
CAPSTONE PROJECT- THE BATTLE OF NEIGHBOURHOODS- PRESENTATION

SIRI DEVARAPALLI



INTRODUCTION

- Background: 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.
- Problem: This project aims to select the safest community in Chicago based on total crimes, explore the community to find the ten most common venues in each neighbourhood and finally cluster the neighbourhoods using k-means clustering.
- Interest: People who are considering to relocate to Chicago will find it easy to identify the safest community in Chicago and common venues in each neighbourhood.

DATA ACQUISITION AND CLEANING

- 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 second source is scraped from a Wikipedia page that contains the list of community areas.
- The third data source is a list of all the neighbourhoods in each community area as found on a Wikipedia page.

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.
- 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.
- 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.
- 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.

METHODOLOGY

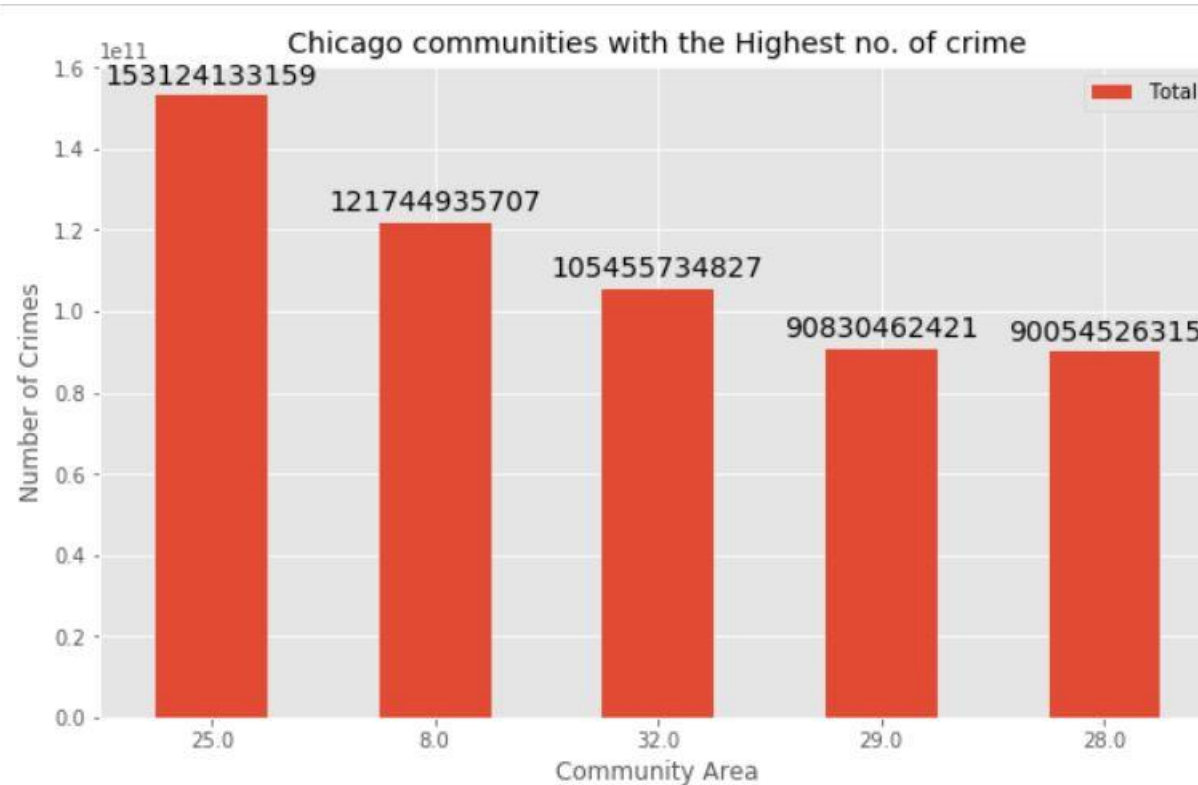
Exploratory data analysis

Statistical summary of crimes

	Community Area	BeatARSON	BeatASSAULT	BeatBATTERY	BeatBURGLARY	BeatCONCEALED CARRY LICENSE VIOLATION	BeatCRIM SEXUAL ASSAULT	BeatDAM
count	77.000000	77.000000	7.700000e+01	7.700000e+01	77.000000	77.000000	77.000000	7.700000
mean	39.000000	6033.285714	2.478528e+05	6.409228e+05	182623.467532	1087.311688	22733.207792	3.930000
std	22.371857	7448.743165	2.776027e+05	7.671015e+05	183967.194760	3534.239385	31020.934660	4.090000
min	1.000000	0.000000	1.192600e+04	2.297400e+04	4474.000000	0.000000	0.000000	1.770000
25%	20.000000	1383.000000	6.153800e+04	1.360570e+05	42009.000000	0.000000	3295.000000	8.320000
50%	39.000000	2736.000000	1.515420e+05	3.945980e+05	130626.000000	0.000000	10118.000000	2.700000
75%	58.000000	10092.000000	3.736560e+05	9.424190e+05	247631.000000	612.000000	27748.000000	5.710000
max	77.000000	37760.000000	1.848355e+06	5.307395e+06	875934.000000	26436.000000	157407.000000	2.590000

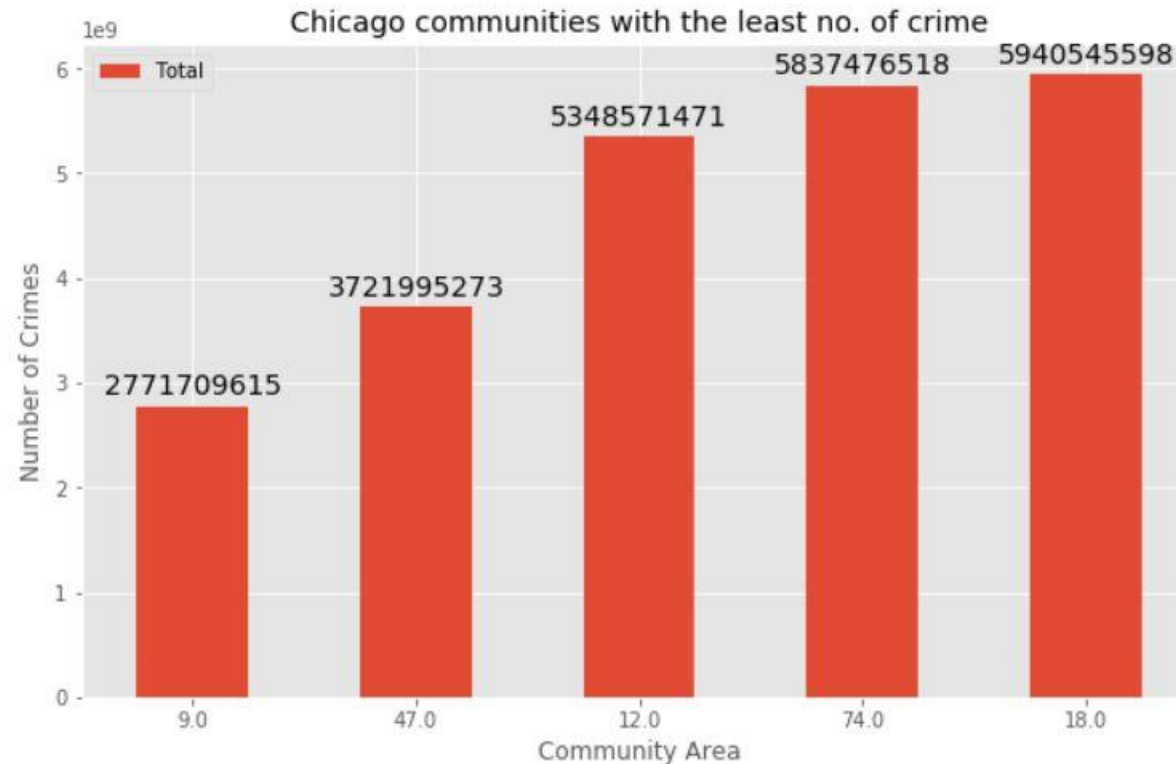
There are a total of 137 crimes of which we will select the major crimes. Theft is the most common crime recorded in 2017. theft is followed by other offense, weapons violation, stalking and sex offense.

COMMUNITY AREAS WITH THE HIGHEST CRIME RATE



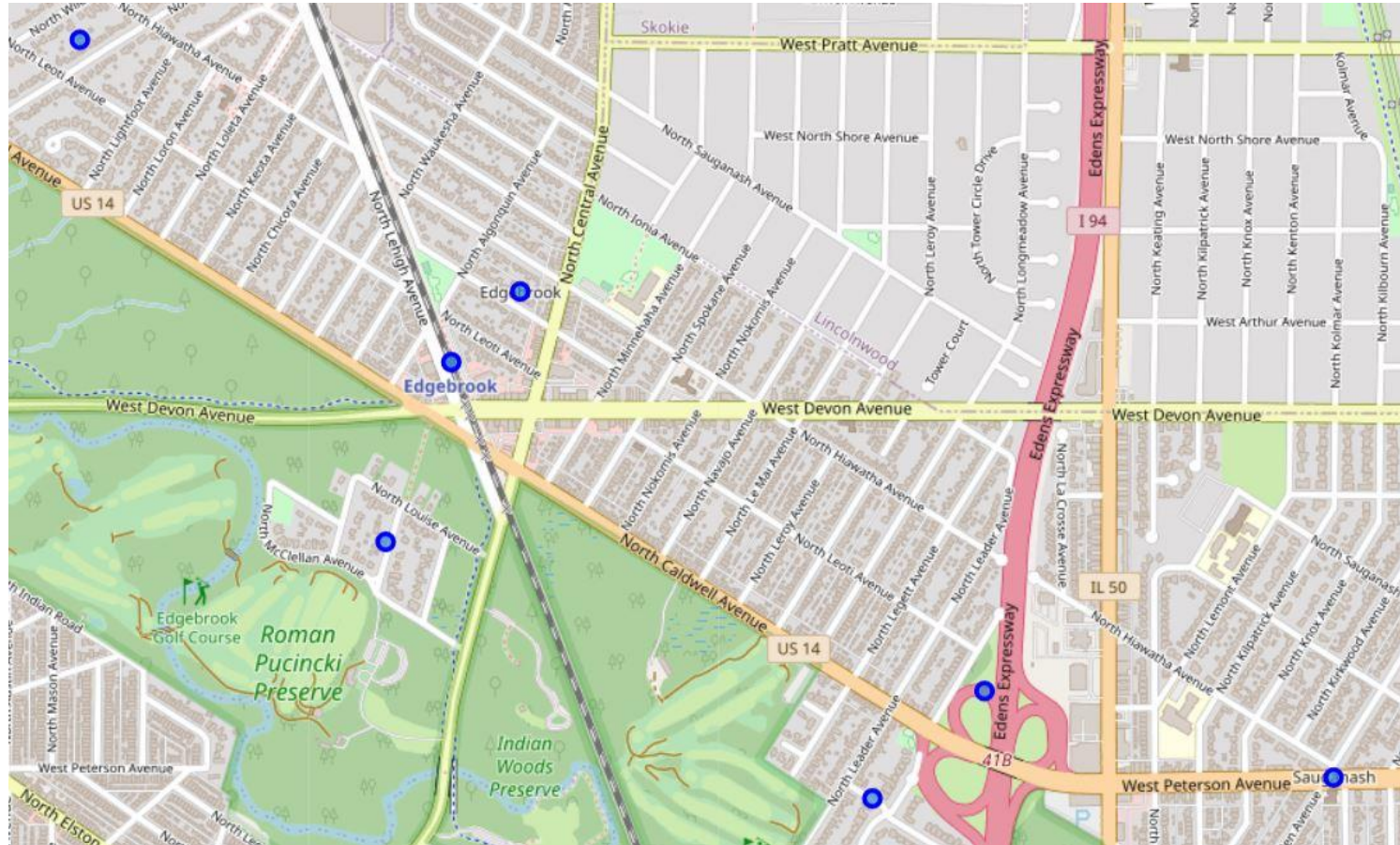
Comparing 5 communities with the highest crime rate it is evident that the community areas 25, 8, 32, 29 and 28 have the highest crime rate compared to other communities.

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

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.

MODELLING

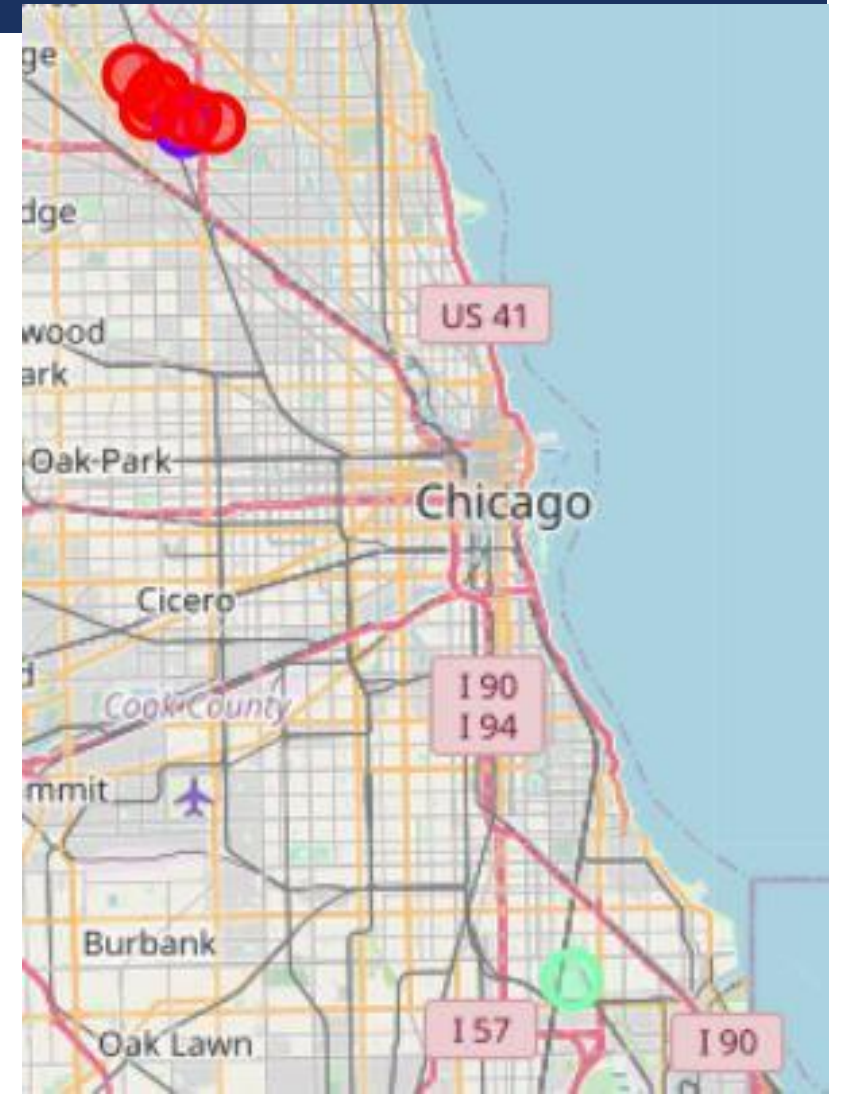
- Using the final dataset we find venues near each neighbourhood in Forest Glen and Burnside in a radius of 500 meters by connecting to the Foursquare API.

	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

- 
- 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.

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. Visualizing the clusters by using a map from folium library.
- Each cluster is colour coded for readability. Red represents the first cluster. Blue represents the second cluster. Green represents the third cluster.



CLUSTER I

	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 Common Venue
1	Edgebrook	Forest Glen	41.999677	-87.764100	0	Spa	Sandwich Place	American Restaurant	Plaza	Vietnamese Restaurant	Diner
2	North Edgebrook	Forest Glen	41.998269	-87.765976	0	Sandwich Place	Park	American Restaurant	Plaza	Hobby Shop	Grocery Store
4	Forest Glen	Forest Glen	41.991752	-87.751674	0	Yoga Studio	Indian Restaurant	Asian Restaurant	Coffee Shop	Fast Food Restaurant	Grocery Store
5	Old Edgebrook	Forest Glen	41.994708	-87.767727	0	Sandwich Place	Salon / Barbershop	Diner	Park	Coffee Shop	Barbershop
6	Wildwood	Forest Glen	42.004691	-87.775924	0	American Restaurant	Nature Preserve	Baseball Field	Theater	Park	Grocery Store
7	Sauganash	Forest Glen	41.990036	-87.742289	0	Park	Indian Restaurant	Asian Restaurant	Basketball Court	Pharmacy	Fast Food Restaurant

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.

CLUSTER 2

	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 Common 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

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.

CLUSTER 3

	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 Common Venue
0	Burnside	Burnside	41.730035	-87.596714	2	Train Station	Home Service	Bus Station	Intersection	Yoga Studio	Gas Station

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

DISCUSSION

- The aim of this project is to help people who want to relocate to the safest community in London, expats can choose the neighbourhoods to which they want to relocate based on the most common venues in it.
- If a person is looking for a neighbourhood with good connectivity and public transportation we can see that Cluster 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 are suitable.
- For a family I feel that the neighbourhoods in Cluster 2 are more suitable due 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.
- The preference of venues in their neighbourhood may vary from person to person, in this way they can select a neighbourhood according to their preference.

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.