(e.g. cse_sorted_df ['per_capita_income_'].quantile(0.75)
- Use geocoders to obtain longitude and latitude for the selected neighborhoods
- Identify and map the locations of crimes (for entire Chicago)
    - Then focus on neighborhoods of interest
- Use Foursquare to locate Apartments and Train Stations
- Review the data produced and make it more useful by eliminating irrelevant rows and by correcting minor issues with the data
- Map Chicago "L" train stations in the selected neighborhoods
- Identify a couple of neighborhoods for further *physical* exploration

4. Results

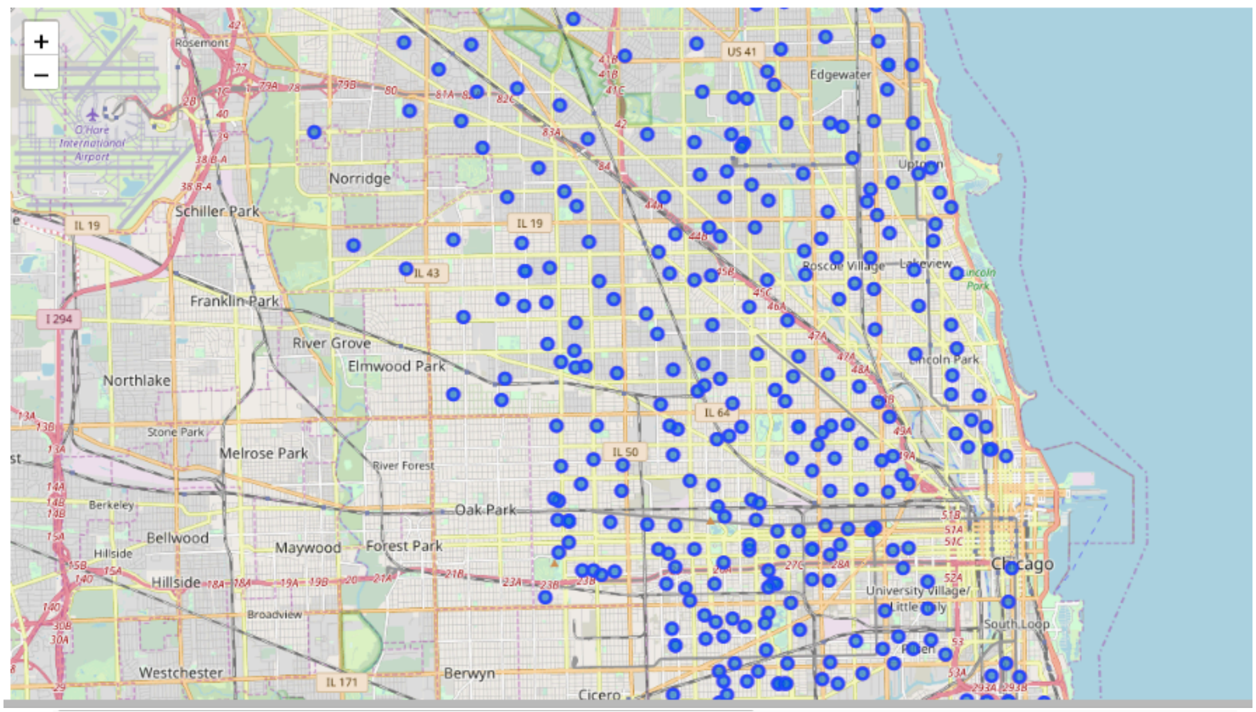
My analysis shows that although there are a great number of schools in Chicago many were in areas of higher crime rates and lower per-capita income neighborhoods. By selecting “good” schools and higher per capita income areas we narrowed down the neighborhoods. Finally, I identified a couple of neighborhoods for further *physical* investigation.

- I started with 566 records for Chicago Public Schools

```
print ('There are {:3d} records for Chicago Public Schools' .format(cps_df.shape[0]))
```

```
There are 566 records for Chicago Public Schools
```

Chicago - All Schools




```
[9]: ith a filter on Adequate Yearly Progress Made? attribute' .format(cps_df[(cps_df['Adequate Yearly Progress Made? '] == 'Yes')].shape[0]))
```

```
There are  72 relevant rows with a filter on Adequate Yearly Progress Made? attribute
```

-
- There are 72 relevant rows with a filter on Adequate Yearly Progress Made? attribute
  - cps_df['Adequate Yearly Progress Made? '] == 'Yes'

```
# let us look for schools where SAFETY_SCORE is in the top 25%
cps_df[cps_df['SAFETY_SCORE'] > cps_df['SAFETY_SCORE'].quantile(0.75)].describe()
```

-

When I applied the filter on Safety Score attribute there are 128 relevant rows
  - I used .quantile against the SAFETY_SCORE attribute to choose schools with a high value
  - cps_df[cps_df['SAFETY_SCORE'] > cps_df['SAFETY_SCORE'].quantile(0.75)].

```
cps_df[cps_df['AVERAGE_STUDENT_ATTENDANCE']>'95.0']['AVERAGE_STUDENT_ATTENDANCE'].describe()
```

```
count        238
unique        29
top       95.50%
freq          25
Name: AVERAGE_STUDENT_ATTENDANCE, dtype: object
```

```
ter on Student Attendance attribute' .format(cps_df[cps_df['AVERAGE_STUDENT_ATTENDANCE']>'95.0']['AVERAGE_STUDENT_ATTENDANCE'].shape[0]))
```

```
There are 238 relevant rows with a filter on Student Attendance attribute
```

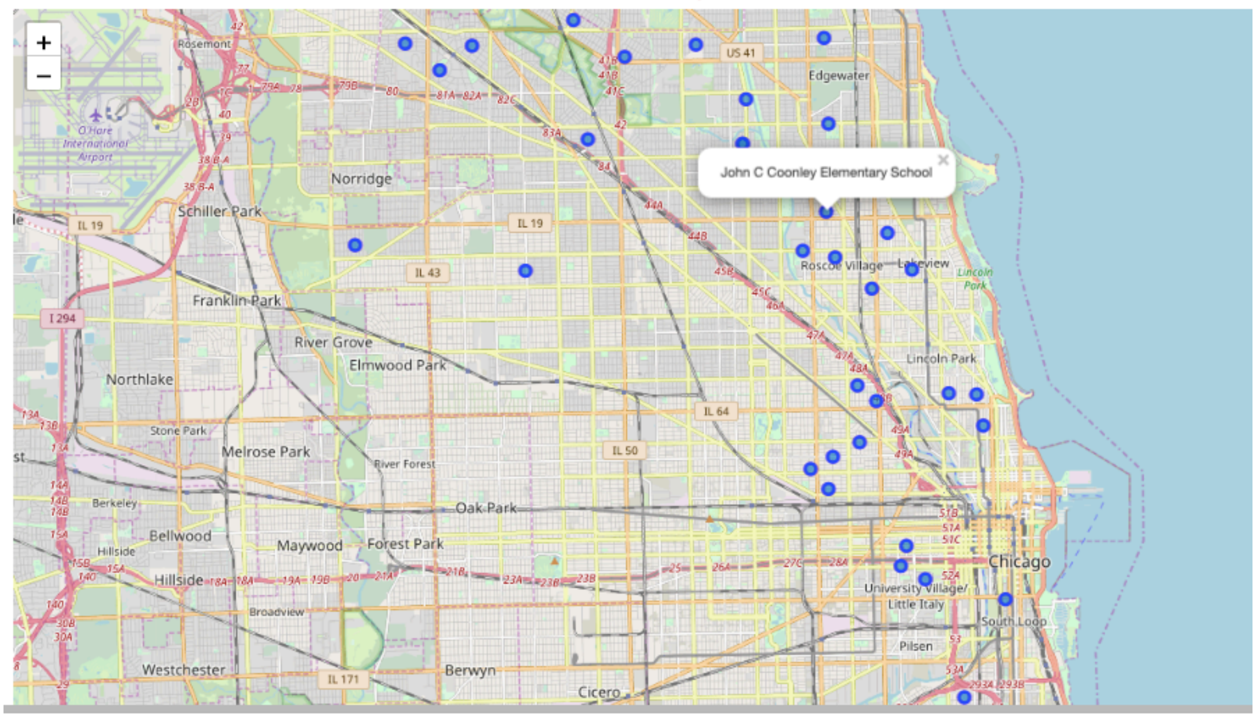
- When I applied the filter on Student Attendance attribute there are 238 relevant rows
    - cps_df['AVERAGE_STUDENT_ATTENDANCE']>'95.0'
- When all applicable filters applied, there are 49 rows to work with. So, in effect, we reduced the raw schools count from 566 to 49 thus far.

```
cps_selected_schools_df = cps_df[
    (cps_df['Adequate Yearly Progress Made? ']          == 'Yes')                          &
    (cps_df['SAFETY_SCORE']                             > cps_df['SAFETY_SCORE'].quantile(0.75))  &
    (cps_df['AVERAGE_STUDENT_ATTENDANCE']               >   '95.0')
]
```

```
print ('There are {:3d} rows with all applicable filters applied' .format(cps_selected_schools_df.shape[0]))
```

```
There are  49 rows with all applicable filters applied
```

Chicago - Selected Schools Only



- I started with 78 rows in the Socio-Economic dataset

```
print ('There are a total of {:3d} rows in the Chicago Socio Economic Dataset' .format(cse_df.shape[0]))
```

```
There are a total of  78 rows in the Chicago Socio Economic Dataset
```

```
# choose neighborhoods that are at the top of the list
cse_sorted_df = cse_sorted_df [cse_sorted_df['per_capita_income_']> cse_sorted_df['per_capita_income_'].quantile(0.60)]
```

```
print ('We will use {:3d} neighborhoods that are in the top quartile of the socio economic data for our analysis' .format(cse_sorted_df.shape[0]))
print ('Notice that the per capita income range in our selected neighborhoods is between ${:6,.0f} and ${:6,.0f} \nwhereas in the original data set the range was betwen ${:6,.0f} and
```

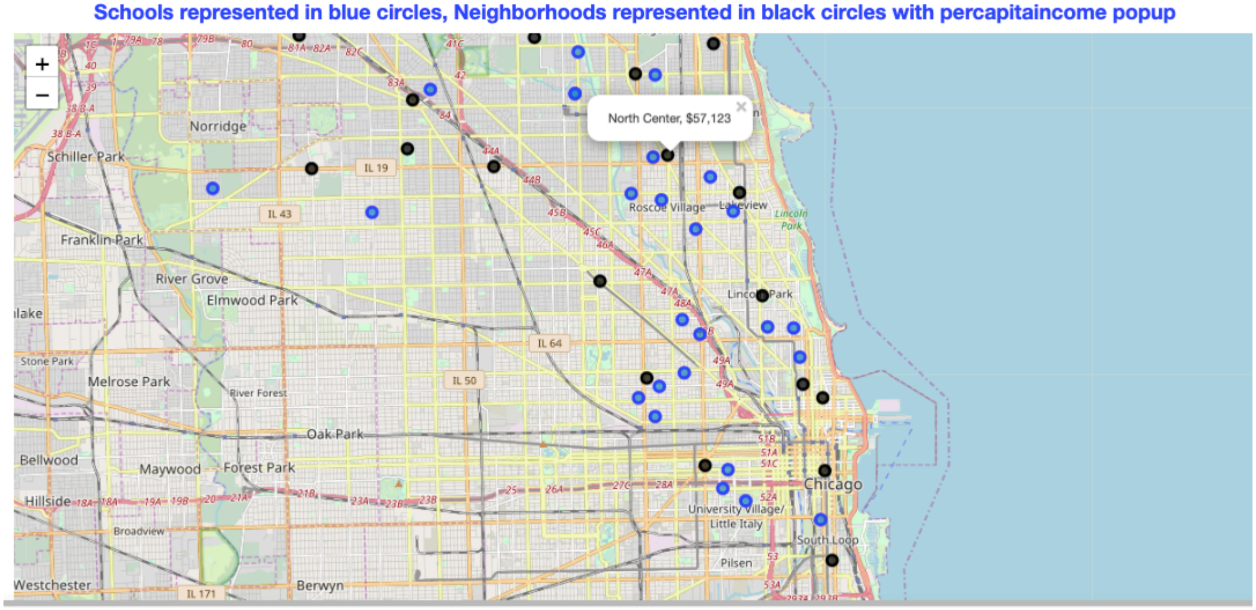
```
We will use  31 neighborhoods that are in the top quartile of the socio economic data for our analysis
Notice that the per capita income range in our selected neighborhoods is between $24,336 and $88,669
whereas in the original data set the range was betwen $ 8,201 and $88,669
```

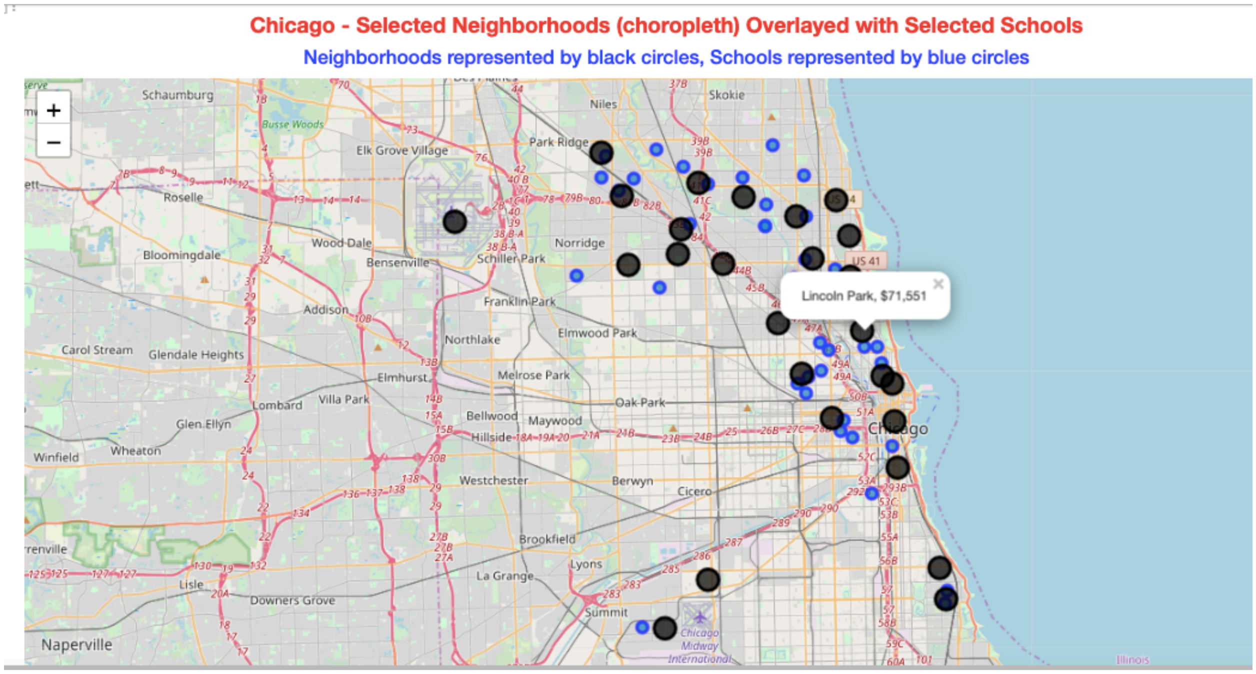
- I will use 31 neighborhoods that are in the top quartile of the socio-economic data for our analysis
- Notice that the per capita income range in
  - the selected neighborhoods is between $24,336 and $88,669 whereas
  - in the original data set the range was between $ 8,201 and $88,669
- Here are the top 20 neighborhoods:

	ca	community_area_name	per_capita_income_
0	8.0	Near North Side	88669
1	7.0	Lincoln Park	71551
2	32.0	Loop	65526
3	6.0	Lake View	60058
4	33.0	Near South Side	59077
5	5.0	North Center	57123
6	28.0	Near West Side	44689
7	12.0	Forest Glen	44164
8	24.0	West Town	43198
9	9.0	Edison Park	40959
10	72.0	Beverly	39523
11	41.0	Hyde Park	39056
12	4.0	Lincoln Square	37524
13	39.0	Kenwood	35911
14	3.0	Uptown	35787
15	74.0	Mount Greenwood	34381
16	77.0	Edgewater	33385
17	10.0	Norwood Park	32875
18	22.0	Logan Square	31908
19	48.0	Calumet Heights	28887

Chicago - Selected Schools and Selected Neighborhoods

Schools represented in blue circles, Neighborhoods represented in black circles with percapitaincome popup

- 

- 

- *For some unknown reason my choropleth map stopped working. I am assuming because of a mismatch in the json data and the key_on parameter. I could not resolve the issue prior to the submission of this assignment. The following observations are from a previous iteration when choropleth map was working.*
- Here are few observations from the above choropleth map:
    - Near North Side, Lincoln Park, Loop have the highest per capita incomes.
    - Notice the bright red colors on choropleth map for Lincoln Park and Loop, but not for Near North Side
    - I verified and found that the geospatial data in the ChicagoGeoJSON file does not have data for Near North Side.
    - Lake View, North Center, West Town, Hyde Park, Edison Park, Norwood Park appear to have many Elementary Schools in their neighborhoods.
- Let us use Foursquare to locate venues in the neighborhoods. Of interest are apartments, train stations and Metro stations.
- From Foursquare documentation the category ids for apartments are:

Residence
4e67e38e036454776db1fb3a

- **Assisted Living**
  5032891291d4c4b30a586d68
- **Home (private)**
  4bf58dd8d48988d103941735
- **Housing Development**
  4f2a210c4b9023bd5841ed28
- **Residential Building (Apartment / Condo)**
  4d954b06a243a5684965b473
- **Trailer Park**
  52f2ab2ebcbc57f1066b8b55

- Similarly, I looked up category ids for Train stations and Metro stations. The category id values I will search for are:
    - Apartments = '4d954b06a243a5684965b473'
    - TrainStation = '4bf58dd8d48988d129951735'
    - Metro = '4bf58dd8d48988d1fd931735'

```
print ('There are {:3d} records in the venues (apartments, train stations) data set for our neighborhoods' .format(venues_df.shape[0]))
```

```
There are 424 records in the venues (apartments, train stations) data set for our neighborhoods
```

- There are 424 records in the venues (apartments, train stations) data set for our neighborhoods

```
venues_df.head()
```

	Neighborhood	Neigh_Lat	Neigh_Long	Venue_Cat	Venue_Name	Venue_Lat	Venue_Long
0	Near North Side	41.900033	-87.634497	Residential Building (Apartment / Condo)	Maple Pointe	41.901623	-87.633804
1	Near North Side	41.900033	-87.634497	Residential Building (Apartment / Condo)	Parc Chestnut	41.897744	-87.635741
2	Near North Side	41.900033	-87.634497	Residential Building (Apartment / Condo)	Chestnut Tower	41.897810	-87.632654
3	Near North Side	41.900033	-87.634497	Residential Building (Apartment / Condo)	1120 N LaSalle	41.902491	-87.632933
4	Near North Side	41.900033	-87.634497	Residential Building (Apartment / Condo)	111 West Maple St	41.901861	-87.631605

```
venues_df.groupby(['Neighborhood','Venue_Cat'])['Venue_Cat'].count()
```

```
Neighborhood  Venue_Cat
Avalon Park   Train                                        1
Beverly       Train Station                                3
CHICAGO       Gym / Fitness Center                         2
              Metro Station                                3
              Office                                       2
                                                          ..
Uptown        Residential Building (Apartment / Condo)    16
              Train                                        2
              Train Station                                2
West Town     Church                                       1
              Residential Building (Apartment / Condo)     1
Name: Venue_Cat, Length: 88, dtype: int64
```

```
venues_df['Venue_Cat'].unique()
```

```
array(['Residential Building (Apartment / Condo)', 'Real Estate Office',
       'Train Station', 'Train', 'Metro Station', 'Office',
       'Mobile Phone Shop', 'Miscellaneous Shop', 'Building',
       'Construction & Landscaping', 'Bank', 'Platform', 'Moving Target',
       'Music Venue', 'Church', "Dentist's Office",
       'Gym / Fitness Center'], dtype=object)
```

- Review the above groupby-count numbers and list of unique venue categories. Here are some observations:
    - Some neighborhoods don't have any apartments within our specified range (e.g. Beverly). This is probably a deficiency in Foursquare data.

- We got some extraneous categories such as Mobile Phone Shops, Real Estate Office, Bank. Music Venue even though we did not ask for these. We will delete such extraneous venues
- For our purposes Train, Train Station, Platform, Metro Station all mean the same. We will call these CTA-L. We will adjust the data frame with the updated category name as "CTA-L"

- There are 393 records in the venues (apartments, train stations) data set for our neighborhoods after deleting irrelevant rows

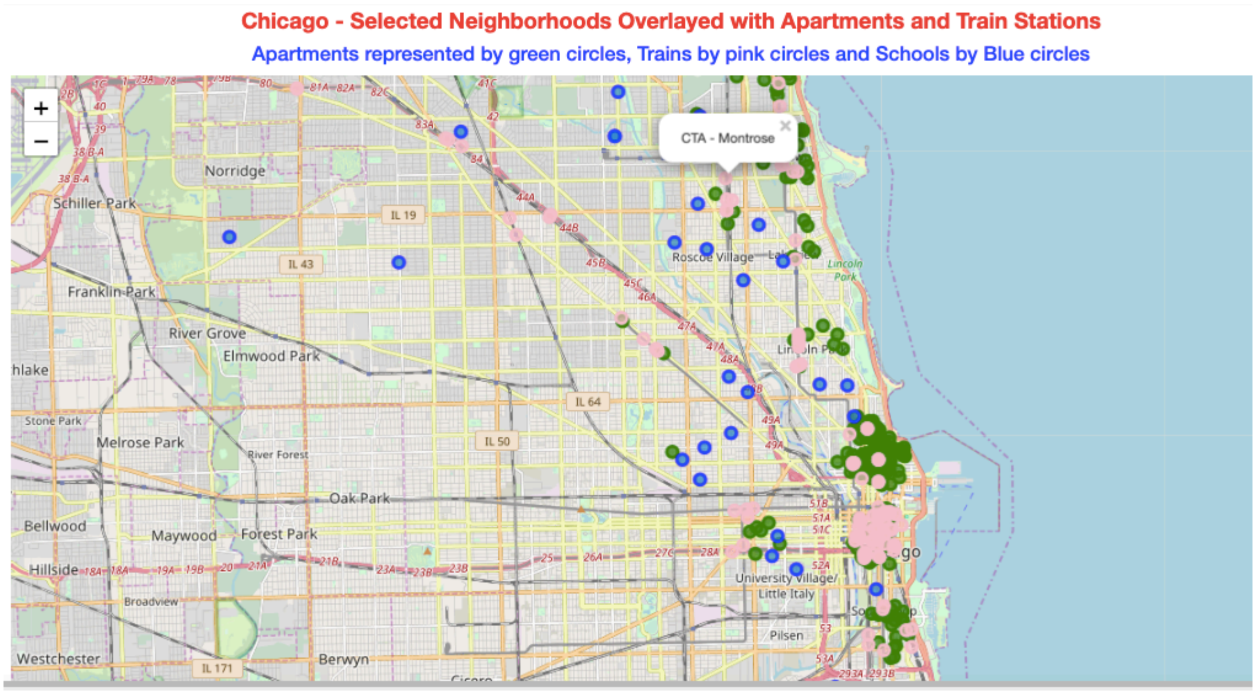
```
print ('There are {:3d} records in the venues (apartments, train stations) data set for our neighborhoods after deleting irrlevant rows' .format(venues_df.s
```

```
There are 393 records in the venues (apartments, train stations) data set for our neighborhoods after deleting irrlevant rows
```

```
venues_df.groupby(['Neighborhood','Venue_Cat'])['Venue_Cat'].count
```

```
Neighborhood       Venue_Cat
Avalon Park        CTA-L                                    1
Beverly            CTA-L                                    3
CHICAGO            CTA-L                                    9
                   Residential Building (Apartment / Condo)    71
Clearing           CTA-L                                    1
Edgewater          CTA-L                                    2
                   Residential Building (Apartment / Condo)     9
Edison Park        CTA-L                                    1
Hyde Park          CTA-L                                    4
                   Residential Building (Apartment / Condo)     6
Irving Park        CTA-L                                    4
Jefferson Park     CTA-L                                    3
Kenwood            CTA-L                                    1
                   Residential Building (Apartment / Condo)     6
Lake View          CTA-L                                    2
                   Residential Building (Apartment / Condo)     5
Lincoln Park       CTA-L                                    9
                   Residential Building (Apartment / Condo)     7
Lincoln Square     Residential Building (Apartment / Condo)     1
Logan Square       CTA-L                                    4
                   Residential Building (Apartment / Condo)     2
Loop               CTA-L                                   66
                   Residential Building (Apartment / Condo)    29
Morgan Park        CTA-L                                    3
Near North Side    CTA-L                                    9
                   Residential Building (Apartment / Condo)    56
Near South Side    CTA-L                                    8
                   Residential Building (Apartment / Condo)    20
```

- As previously observed, not all neighborhoods have apartment venues.

- 

-

```
print ('There are {:3d} records in the Chicago Crimes data set' .format(ccrime_df.shape[0]))
```

```
There are 533 records in the Chicago Crimes data set
```

-

```
ccrime_df.columns
```

```
Index(['ID', 'CASE_NUMBER', 'DATE', 'BLOCK', 'IUCR', 'PRIMARY_TYPE',
       'DESCRIPTION', 'LOCATION_DESCRIPTION', 'ARREST', 'DOMESTIC', 'BEAT',
       'DISTRICT', 'WARD', 'COMMUNITY_AREA_NUMBER', 'FBICODE', 'X_COORDINATE',
       'Y_COORDINATE', 'YEAR', 'UPDATEDON', 'LATITUDE', 'LONGITUDE',
       'LOCATION'],
      dtype='object')
```

```
print ('Notice that we don\'t have community area anme in the above data set')
```

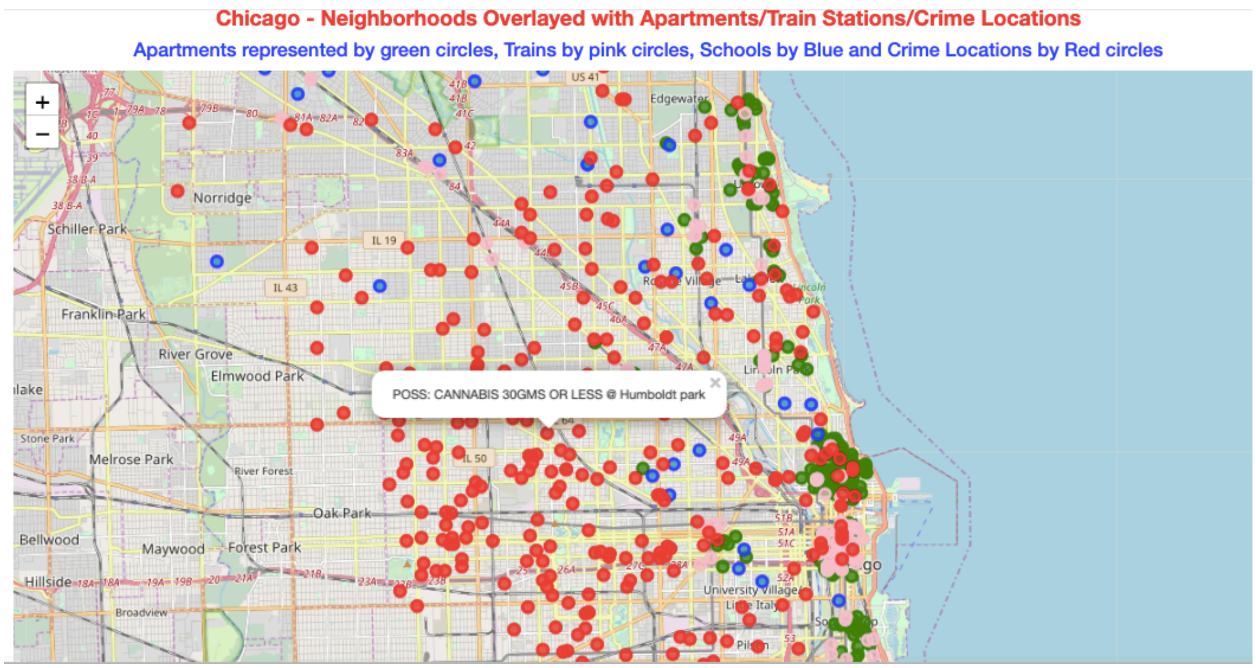
```
Notice that we don't have community area anme in the above data set
```

-
- I will join the Socio-Economic data and the Crime data to match the Community Area Number and get the Community_area_name

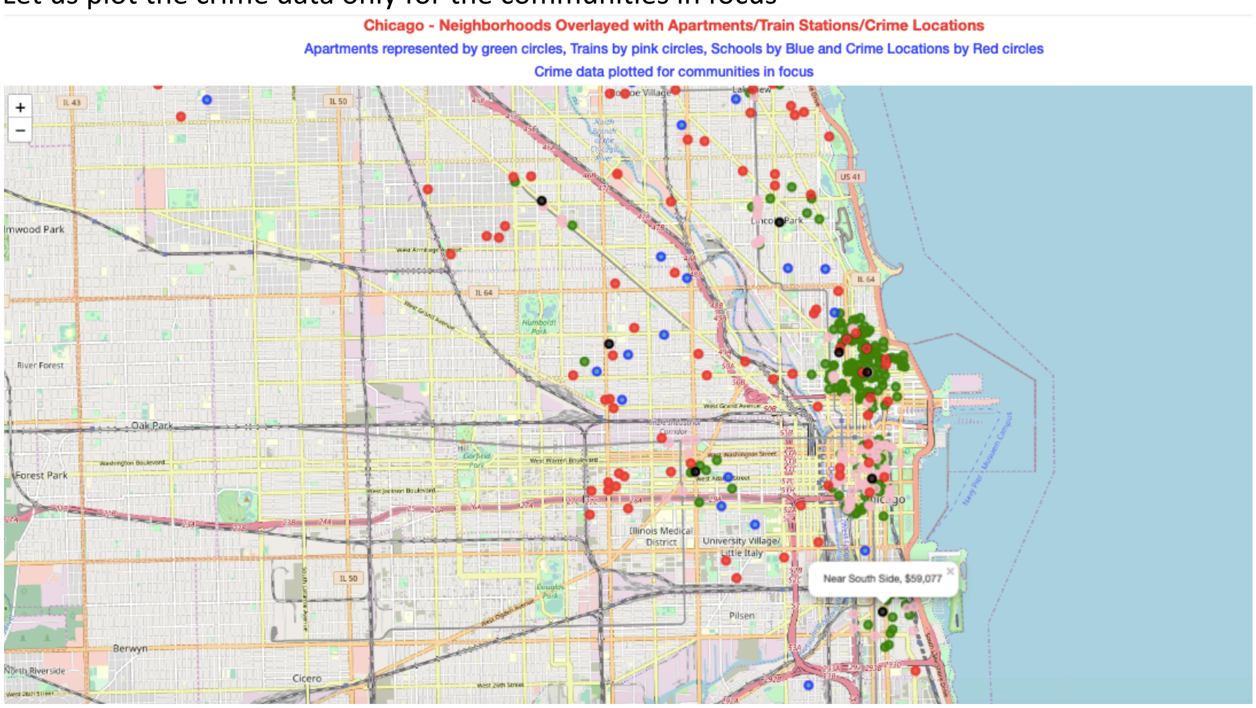
```
print ('There are {:3d} records in the Chicago Crimes data set after dropping rows without Lat/Long' .format(ccrime_relevant_df.shape[0]))
```

```
There are 529 records in the Chicago Crimes data set after dropping rows without Lat/Long
```

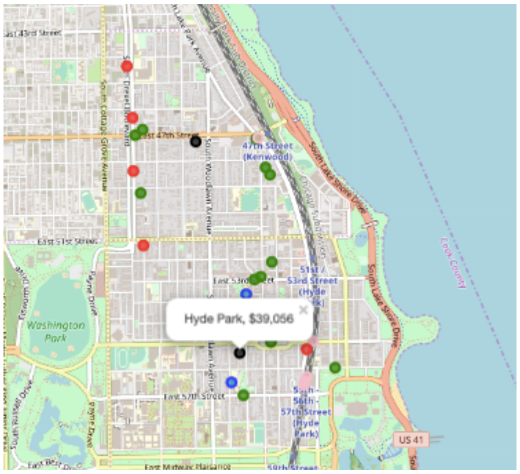
-
- Let us plot the crime locations for a visual check:



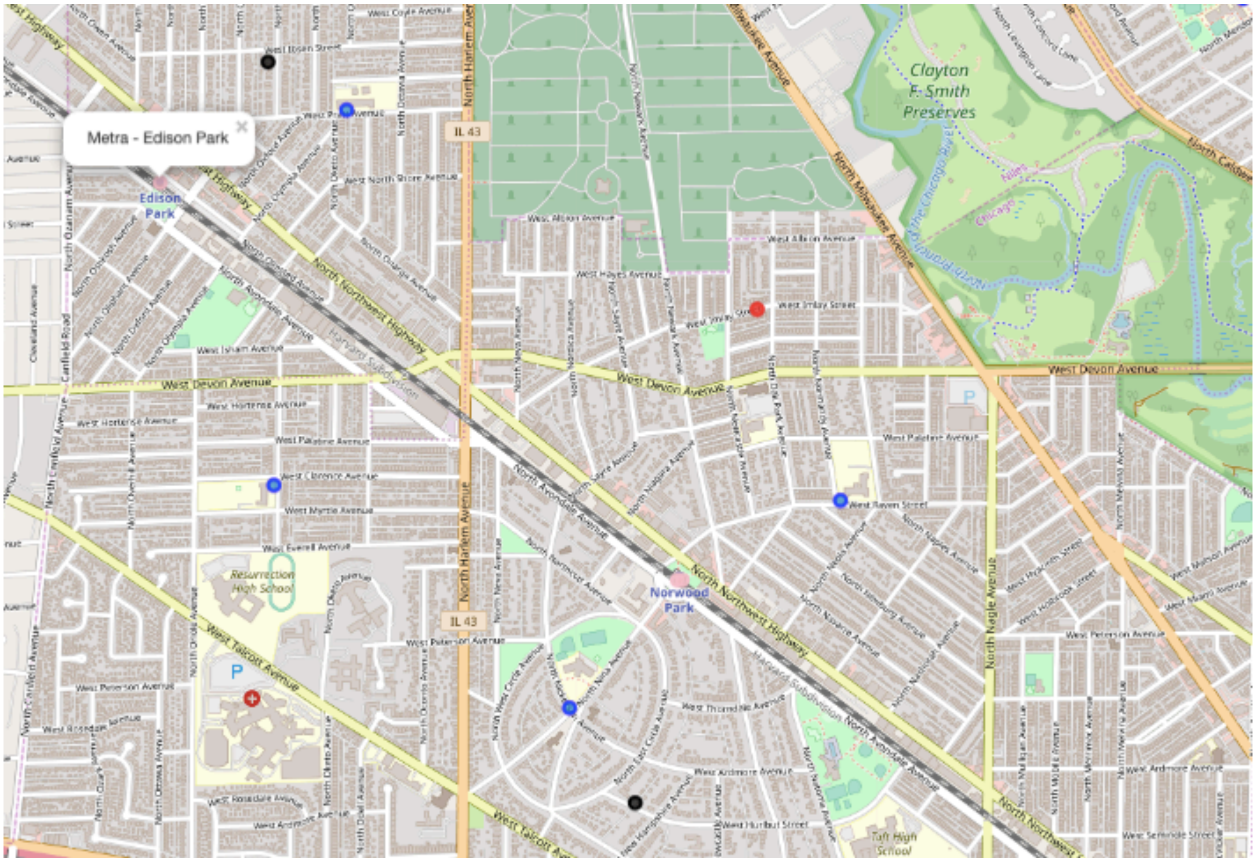
- Observation: There are more crime locations in "other" parts of Chicago than in the parts that we have been focusing on. In the above map the neighborhoods we are interested in have the schools (blue), train stations (pink) and apartments (green) displayed.
- Let us plot the crime data only for the communities in focus



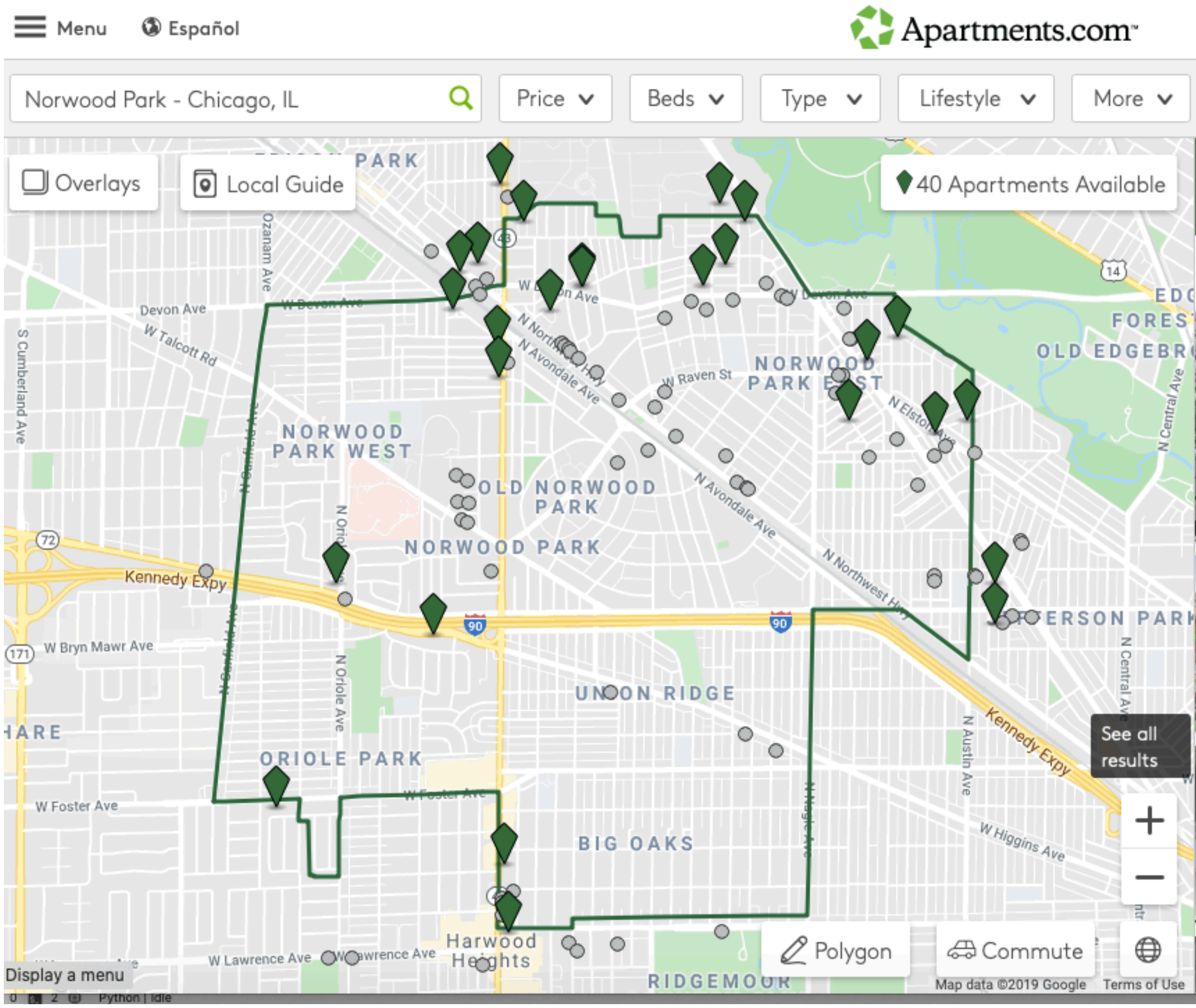
- Observations:

- 

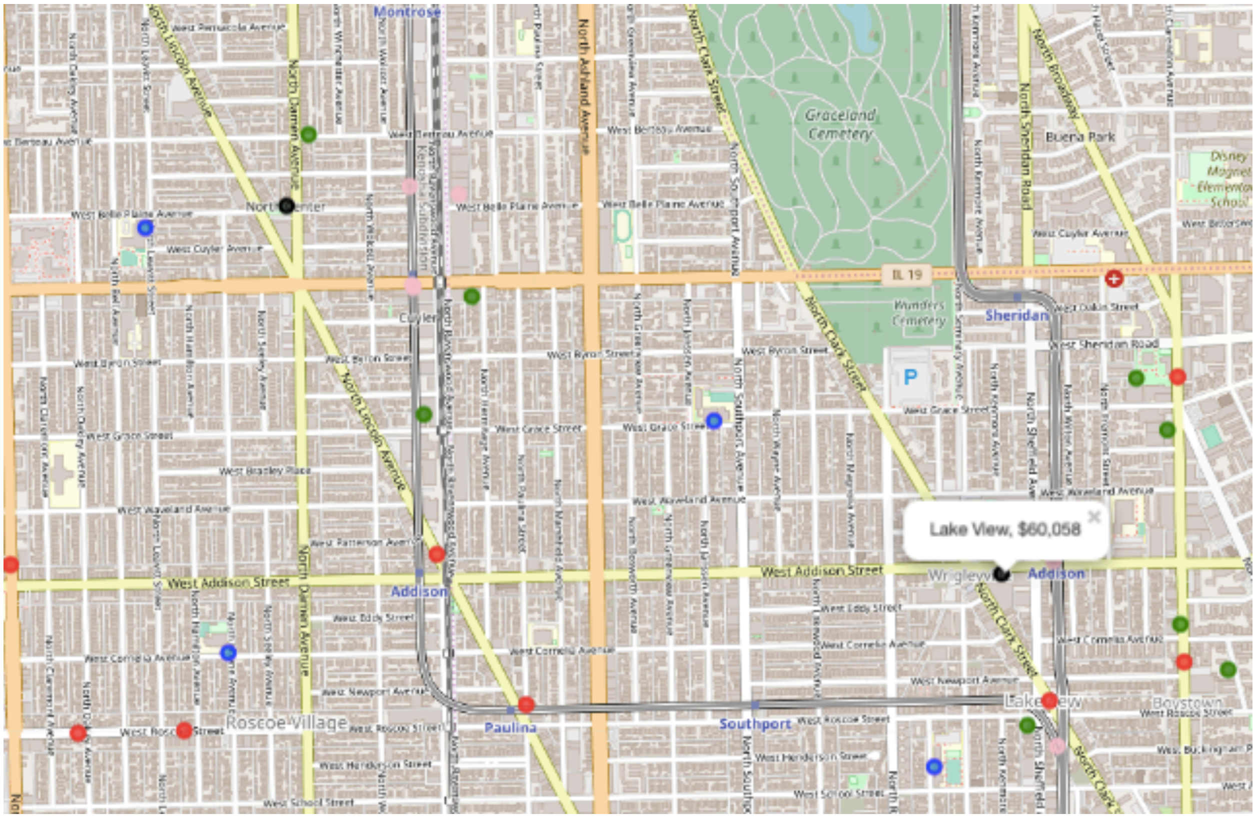
- Hyde Park is a candidate for further investigation
    - Area has low crime
    - Area has a good school: William H Ray Elementary School
    - Easy access to South Shore Commuter Train
    - Solstice on the Park https://www.apartments.com/solstice-on-the-park-chicago-il/v2f4lpe/
    - The BlackWood apartments https://www.apartments.com/the-blackwood-chicago-il/eddsdwm/
    - The Versailles Apts https://www.apartments.com/the-versailles-apartments-chicago-il/6dg2tts/

- 

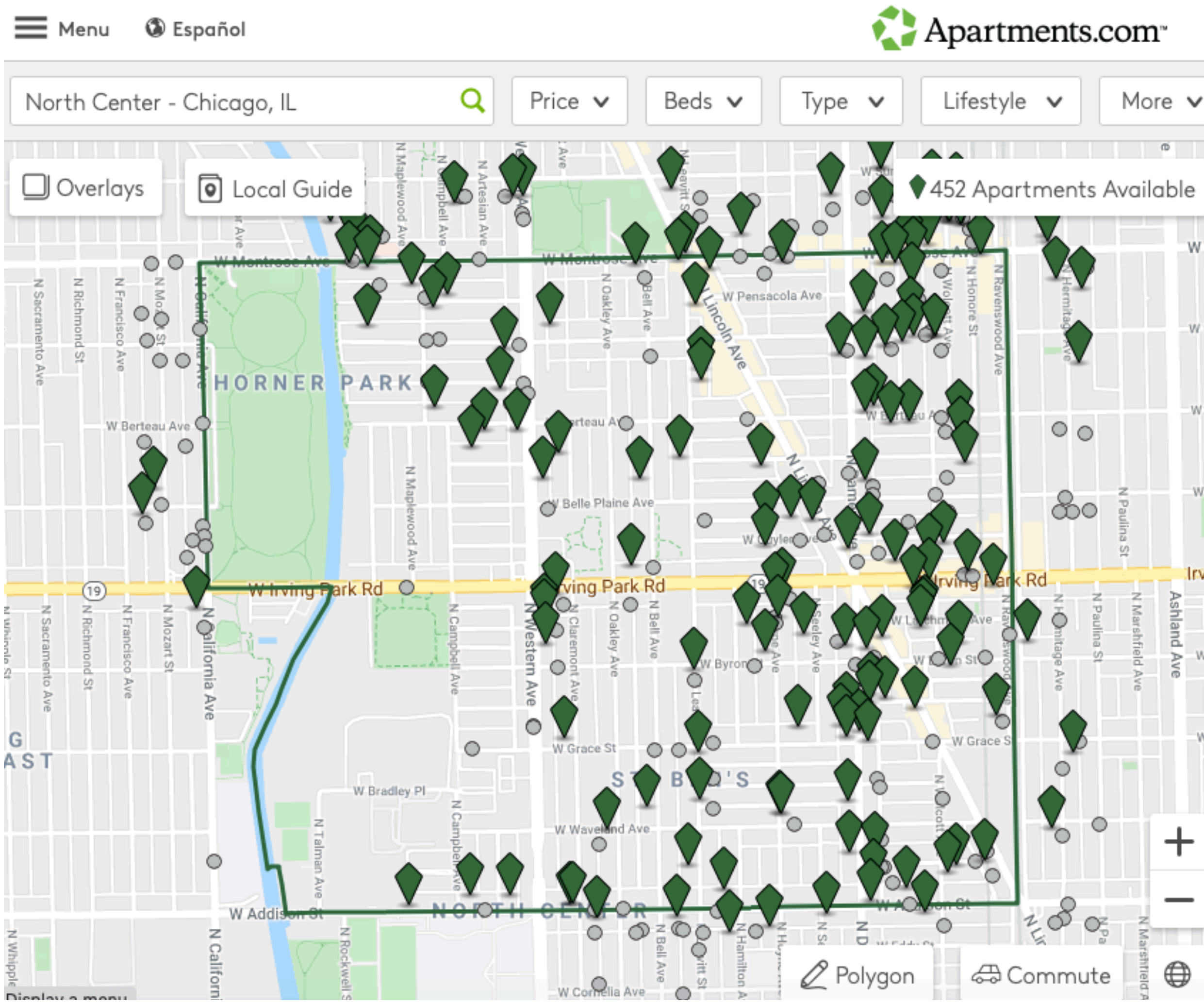
- Edison Park and Norwood Park are also candidates for further investigation:
    - Area has low crime
    - Good schools: Norwood Park Elementary School, Edison Park Elementary School
    - Easy access to Metro: Edison Park and Norwood Park (Metra 624)
    - Foursquare did not show any apartments, but a search on www.apartments.com had a ton of hits



Menu
Español
Apartments.com
Norwood Park - Chicago, IL
Price
Beds
Type
Lifestyle
More
Overlays
Local Guide
40 Apartments Available
NORWOOD PARK WEST
OLD NORWOOD PARK
NORWOOD PARK
NORWOOD PARK EAST
UNION RIDGE
ORIOLE PARK
BIG OAKS
Harwood Heights
RIDGEMOOR
OLD EDGEBR
Devon Ave
W Devon Ave
W Talcott Rd
Kennedy Expy
W Bryn Mawr Ave
W Foster Ave
W Lawrence Ave
W Raven St
N Northwest Hwy
N Avondale Ave
N Elston Ave
N Oriole Ave
N Austin Ave
N Central Ave
W Higgins Ave
S Cumberland Ave
Ozanam Ave
See all results
Polygon
Commute
Display a menu
Map data ©2019 Google
Terms of Use

- 

- North Center and Lake View neighborhoods deserve further investigation:
    - Augustus H Burley Elementary School, Hawthorne Elementary School, James G Blaine Elementary School, John C Coonley Elementary School, John J Audobon Elementary School are all in the vicinity
    - CTA: Addison, CTA Montrose, CTA Irwing Park are all close by
    - Foursquare picked up a few apartments such as: 828 W Grace, Halstead Flats, Elaine Place, 800 West Cornelia


Menu
Español
Apartments.com
North Center - Chicago, IL
Price
Beds
Type
Lifestyle
More
Overlays
Local Guide
452 Apartments Available
W Montrose Ave
HORNER PARK
W Berteau Ave
W Pensacola Ave
N Lincoln Ave
W Belle Plaine Ave
W Irving Park Rd
N California Ave
N Western Ave
W Byron St
W Grace St
W Bradley Pl
W Waveland Ave
W Addison St
NORTH CENTER
W Cornelia Ave
Ashland Ave
Polygon
Commute

5. Observations and Recommendations

- Foursquare did not display many apartments, whereas a search on www.apartments.com produced a huge list of candidate apartments in each of the neighborhoods that I focused on. It is possible that Foursquare is not the most appropriate tool in this case. However, for the purposes of the assignment it did provide us with sufficient number of hits on apartments and train stations.
- For the purposes of this assignment the data sets used were OK. However, the data we used is from public sources and dated. In a real-life-scenario (and for business purposes) we should look for exhaustive and current data.
- Quality of data:
    - Even though we were looking for apartments and train stations Foursquare returned extraneous / irrelevant data. We were able to inspect for such anomalies and take corrective action by deleting (e.g. Mobile Phone Shops) or altering the data (Train same as Train Station). One wonders "what if I am not able to catch bad data?"
- ChicagoGeoJson data was working while I was developing the Notebook, but in the last minute the choropleth maps stopped displaying. Dependence on third party datasets whose structure may change will cause stability issues with solutions. One should pick data sources from reputable/reliable sources.

6. Conclusion

Purpose of this project was to identify neighborhoods in the Chicago area that are candidates to move to.

Certain basic conditions had to be met: good schools, low-crime area, higher per-capita income locality, etc. I used publicly available information to narrow down the neighborhoods. At each stage, I used visual aids – Folium , Choropleth maps – to validate and draw observations.

Once the neighborhoods were selected, I used Foursquare data to identify apartments and train stations. I observed that the Foursquare data was not rich in apartment information in some localities. However, this was not an impediment because apartment information is available through alternative sources.

This exercise is expected to assist stakeholders to narrow the search for neighborhoods to move into. Final decision on choosing an apartment will be made by stakeholders based on visiting apartments in the neighborhoods of interest.

I am sure many other people move frequently and are in search for good localities to move into. I hope such an analysis would benefit others also. In the future, we can harden the solution to other cities also.

7. GitHub Links

Here is the link to my Notebook on my GitHub repository:
https://github.com/RajaramChalla/TorontoClusters/blob/master/Capstone_Final_Assignment%20-%20Chicago%20Neig.ipynb

Please use nbviewer to view the Notebook
https://nbviewer.jupyter.org/ Paste the above Notebook link into the text box on this viewer web page.

Here is the link to the presentation file:
https://github.com/RajaramChalla/TorontoClusters/blob/master/Capstone%20Final%20Assignment%20-%20Presentation.pptx

Here is the link to this report
https://github.com/RajaramChalla/TorontoClusters/blob/master/Capstone_Project_Final_Assignment_Report_by_Rajaram_Challa.docx

Here is the pdf version of this report
https://github.com/RajaramChalla/TorontoClusters/blob/master/Capstone_Project_Final_Assignment_Report_by_Rajaram_Challa.pdf

8. References

8.1 Chicago Public Schools (CPS) - Progress Report Cards

The city of Chicago released a dataset showing all school level performance data used to create School Report Cards. The dataset is available from the Chicago Data Portal: https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t

A static copy of this database (.csv format) can be downloaded from: https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv

This dataset includes a large number of metrics. A pdf document that describes the CPS data is at: https://data.cityofchicago.org/api/assets/AAD41A13-BE8A-4E67-B1F5-86E711E09D5F?download=true

Here is an extract of description of the attributes. Attributes used in my analysis are in red font.

Column Name	Description	Type
School ID		Number
Name of School		Plain Text
Elementary, Middle, or High School	ES indicates elementary school; MS indicates middle school; and HS indicates high school	Plain Text
Street Address		Plain Text
City		Plain Text
State		Plain Text
ZIP Code		Number
Phone Number		Plain Text
Link		Website URL
Network Manager		Plain Text
Collaborative Name		Plain Text
Adequate Yearly Progress Made?		Plain Text
Track Schedule		Plain Text
CPS Performance Policy Status	This reflects whether or not your school is on probation this year. More information on the CPS Performance Policy may be found online at www.cps.edu. N/A: Not applicable NDA: No data available	Plain Text
CPS Performance Policy Level	Level 1 indicates the highest performing schools. Level 2 indicates a middle-performing school that needs improvement. Level 3 indicates the lowest performing schools. NDA indicates no data available.	Plain Text
Healthy Schools Certified?		Plain Text
Safety Icon	Student Perception/Safety category from 5 Essentials survey	Plain Text

Column Name	Description	Type
Safety Score	Student Perception/Safety score from 5 Essentials survey. These scores range from 1 to 99. For more information on these scores, please see the My Voice, My School survey available at www.ccsrsurvey.uchicago.edu/2011.	Number
Parent Environment Score	Parent Perception/Environment score from parent survey. These scores range from 30 to 70. For more information on these scores, please see the parent survey information available at www.cps.edu.	Plain Text
…	…	…
Average Student Attendance	Average daily student attendance %	Number
Rate of Misconducts (per 100 students)	# of misconducts per 100 students	Number
…	…	…
Latitude		Number
Longitude		Number
Community Area Number		Number
Community Area Name		Plain Text
Ward		Number
Police District		Number
Location		Location

8.2 Chicago Socio Economic Indicators

Description and a discussion on the Socio-Economic Indicators is available at: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

The associated data file is located at 'https://data.cityofchicago.org/resource/jcxq-k9xf.csv'

Here is an extract of description of the attributes. Attributes used in my analysis are in red font.

Column Name	Description	Type
Community Area Number		Number
COMMUNITY AREA NAME		Plain Text
PERCENT OF HOUSING CROWDED	Percent occupied housing units with more than one person per room	Number
PERCENT HOUSEHOLDS BELOW POVERTY	Percent of households living below the federal poverty level	Number
PERCENT AGED 16+ UNEMPLOYED	Percent of persons over the age of 16 years that are unemployed	Number
PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	Percent of persons over the age of 25 years without a high school education	Number
PERCENT AGED UNDER 18 OR OVER 64	Percent of the population under 18 or over 64 years of age (i.e., dependency)	Number

Column Name	Description	Type
PER CAPITA INCOME	Community Area Per capita income is estimated as the sum of tract-level aggregate incomes divided by the total population	Number
HARDSHIP INDEX	Score that incorporates each of the six selected socioeconomic indicators (see dataset description)	Number

8.3 Chicago Crime Data

The above referenced Chicago City web site also has data on the crime statistics. The associated data file is located at https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

While this dataset is quite large - over 1.5GB in size with over 6.5 million rows, for the purposes of this assignment a much smaller sample of this dataset was use, available at the above link. A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at:https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Of the attributes compiled by the city: 'ID', 'CASE_NUMBER', 'DATE', 'BLOCK', 'IUCR', 'PRIMARY_TYPE', 'DESCRIPTION', 'LOCATION_DESCRIPTION', 'ARREST', 'DOMESTIC', 'BEAT', 'DISTRICT', 'WARD', 'COMMUNITY_AREA_NUMBER', 'FBICODE', 'X_COORDINATE', 'Y_COORDINATE', 'YEAR', 'UPDATEDON', 'LATITUDE', 'LONGITUDE', 'LOCATION'
For our analysis we will use: 'DESCRIPTION', 'COMMUNITY_AREA_NUMBER', 'LATITUDE', 'LONGITUDE'. The community area name is not available in this data set. We will use 'COMMUNITY_AREA_NUMBER' from the crime data set to join with the Socio-Economic Data set to get the Community_Area_Name.

8.4 Chicago Rapid Transit System The Chicago "L"

[Chicago Transit Authority CTA (CTA-L)]