Chicago data analysis

Problematic

XYZ

is a real estate investor. They are aiming to invest in several buildings in Chicago, IL to expand their business. They would like to invest only in the best neighbourhoods of Chicago, according to 3 criteria: Crime Rate, Amenity Rate and SEH rate

1

Data Acquisition and Preprocessing

Data Sources

1.Chicago community area wikipedia page

Used for web scraping

List of community areas [edit]

Chicago community areas by number, population, and area

Number[8]	Name[8]	2017 population[9]	Area (sq mi.)[10]	Area (km^2)	2017 population density (/sq mi.)	2017 population density (/km^2)
01	Rogers Park	55,062	1.84	4.77	29,925.00	11,554.11
02	West Ridge	76,215	3.53	9.14	21,590.65	8,336.20
03	Uptown	57,973	2.32	6.01	24,988.36	9,648.06
04	Lincoln Square	41,715	2.56	6.63	16,294.92	6,291.50
05	North Center	35,789	2.05	5.31	17,458.05	6,740.59
06	Lake View	100,470	3.12	8.08	32,201.92	12,433.23
07	Lincoln Park	67,710	3.16	8.18	21,427.22	8,273.10
08	Near North Side	88,893	2.74	7.10	32,442.70	12,526.20
09	Edison Park	11,605	1.13	2.93	4,235.40	1,635.30
10	Norwood Park	37,089	4.37	11.32	8,487.19	3,276.92
11	Jefferson Park	26,808	2.33	6.03	11,505.58	4,442.33
12	Forest Glen	19,019	3.20	8.29	5,943.44	2,294.78
13	North Park	18,842	2.52	6.53	7,476.98	2,886.88
14	Albany Park	51,992	1.92	4.97	27,079.17	10,455.33
15	Portage Park	64,307	3.95	10.23	16,280.25	6,285.84
16	Irving Park	54,606	3.21	8.31	17,011.21	6,568.06
17	Dunning	43,689	3.72	9.63	11,744.35	4,534.52
18	Montclare	13,830	0.99	2.56	13,969.70	5,393.73
19	Belmont Cragin	79,910	3.91	10.13	20,437.34	7,890.90
20	Hermosa	24,144	1.17	3.03	20,635.90	7,967.57
21	Avondale	37,368	1.98	5.13	18,872.73	7,286.80
22	Logan Square	73,046	3.59	9.30	20,347.08	7,856.05
23	Humboldt Park	56,427	3.60	9.32	15,674.17	6,051.83
24	West Town	84,502	4.58	11.86	18,450.22	7,123.67
25	Austin	95,260	7.15	18.52	13,323.08	5,144.07
26	West Garfield Park	17,163	1.28	3.32	13,408.59	5,177.09

Data Sources

2.Chicago crime data from 2018

Used for crime analysis

CHICAGO DATA PORTAL

Browse Tuto

Crimes - 2018

Based on Crimes - 2001 to Present

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City

More View

ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description
11561837	JC110056	12/31/2018 11:59:00 ...	013XX W 72ND ST	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 3...	
11556487	JC104662	12/31/2018 11:59:00 ...	112XX S SACRAMENTO ...	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET
11552699	JC100043	12/31/2018 11:57:00 ...	084XX S SANGAMON ST	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT
11552724	JC100006	12/31/2018 11:56:00 ...	018XX S ALLPORT ST	0440	BATTERY	AGG: HANDS/FIST/FEET NO/MINOR I...	OTHER
11552731	JC100031	12/31/2018 11:55:00 ...	078XX S SANGAMON ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT
11552715	JC100026	12/31/2018 11:49:00 ...	052XX W GLADYS AVE	041A	BATTERY	AGGRAVATED - HANDGUN	STREET
11552741	JC100011	12/31/2018 11:48:00 ...	079XX S LAFLIN ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT
11552602	JC100089	12/31/2018 11:47:00 ...	018XX W BELMONT AVE	0460	BATTERY	SIMPLE	VEHICLE - OTHER RIDE SHARE SERVICE (E.G., UBER, LYFT)
11554852	JC101652	12/31/2018 11:45:00 ...	032XX W EVERGREEN AVE	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT
11553488	JC101094	12/31/2018 11:45:00 ...	032XX N SHEFFIELD AVE	0890	THEFT	FROM BUILDING	BAR OR TAVERN
11552570	JB574407	12/31/2018 11:44:00 ...	047XX N RACINE AVE	1330	CRIMINAL TRESPASS	TO LAND	MOVIE HOUSE/THEATER
11552603	JC100036	12/31/2018 11:43:00 ...	071XX S VINCENNES AVE	143A	WEAPONS VIOLATION	UNLAWFUL POSS OF HANDGUN	GAS STATION
11552568	JC100004	12/31/2018 11:42:00 ...	030XX N MARMORA AVE	2022	NARCOTICS	POSS: COCAINE	ALLEY
11552739	JC100117	12/31/2018 11:40:00 ...	088XX S COTTAGE GRO...	1477	WEAPONS VIOLATION	RECKLESS FIREARM DISCHARGE	STREET
11552597	JC100013	12/31/2018 11:40:00 ...	010XX W PRATT BLVD	0460	BATTERY	SIMPLE	PARK PROPERTY
11555648	JC102651	12/31/2018 11:30:00 ...	001XX W HUBBARD ST	0870	THEFT	POCKET-PICKING	BAR OR TAVERN
11553486	JC101095	12/31/2018 11:30:00 ...	032XX N Sheffield Ave	1150	DECEPTIVE PRACTICE	CREDIT CARD FRAUD	BAR OR TAVERN
11553209	JC100784	12/31/2018 11:30:00 ...	036XX W BELLE PLAINE ...	1320	CRIMINAL DAMAGE	TO VEHICLE	VEHICLE NON-COMMERCIAL
11553037	JC100611	12/31/2018 11:30:00 ...	012XX W 102ND PL	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE
11552657	JC100019	12/31/2018 11:30:00 ...	004XX S CLINTON ST	0313	ROBBERY	ARMED: OTHER DANGEROUS WEAPON	CTA PLATFORM
11552637	JB574400	12/31/2018 11:30:00 ...	085XX S MORGAN ST	1822	NARCOTICS	MANU/DEL:CANNABIS OVER 10 GMS	STREET

< Previous Next >

Showing rows 1 to 100 out of 268,293

Data Sources

3.Public Health Statistics of Chicago

Used for SEH rate

	Community Area	Community Area Name	Birth Rate	General Fertility Rate	Low Birth Weight	Prenatal Care Beginning in First Trimester	Preterm Births	Teen Birth Rate	Assault (Homicide)	Breast cancer in females	...	Childhood Lead Poisoning	Gonorrhea in Females	Gonorrhea in Males	Tuberculosis	B Po l
0	1	Rogers Park	16.4	62.0	11.0	73.0	11.2	40.8	7.7	23.3	...	0.5	322.5	423.3	11.4	
1	2	West Ridge	17.3	83.3	8.1	71.1	8.3	29.9	5.8	20.2	...	1.0	141.0	205.7	8.9	
2	3	Uptown	13.1	50.5	8.3	77.7	10.3	35.1	5.4	21.3	...	0.5	170.8	468.7	13.6	
3	4	Lincoln Square	17.1	61.0	8.1	80.5	9.7	38.4	5.0	21.7	...	0.4	98.8	195.5	8.5	
4	5	North Center	22.4	76.2	9.1	80.4	9.8	8.4	1.0	16.6	...	0.9	85.4	188.6	1.9	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
72	73	Washington Heights	12.0	61.0	19.6	75.4	16.2	65.0	38.0	47.9	...	1.5	1298.2	1274.2	3.0	
73	74	Mount Greenwood	12.5	59.0	8.4	94.5	15.1	7.7	2.2	34.6	...	0.0	NaN	.	0.0	
74	75	Morgan Park	13.2	67.5	10.6	74.5	12.3	46.7	19.9	32.4	...	1.3	800.5	741.1	2.6	
75	76	O'Hare	15.8	70.0	3.5	82.0	5.0	15.9	5.6	20.5	...	0.5	NaN	.	6.3	
76	77	Edgewater	12.1	48.1	7.5	76.1	7.4	15.1	5.8	18.5	...	0.9	120.1	427.5	10.5	

77 rows × 29 columns

Data Sources

4.FOURSQUARE API

Used for amenity rate

	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	...	Vineyard	Warehouse Store	Waterfront	Weigl Los Cente
AUSTIN	2.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
NEAR NORTH SIDE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
LOOP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0
NEAR WEST SIDE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
NORTH LAWNDALE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
MONTCLARE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
MOUNT GREENWOOD	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
FOREST GLEN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
BURNSIDE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
EDISON PARK	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0

77 rows × 363 columns

Data Preprocessing

Crime Database

22 columns initially

11 were removed.

Some rows deleted because no Location

SEH database

29 columns initially

8 kept because others too specific

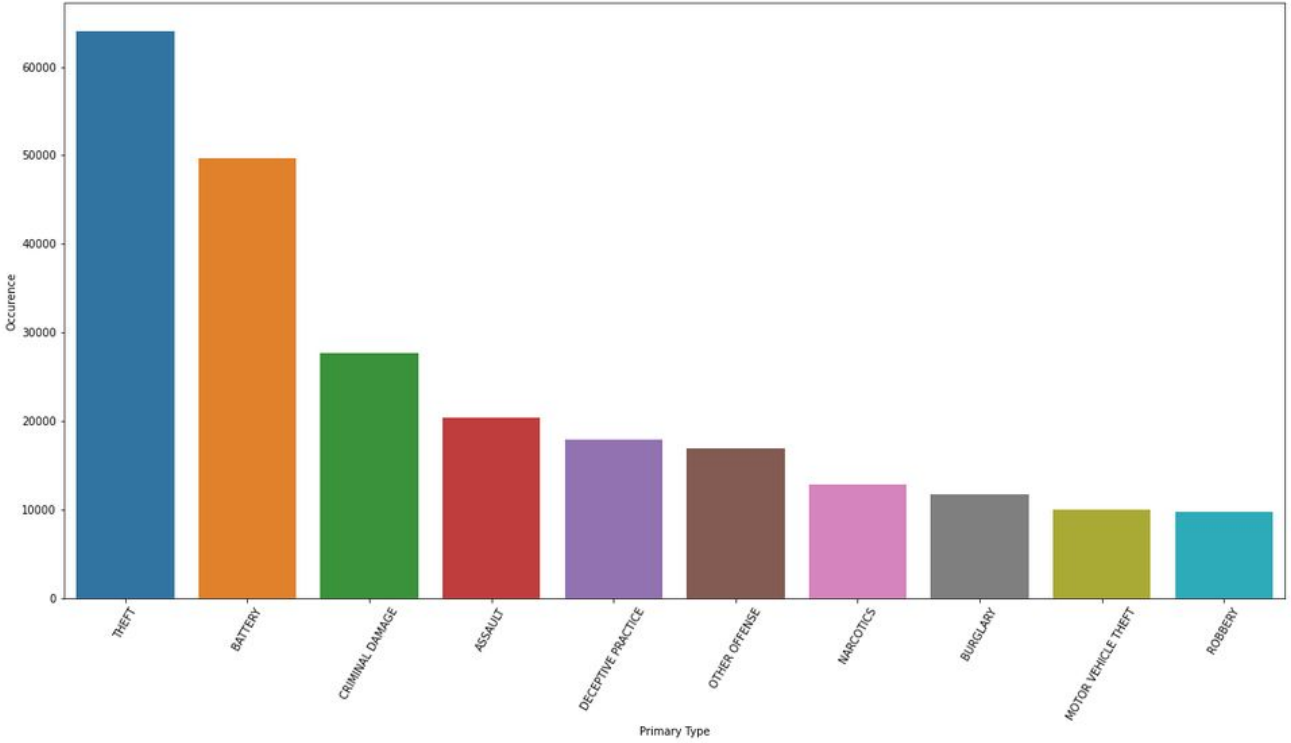
Foursquare API

It returns a json file, the main part of the work using it is extracting the information out of the json file, to create a dataframe

2

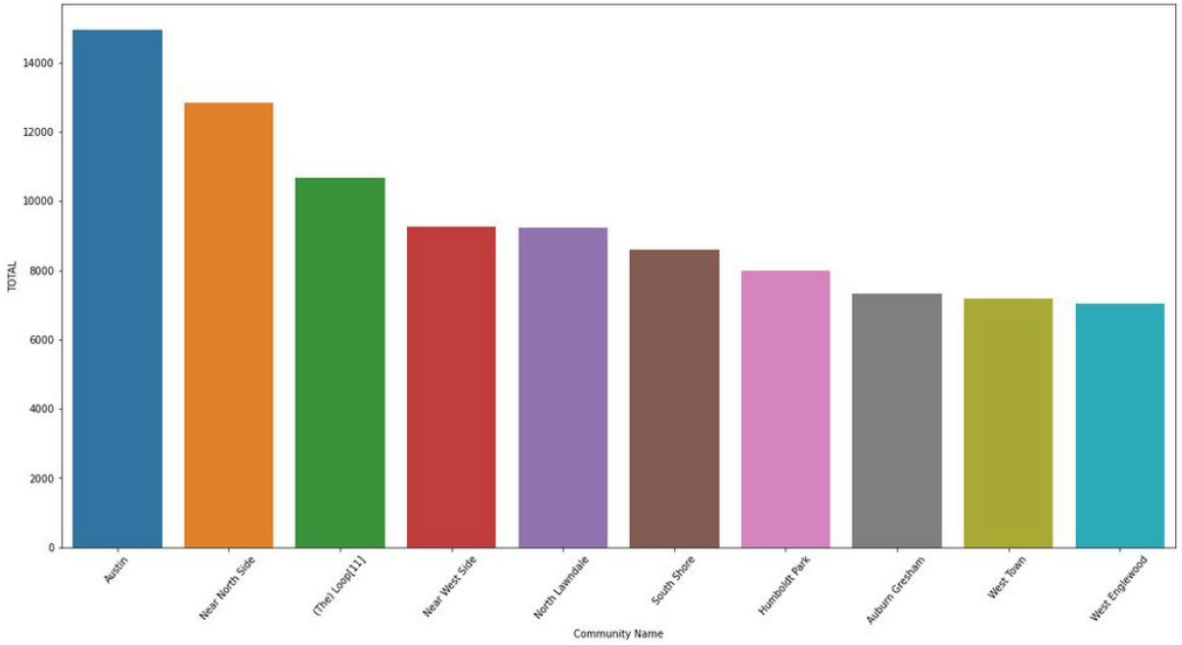
Data Visualization

Type of crime occurrences in Chicago



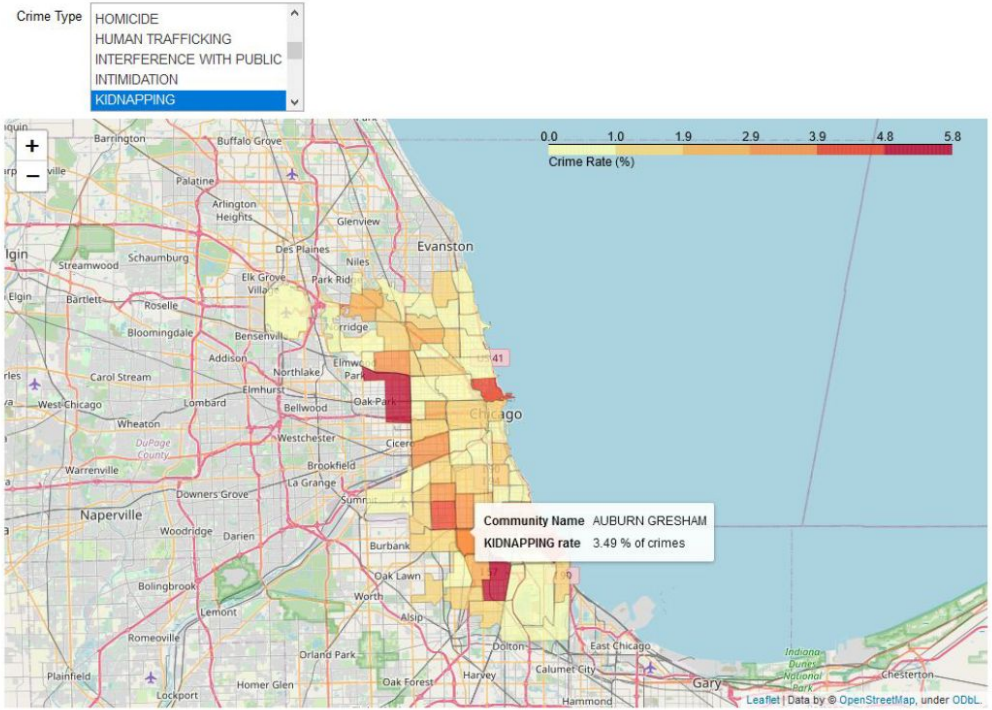
Most of the crimes are theft or battery

Number of crime occurrence by Community areas



The areas where most crime are committed are shown here. 1.Austin, 2. Near North Side and 3. The Loop

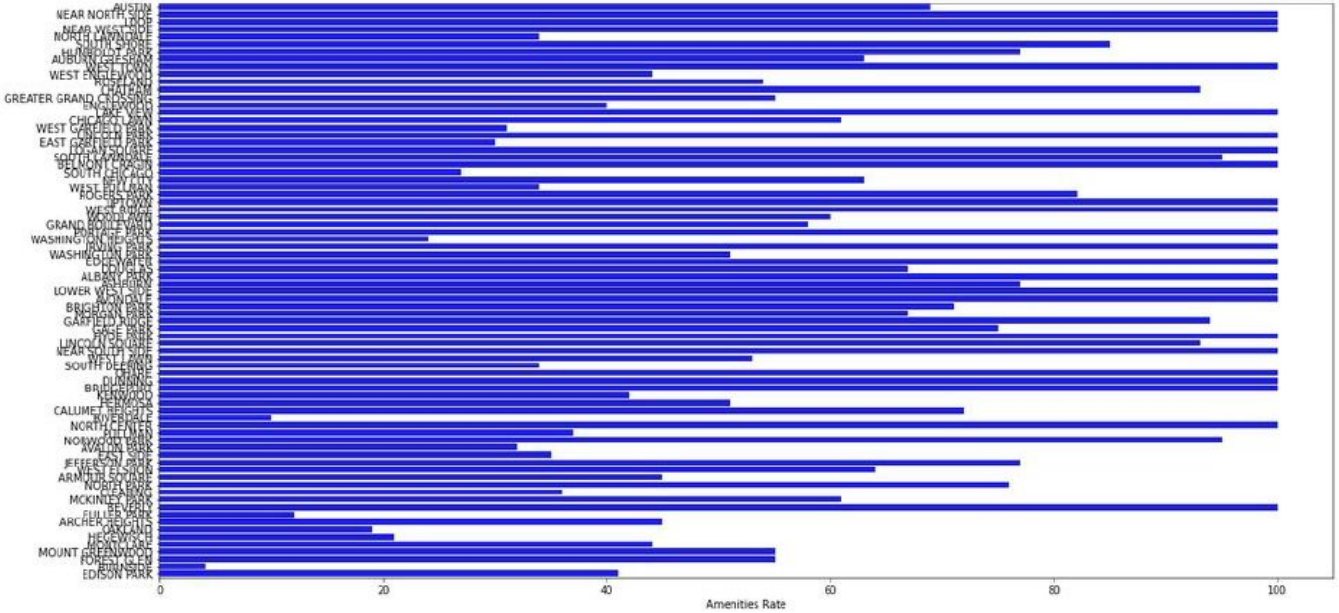
Choropleth map of the crime rate per crime type by community area



Finally here is an interactive map showing the different crime rate by community area, and categorized by crime type

Number of amenities per community area

```
Out[41]: <AxesSubplot:xlabel='Amenities Rate'>
```



Normalized dataframe for the SEH rate

Community Area Name	Birth Rate	General Fertility Rate	Teen Birth Rate	Cancer (All Sites)	Infant Mortality Rate	Below Poverty Level	Per Capita Income	Unemployment	SEH Rate
North Center	1.000000	0.721726	0.938581	0.813302	0.943128	0.931389	0.547896	0.991620	6.887644
Near South Side	0.923077	0.672619	0.570934	0.714702	0.843602	0.862779	0.662080	0.958101	6.207894
O'Hare	0.492308	0.629464	0.873702	0.892649	0.976303	0.890223	0.265389	0.986034	6.006072
Lincoln Park	0.292308	0.163690	0.993080	0.808635	0.957346	0.850772	0.799562	0.991620	5.857013
Lake View	0.315385	0.163690	0.874567	0.960327	0.966825	0.873070	0.631989	0.986034	5.771886
...	...	...	...	...	...	...	...	...	...
Douglas	0.069231	0.215774	0.715398	0.126021	0.436019	0.605489	0.185214	0.650838	3.003983
Washington Park	0.761538	0.660714	0.296713	0.195449	0.156398	0.382504	0.057893	0.469274	2.980484
West Garfield Park	0.823077	0.903274	0.017301	0.000000	0.170616	0.361921	0.030727	0.413408	2.720324
Riverdale	0.238462	0.273810	0.453287	0.193699	0.658768	0.000000	0.000000	0.379888	2.197913
Fuller Park	0.192308	0.486607	0.412630	0.190198	0.000000	0.101201	0.006117	0.000000	1.389061

77 rows × 9 columns

3

Clustering with K-means algorithm

1st step

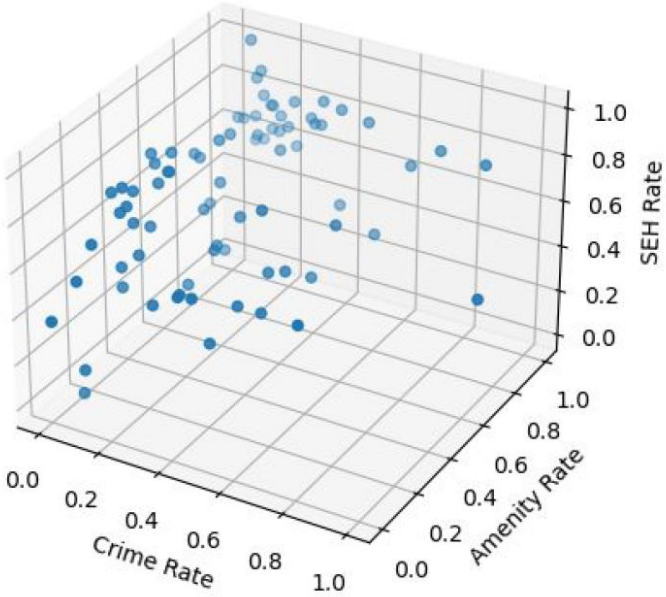
Grouping every information required into one dataframe

	Crime Rate	Amenity Rate	SEH Rate
AUSTIN	1.000000	0.684211	0.397198
NEAR NORTH SIDE	0.856631	1.000000	0.723216
LOOP	0.709677	1.000000	0.732392
NEAR WEST SIDE	0.612903	1.000000	0.632709
NORTH LAWNDALE	0.612903	0.336842	0.369960
...	...	...	...
MONTCLARE	0.023297	0.357895	0.726861
MOUNT GREENWOOD	0.021505	0.536842	0.718946
FOREST GLEN	0.016129	0.526316	0.767015
BURNSIDE	0.008961	0.000000	0.377331
EDISON PARK	0.000000	0.336842	0.711122

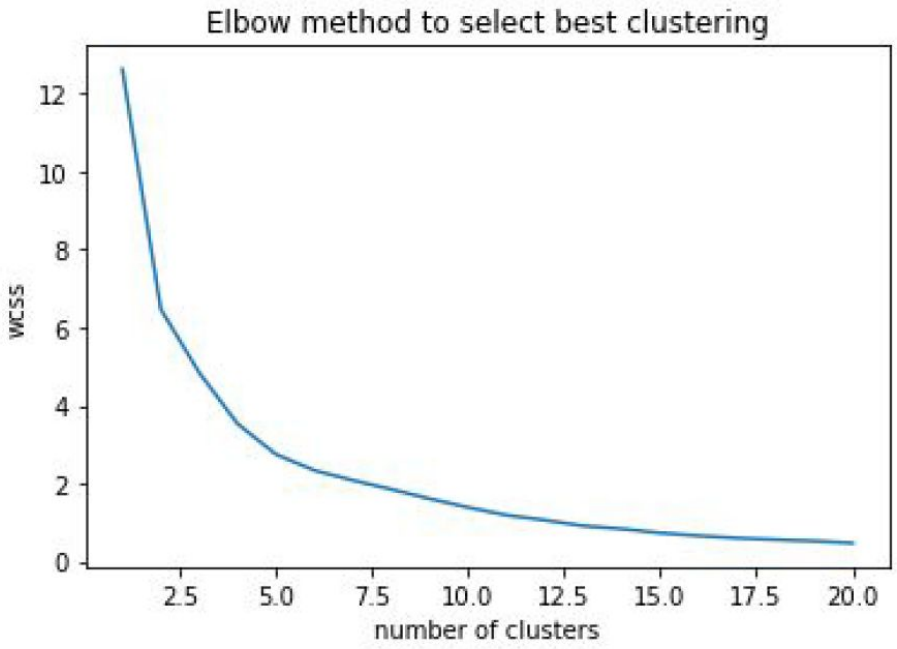
77 rows × 3 columns

Here, the dataframe contains every information standardized for the k-means clustering

Visualizing the points in 3D before clustering

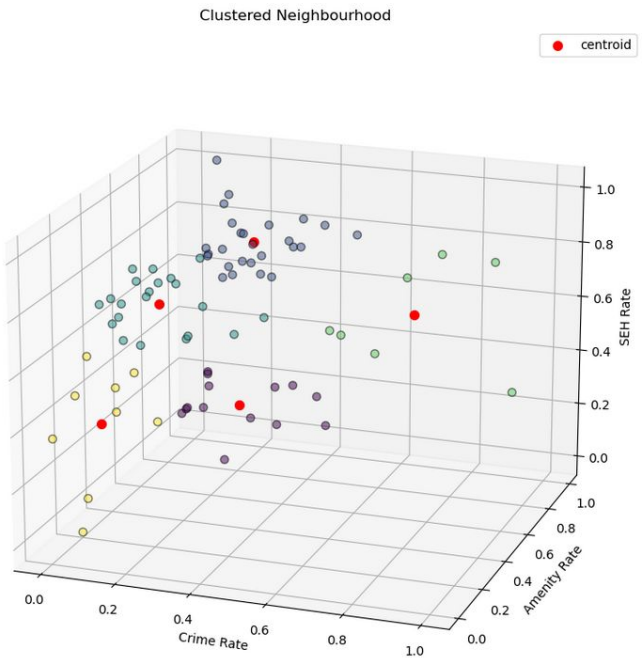


Applying the elbow method to find the best k



Here, the best k could be either 2 or 5. However, 2 clusters are not enough, so 5 is selected

Visualizing the clusters in 3d with the k chosen



5 clusters are displayed, all include the centroids

5th step

Finding the common trait between each clusters

	Crime Rate	Amenity Rate	SEH Rate	Cluster
NORTH LAWNDALE	0.612903	0.336842	0.369960	0
AUBURN GRESHAM	0.482079	0.642105	0.320611	0
WEST ENGLEWOOD	0.462366	0.389474	0.326521	0
ROSELAND	0.449821	0.547368	0.398061	0
GREATER GRAND CROSSING	0.412186	0.515789	0.399293	0
ENGLEWOOD	0.403226	0.357895	0.355040	0
WEST GARFIELD PARK	0.367384	0.263158	0.242110	0
EAST GARFIELD PARK	0.315412	0.252632	0.425576	0
SOUTH CHICAGO	0.275986	0.242105	0.423360	0
WEST PULLMAN	0.256272	0.284211	0.397083	0
WOODLAWN	0.218638	0.536842	0.409870	0
GRAND BOULEVARD	0.202509	0.578947	0.397419	0
WASHINGTON PARK	0.173835	0.463158	0.289424	0
DOUGLAS	0.164875	0.684211	0.293698	0

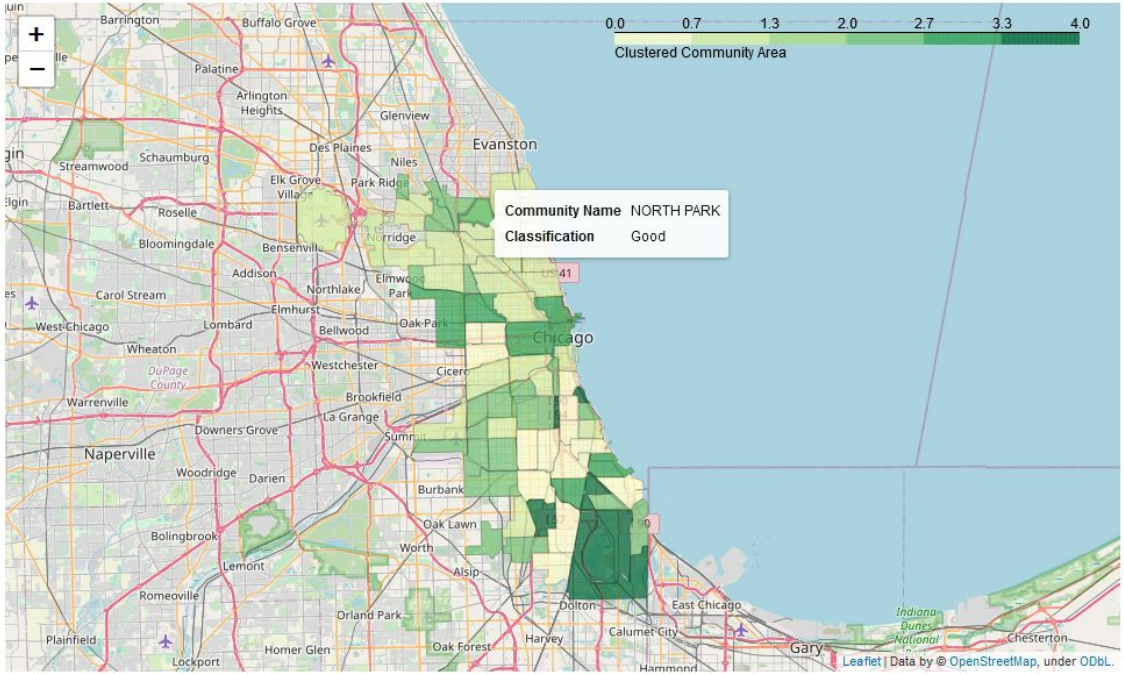
	Crime Rate	Amenity Rate	SEH Rate	Cluster
CHICAGO LAWN	0.367384	0.557895	0.627607	2
NEW CITY	0.268817	0.600000	0.535169	2
ASHBURN	0.145161	0.694737	0.581548	2
BRIGHTON PARK	0.130824	0.715789	0.749180	2
MORGAN PARK	0.127240	0.631579	0.495481	2
WEST LAWN	0.103943	0.484211	0.745243	2
KENWOOD	0.082437	0.410526	0.546181	2
HERMOSA	0.078853	0.463158	0.699800	2
CALUMET HEIGHTS	0.077061	0.747368	0.429901	2
EAST SIDE	0.057348	0.326316	0.677891	2
JEFFERSON PARK	0.057348	0.726316	0.640421	2
WEST ELSDON	0.051971	0.578947	0.751688	2
ARMOUR SQUARE	0.051971	0.368421	0.576640	2
NORTH PARK	0.051971	0.705263	0.668448	2
CLEARING	0.048387	0.305263	0.661555	2
MCKINLEY PARK	0.044803	0.568421	0.671773	2
ARCHER HEIGHTS	0.039427	0.389474	0.697873	2
MONTCLARE	0.023297	0.357895	0.726861	2
MOUNT GREENWOOD	0.021505	0.536842	0.718946	2
FOREST GLEN	0.016129	0.526316	0.767015	2
EDISON PARK	0.000000	0.336842	0.711122	2

	Crime Rate	Amenity Rate	SEH Rate	Cluster
WEST TOWN	0.471326	1.000000	0.773020	1
LAKE VIEW	0.379928	1.000000	0.797083	1
LINCOLN PARK	0.318996	1.000000	0.812564	1
LOGAN SQUARE	0.311828	1.000000	0.706920	1
SOUTH LAWNDALE	0.290323	1.000000	0.704243	1
BELMONT CRAGIN	0.277778	1.000000	0.724823	1
ROGERS PARK	0.238351	0.884211	0.640016	1
UPTOWN	0.227599	1.000000	0.583013	1
WEST RIDGE	0.220430	1.000000	0.777153	1
PORTAGE PARK	0.198925	1.000000	0.658743	1
IRVING PARK	0.173835	1.000000	0.698676	1
EDGEWATER	0.168459	1.000000	0.627904	1
ALBANY PARK	0.146953	1.000000	0.734974	1
LOWER WEST SIDE	0.143369	1.000000	0.630634	1
AVONDALE	0.139785	1.000000	0.736652	1
GARFIELD RIDGE	0.123656	0.947368	0.628112	1
GAGE PARK	0.123656	0.789474	0.728314	1
HYDE PARK	0.114695	1.000000	0.575905	1
LINCOLN SQUARE	0.114695	1.000000	0.770436	1
NEAR SOUTH SIDE	0.105735	1.000000	0.876377	1
OHARE	0.091398	1.000000	0.839673	1
DUNNING	0.086022	1.000000	0.668504	1
BRIDGEPORT	0.086022	1.000000	0.561548	1
NORTH CENTER	0.071685	1.000000	1.000000	1
NORWOOD PARK	0.062724	0.936842	0.693322	1
BEVERLY	0.044803	1.000000	0.644411	1

The main objective here is to understand the clustering segmentation done by the k means algorithm to be able to label the clusters

Last Step

Showing a map allowing XYZ to have the best overview for their problem



On this map, label clustered are shown, and each color represent a cluster. XYZ can then chose the neighbourhood they prefer and matching their high standard needs.

Conclusion

- From one defined problem, data needed have been identified then sourced.
- After cleaning the data, different chart and map have been displayed to help the company understanding of the data
- To end with, an interactive map have been made, allowing XYZ to choose the best neighbourhood suiting their needs.

THANK YOU