Capstone Project-The battle of Neighbourhoods

Siri Devarapalli May 2020

1.Introduction

1.1 Background

The average American moves about ten times in their lifetime. This brings us to a question: Do people move until they can find a safe place to stay or do they settle down for other factors such as proximity to office, malls, etc. we mostly try our best to search for a safe neighbourhood and one which fits our criteria. One might want to stay close to their work place or some families prefer to stay near the school of their children so that they are comfortable.

To minimize the chance of feeling any discomfort after shifting we should research the neighbourhood thoroughly. We should consider our priorities and make an informed choice so that we don't regret it after getting there. One common concern is safety. All of us want to live safe communities.

1.2 Problem

The crime statistics dataset of Chicago found on cityofchicago.org has crimes in each district from 2012 to 2017. The year 2017 will be taken as the latest data . the crime rates in each district may have changed over time. This project aims to select the safest district in Chicago based on total crimes, explore the district to find the ten most common venues in each district and finally cluster the neighbourhoods using k-means clustering.

1.3 Interest

People who are considering to relocate to Chicago will find it easy to identify the safest district in Chicago and common venues in each district.

2.Data acquisition and Cleaning

2.1 Data Acquisition

The data acquired for this project is from three sources. The first data source uses a Chicago crime that shows all the crimes from 2001 till 2020 in Chicago. The dataset contains the following columns:

1. ID

2. Case number

3. Date

4. Primary type

5. Beat

6. District

7. Community area

8. FBI code

The second source is scraped from a Wikipedia page that contains the list of community areas. This page also contains additional information about each community area.

1. Serial number

2. Name

3. Area (sq. kms)

4. 2017 population density (per square km)

The third data source is a list of all the neighbourhoods in each community area as found on a Wikipedia page. The dataset is created from scratch using the data from the website. The following are the columns:

1. Neighbourhood

2. Community area

3. Latitude

4. Longitude

2.2 Data Cleaning

The data preparation for each of the three sources was done separately. From the crime in Chicago dataset the crimes of year 2017 are only selected.

	ID	Case Number	Date	Primary Type	Beat	District	Community Area	FBI Code	Year
0	11034701	JA366925	01-01-2001	DECEPTIVE PRACTICE	412	4	45.0	11	2001
1	11227287	JB147188	10-08-2017	CRIM SEXUAL ASSAULT	2222	22	73.0	2	2017
2	11227583	JB147595	03/28/2017 02:00:00 PM	BURGLARY	835	8	70.0	5	2017
3	11227293	JB147230	09-09-2017	THEFT	313	3	42.0	6	2017
4	11227634	JB147599	08/26/2017 10:00:00 AM	CRIM SEXUAL ASSAULT	122	1	32.0	2	2017

Fig1. The first dataset before pre-processing.

The second data is scraped from a Wikipedia page .

	Number[8]	Name[8]	2017 population[9]	Area (sq mi.)[10]	Area (km2)	2017 population	2017 population.1
0	NaN	NaN	NaN	NaN	NaN	density (/sq mi.)	density (/km2)
1	1.0	Rogers Park	55062.0	1.84	4.77	29925	11554.1
2	2.0	West Ridge	76215.0	3.53	9.14	21590.7	8336.2
3	3.0	Uptown	57973.0	2.32	6.01	24988.4	9648.06
4	4.0	Lincoln Square	41715.0	2.56	6.63	16294.9	6291.5

Fig2 The second dataset before pre-processing.

We make sure that the community area numbers are same so that we can merge the data frames using these numbers(1-77). To identify the community area with the least crimes in the year 2017.After visualising the crime in each community area we can find the community with the least crime rate and hence tag the safest community area.

	Community Area	BeatARSON	BeatASSAULT	BeatBATTERY	BeatBURGLARY	BeatCONCEALED CARRY LICENSE VIOLATION	BeatCRIM SEXUAL ASSAULT	BeatCRIMINAL DAMAGE
0	1.0	2423	606326	1549923	419568	0	72774	1280483
1	2.0	2411	535952	1263620	420465	0	37788	1021174
2	3.0	1914	479427	1173212	285036	0	87848	571137
3	4.0	4063	225057	511664	205114	0	38012	364044
4	5.0	1921	109399	207378	224630	0	15350	224707
5	6.0	3858	373822	1363423	550871	1933	92436	689466
6	7.0	1935	236321	614846	348812	3870	37230	656867
7	8.0	10971	774877	2367928	376467	1824	157407	1023396
8	9.0	4834	22560	67680	4833	0	4835	54793
9	10.0	1611	117757	253221	90371	0	16123	187085

Fig3 The merged dataset .

The third source of data is created from the list of neighbourhoods from scratch using the list available on Wikipedia. This data contains all the neighbourhoods in the safest community areas.

	Neighborhood	Community	Latitude	Longitude
0	Burnside	Burnside		
1	Edgebrook	Forest Glen		
2	North Edgebrook	Forest Glen		
3	South Edgebrook	Forest Glen		
4	Forest Glen	Forest Glen		
5	Old Edgebrook	Forest Glen		
6	Wildwood	Forest Glen		
7	Sauganash	Forest Glen		

Fig4 The third dataset which contains neighbourhoods.

The coordinates are generated using Google maps API geo-encoding . The new dataset is used to generate the venues for each neighbourhood using the Foursquare API. These neighbourhoods are grouped using K means clustering.

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Statistical Summary of Crimes

The describe function is used to get the statistical summary of crimes in Chicago. It returns the mean, standard deviation, minimum, maximum, 25% quartile, 50% quartile, 75% quartile for each of the crimes.

	Community Area	BeatARSON	BeatASSAULT	BeatBATTERY	BeatBURGLARY	BeatCONCEALED CARRY LICENSE VIOLATION	BeatCRIM SEXUAL ASSAULT	Beat DAM
count	77.000000	77.000000	7.700000e+01	7.700000e+01	77.000000	77.000000	77.000000	7.70
mean	39.000000	6033.285714	2.478528e+05	6.409228e+05	182623.467532	1087.311688	22733.207792	3.93
std	22.371857	7448.743165	2.776027e+05	7.671015e+05	183967.194760	3534.239385	31020.934660	4.09
min	1.000000	0.000000	1.192600e+04	2.297400e+04	4474.000000	0.000000	0.000000	1.77
25%	20.000000	1383.000000	6.153800e+04	1.360570e+05	42009.000000	0.000000	3295.000000	8.32
50%	39.000000	2736.000000	1.515420e+05	3.945980e+05	130626.000000	0.000000	10118.000000	2.70
75%	58.000000	10092.000000	3.736560e+05	9.424190e+05	247631.000000	612.000000	27748.000000	5.71
max	77.000000	37760.000000	1.848355e+06	5.307395e+06	875934.000000	26436.000000	157407.000000	2.59

Fig5. Descriptive analysis of crimes in Chicago

3.1.2 Community areas with the highest crime rate

Comparing 5 communities with the highest crime rate it is evident that the community areas 25, 8, 32, 29 and 28 have the highest crime rate compared to other communities.

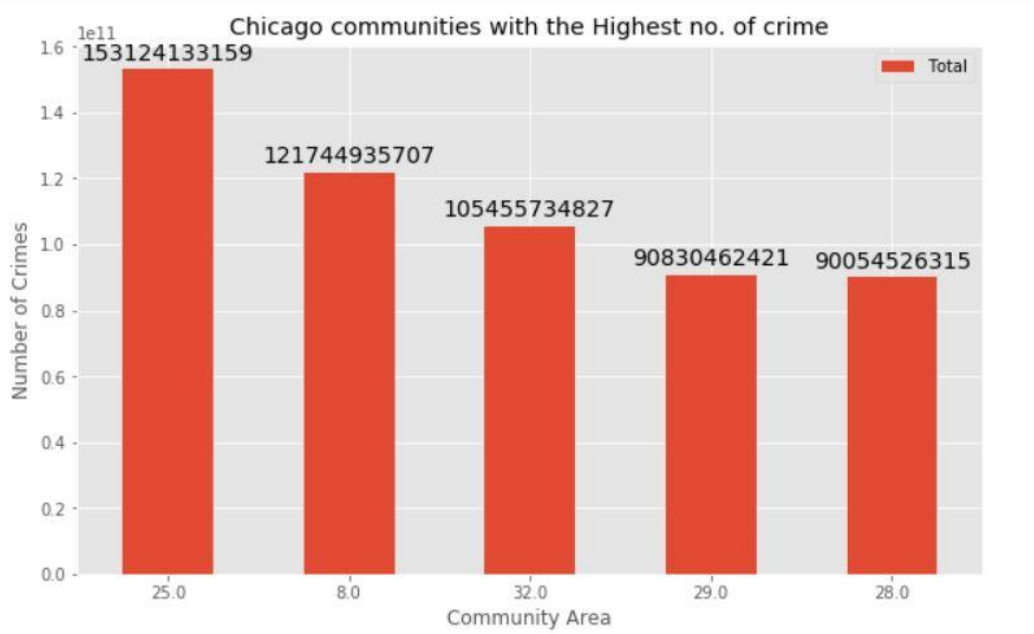


Fig6.Communities with the highest crime rates

3.1.3 Community areas with the lowest crime rate

Comparing the communities with the lowest crime rate in 2017 we find that the community areas 9, 47, 12, 74 and 18 have the lowest crime rates.

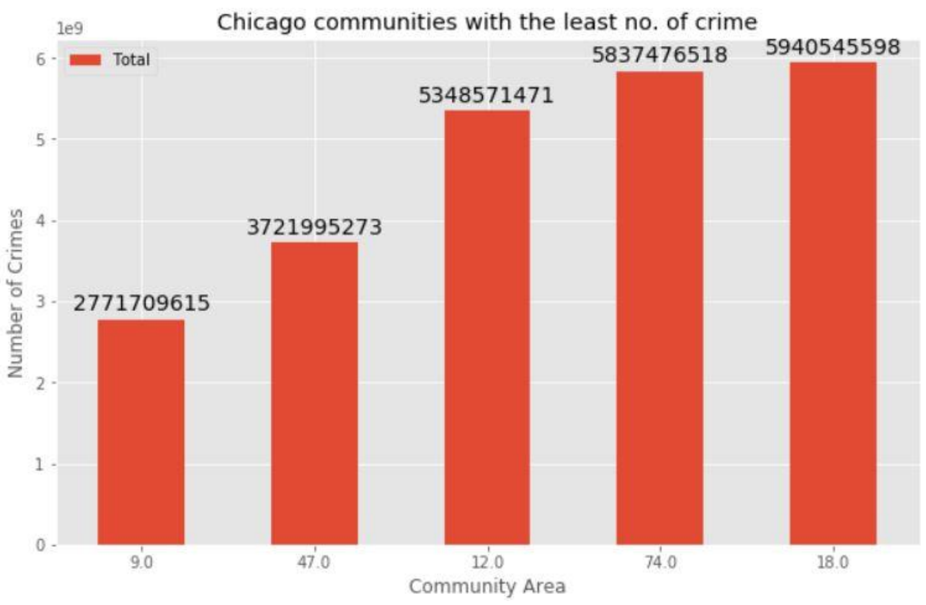


Fig7 Communities with lowest crime rate

Community 9 is Edison park. It has the lowest crime rate but it is very small and is on the outskirts of Chicago. It’s population is also very low.

```
Community Area                              9
name                              Edison Park
population in 2017                      11605
area(sq mi.)                             1.13
Area(km2)                                2.93
2017 population density/sq mi.         4235.4
2017 popualation density/km2           1635.3
Name: 9, dtype: object
```

Fig8. A description of community area 9.

So, we will not consider this community area. We will consider the next community areas which are 47 and 12

```
Community Area                           47
name                               Burnside
population in 2017                     2254
area(sq mi.)                           0.61
Area(km2)                              1.58
2017 population density/sq mi.      3695.08
2017 popualation density/km2        1426.68
Name: 47, dtype: object
Community Area                                12
name                               Forest Glen
population in 2017                       19019
area(sq mi.)                               3.2
Area(km2)                                 8.29
2017 population density/sq mi.         5943.44
2017 popualation density/km2           2294.78
Name: 12, dtype: object
```

Fig 9. A description of community areas 47 and 12.

3.1.4 Neighbourhoods in community areas Burnside and Forest Glen

There are 8 neighbourhoods in the community areas of Burnside and Forest Glen. They are visualized on a map using folium on python.

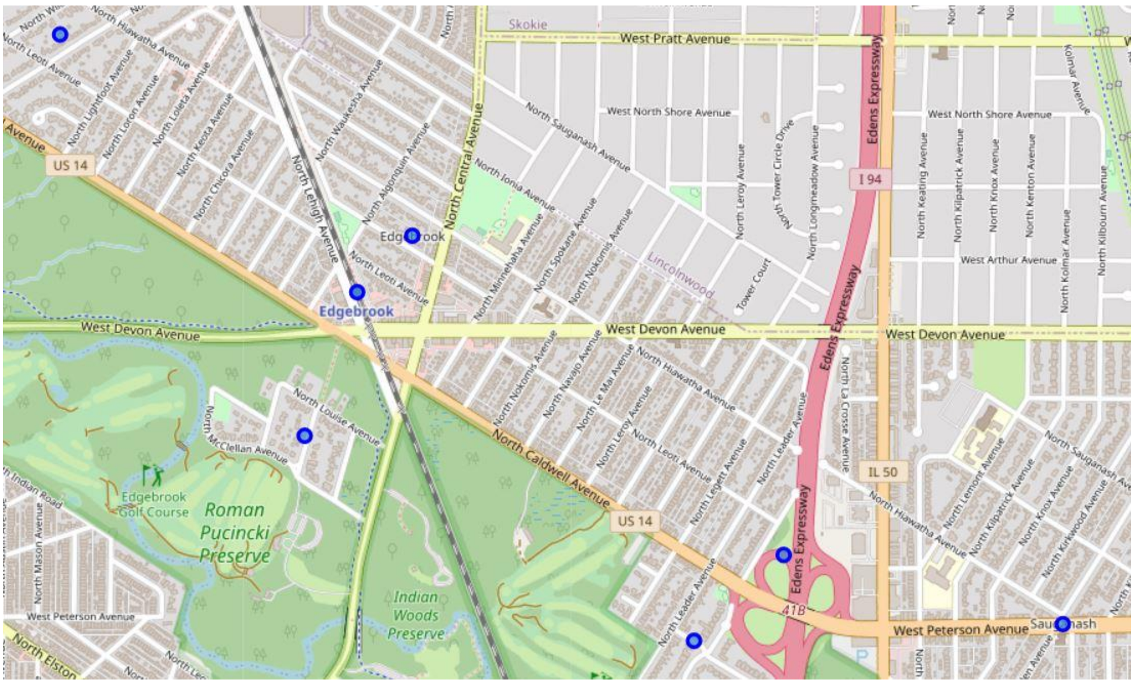


Fig 10. Map showing neighbourhoods in Forest Glen.

3.2 Modelling

Using the final dataset we find venues near each neighbourhood in Forest Glen and Burnside in a radius of 500 meters by connecting to the Foursquare API. This returns a json file containing all the venues in each neighbourhood which is changed into a data frame by pandas. This data frame contains all the venues along with their coordinates and category.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Burnside	41.730035	-87.596714	Captain Clean	41.728278	-87.598975	Home Service
1	Burnside	41.730035	-87.596714	93rd St. & Cottage Grove Ave.	41.728274	-87.600860	Intersection
2	Burnside	41.730035	-87.596714	Cta Training Center	41.733666	-87.595408	Bus Station
3	Burnside	41.730035	-87.596714	Metra - 91st Street (Chesterfield)	41.730079	-87.601962	Train Station
4	Edgebrook	41.999677	-87.764100	Chocolate Shoppe Ice Cream	41.997200	-87.762554	Ice Cream Shop

Fig 11. Venue details of each neighbourhood.

One hot encoding is done on the venues(One hot encoding is a process by which categorical variables are converted into a form that can be provided to ML algorithms to do a better job in prediction). The venues data is then grouped by the neighbourhood and mean of the venues are calculated, finally the top ten venues are selected.

To help people find similar neighbourhoods we use k mean clustering (a form of unsupervised machine learning that clusters data based on predefined cluster size). We will use a cluster size of 3 which will divide 8 neighbourhood into 3 clusters. The reason to cluster the neighbourhoods is to make it easier for people to eliminate irrelevant neighbourhoods based on amenities and venues in each neighbourhood.

4. Results

After running the k means clustering algorithm we can access each cluster to see which neighbourhoods were assigned to each of the three clusters. The first cluster has the following neighbourhoods.

	Neighborhood	Community	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6t C V
1	Edgebrook	Forest Glen	41.999677	-87.764100	0	Spa	Sandwich Place	American Restaurant	Plaza	Vietnamese Restaurant	D
2	North Edgebrook	Forest Glen	41.998269	-87.765976	0	Sandwich Place	Park	American Restaurant	Plaza	Hobby Shop	G S
4	Forest Glen	Forest Glen	41.991752	-87.751674	0	Yoga Studio	Indian Restaurant	Asian Restaurant	Coffee Shop	Fast Food Restaurant	G C
5	Old Edgebrook	Forest Glen	41.994708	-87.767727	0	Sandwich Place	Salon / Barbershop	Diner	Park	Coffee Shop	B S
6	Wildwood	Forest Glen	42.004691	-87.775924	0	American Restaurant	Nature Preserve	Baseball Field	Theater	Park	G S
7	Sauganash	Forest Glen	41.990036	-87.742289	0	Park	Indian Restaurant	Asian Restaurant	Basketball Court	Pharmacy	F R

Fig 12. cluster 1

Cluster one is the biggest cluster. It has 6 neighbourhoods. Upon observation we can find out that it's most common venues are restaurants, parks, shops and fitness studios.

	Neighborhood	Community	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 Commo Venue
3	South Edgebrook	Forest Glen	41.989608	-87.754688	1	Moving Target	Other Great Outdoors	Golf Course	Gas Station	Ice Cream Shop	Home Service

Fig 13. Cluster 2

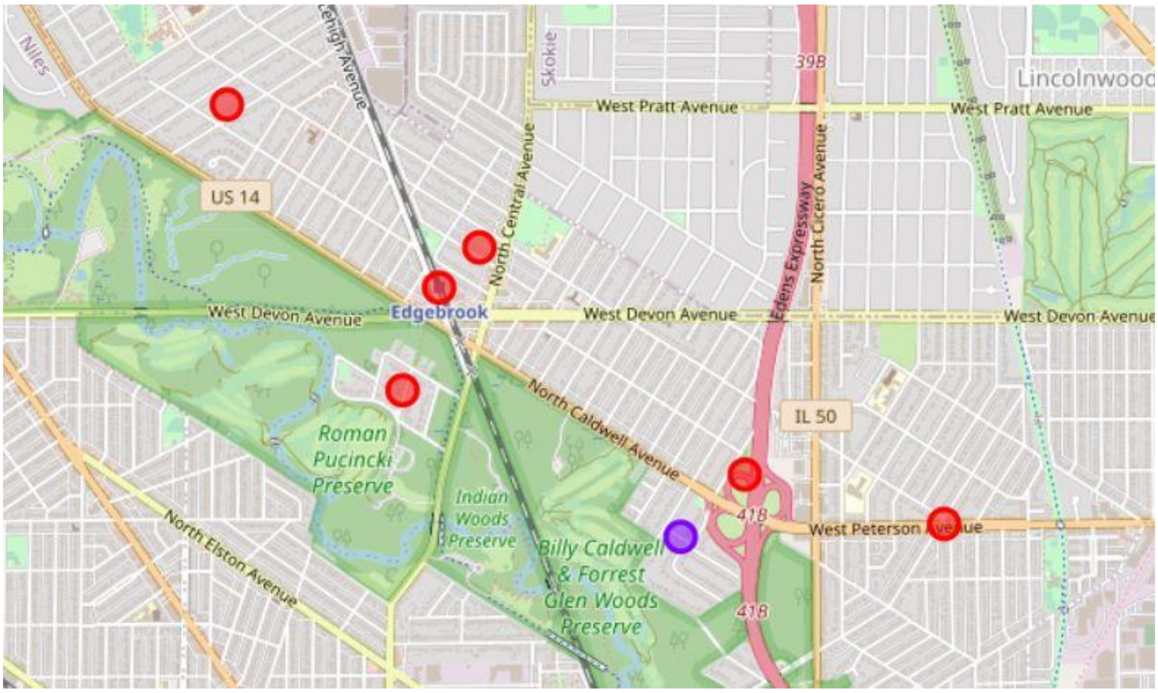
The second cluster consists of one neighbourhood. The venues are a target store, the great outdoors, a gold course, gas station and a home service.

	Neighborhood	Community	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 Mo Comm Venue
0	Burnside	Burnside	41.730035	-87.596714	2	Train Station	Home Service	Bus Station	Intersection	Yoga Studio	Gas Station

Fig 14 Cluster 3

The third cluster has one neighbourhood. The venues are a train station, home service, bus station, Intersection, fitness centre and a gas station.

Visualizing the clustered neighbourhoods on a folium map



Each cluster is colour coded for readability. Red represents the first cluster. Blue represents the second cluster. Green represents the third cluster.

5. Discussion

The aim of this project is to help people who want to relocate to the safest community in London, expats can chose the neighbourhoods to which they want to relocate based on the most common venues in it. For example if a person is looking for a neighbourhood with good connectivity and public transportation we can see that Clusters 3 has Train stations and Bus stops as the most common venues. If a person is looking for a neighbourhood with stores and restaurants in a close proximity then the neighbourhoods in the first cluster is suitable. For a family I feel that the neighbourhoods in Cluster 2 is more suitable dues to the common venues in that cluster, these neighbourhoods have common venues such as golf course, Gym/Fitness centre, Restaurants, Home service and great outdoors which is ideal for a family.

6.Conclusion

This project helps a person get a better understanding of the neighbourhoods with respect to the most common venues in that neighbourhood. It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a neighbourhood. We have just taken safety as a primary concern to shortlist the community area in Chicago. The future of this project includes taking other factors such as cost of living in the areas into consideration to shortlist the borough based on safety and a predefined budget.