Data Analysis of the Chicago community areas

Valentin FORMONT, 01/02/2021

INTRODUCTION:

The company XYZ is a real estate investor. They are aiming to invest in several buildings in Chicago, IL to expand their business. XYZ has a company guideline of investing only in the so-called 'best' neighbourhoods of cities, which have built their reputation as a good real estate company. Regardless of the city, they offer safe and dynamic places to live for their customers.

To do so, they uses 3 indicators:

- Crime rate

- SEH rate (socio-economic and human rate)

- Amenity rate

XYZ intends to only offer places with low crime rate, high SEH rate and high amenity rate. However, they lack information about the city of Chicago, and are not able to tell which areas they should invest in. They open a business opportunity to those who are ready to work on this subject.

1. DATA ACQUISITION AND PREPROCESSING:

1.1 Data sources:

In order to answer this business opportunity, it is needed to leverage several databases.

To fit the community area number with the formal names, a part of web scraping is done on the Chicago community area wikipedia page.

One is the Chicago crime data from 2018, where every single event is reported. This is a 268282 rows × 22 columns csv. For this analysis, what will be relevant is the community area where a crime happened. But we will have a further analysis using the type of crime committed in the area. This will define the safe and risky places. The crime analysis will include maps so XYZ can understand quickly which areas are subject to crime and which aren't.

Then , using the Public Health Statistics of Chicago, a SEH factor will be estimated, based on mixing health information and socio-economic data. To be accurate, it comprises the birth rate, the number of cancers, the infant mortality rate, the number of people unemployed, the number of people below the level of poverty and the per capita income. Literally thousands of other factors could have been taken into account, but for this analysis, it has been restricted to these ones

Then, the Foursquare API is used to get the number of amenities for each neighbourhood. This way, it shows the amenity rate for each neighbourhood.

To end with, with these 3 factors, a K Means algorithm will be applied to find the different kinds of neighbourhood. The goal of this clustering is to find the 'ideal' community areas for XYZ. Thus, a heatmap will be created with the result of the clustering algorithm, allowing the company to have an overview of the city with its clusters.

1.2 Data Preprocessing:

As the project was elaborated and improved while it was being done, there are analyses for the 3 different factors, treated as single entities, then a common analysis.

For the crime table, initially 22 columns were included in the dataset. However, half of it was irrelevant or duplicated info, and 11 were removed:

['Block','Beat','District','Ward','FBI Code', 'X Coordinate', 'Y Coordinate','Updated On', 'Year','ID','IUCR']

‘Block’,’Beat’,’District’ and ‘Ward’ are either related to the police patrol number, or are information too detailed for the analysis. ‘Updated on’, ‘Year’, ‘ID’, and ‘IUCR’ are irrelevant for this analysis. Finally X and Y Coordinate are just duplicates of the Location column.

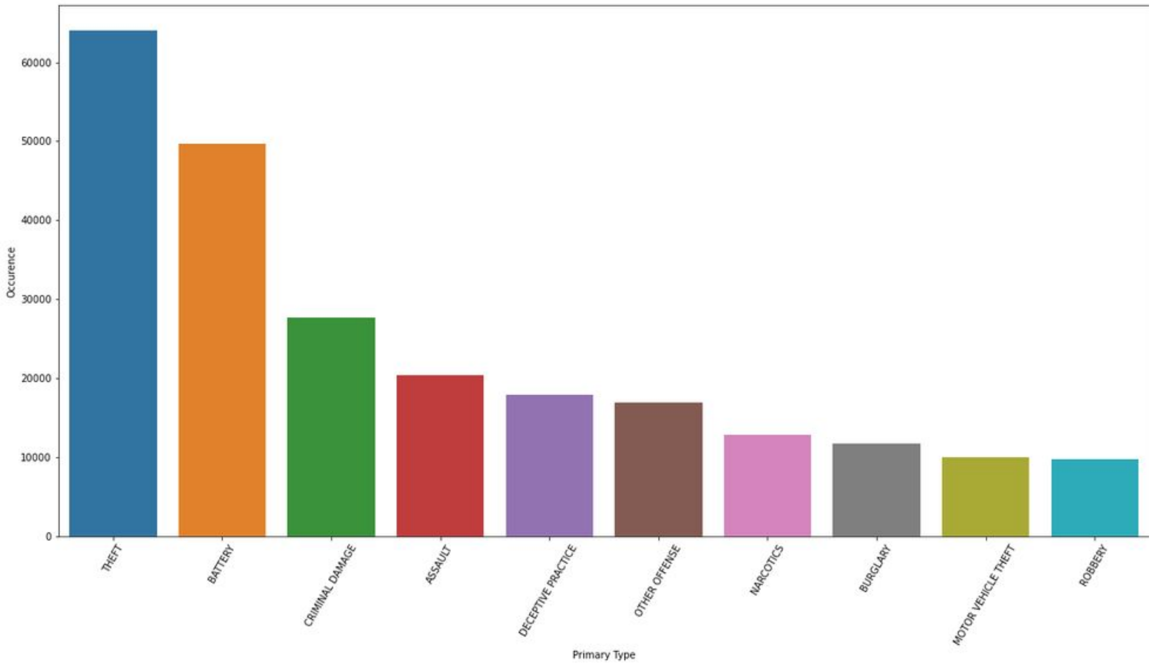
A small fraction of the data had missing locations, thus, were removed.

The Public Health Statistics of Chicago dataset, included 29 columns. Only 8 were kept. Most of the dumped data were too specifical to be taken into such a general analysis (ie. Preterm Births, or Childhood lead poisoning). Other features could have been used, but the aim of this project was not to do a Health in depth analysis, but just extracting a SEH factor from it.

2.DATA VISUALIZATION

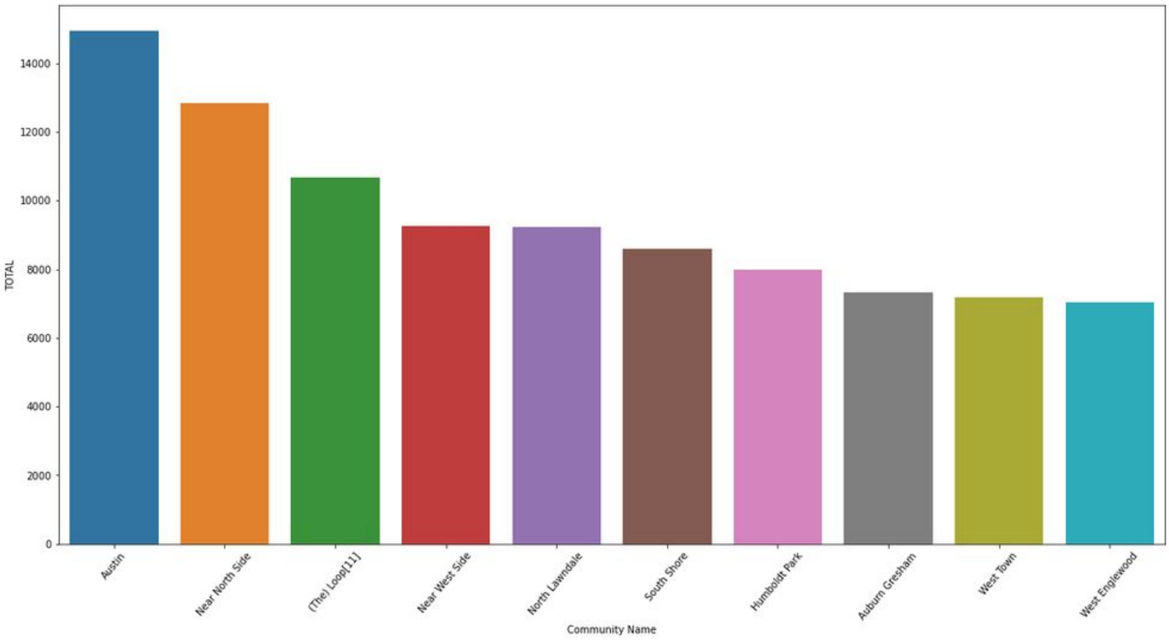
2.1 Crime related data:

The first part of this analysis comprises the visualization of the crime data. The very first thing made was to plot the different kinds of crime on a chart, to have a better understanding of the full picture. The result represents the number of occurence a kind of crime that appeared in the dataframe.



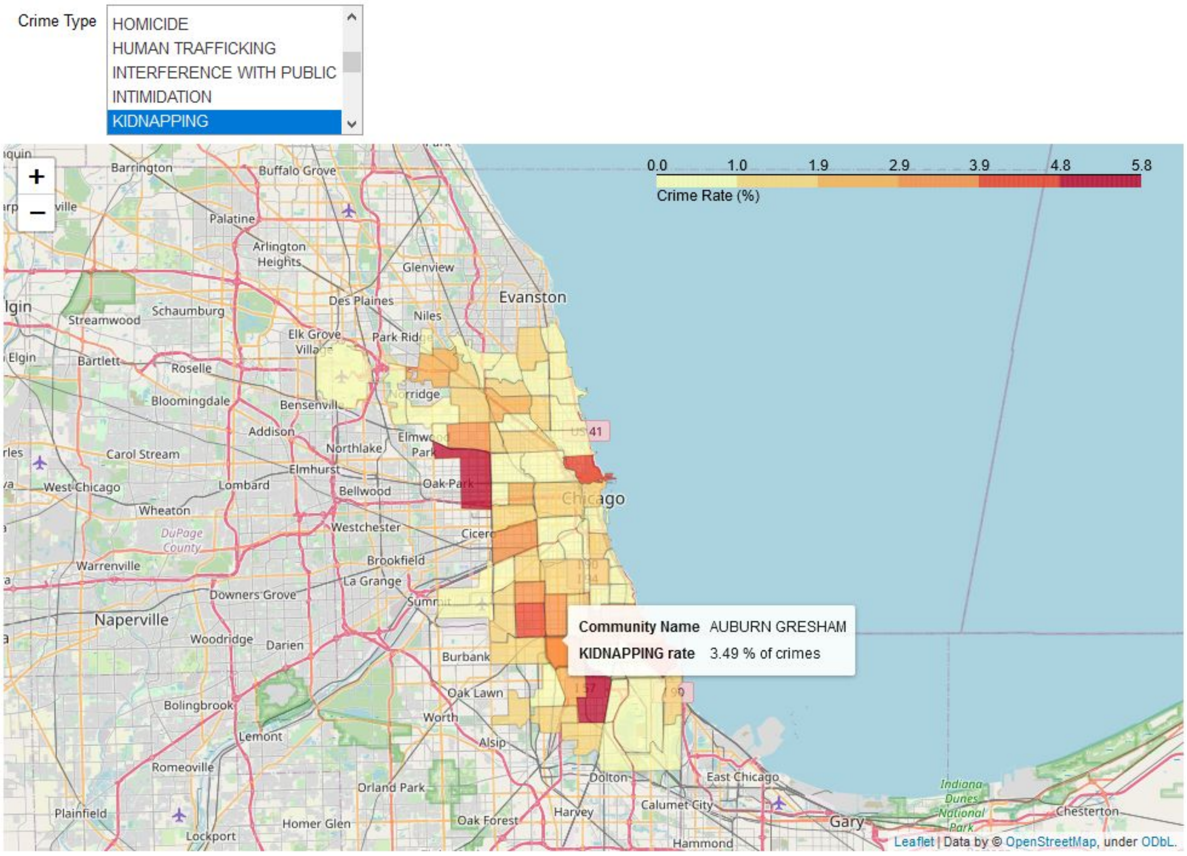
The final objective of the crime section is to have a choropleth map of Chicago related to the crime density. In order to have such a map, it is needed to group the data frame by community area.

Once the operation of one hot encoding and grouping is done , the graph reveals the most dangerous neighbourhoods in Chicago



Once it is done, each of these values are turned into a percentage, and the Choropleth map can be displayed

The map can be explored by hovering on the different neighbourhoods. It clearly shows the area where the crime rate in total is higher and lower. To make this map even more detailed, the choice of crime type is let to the user.



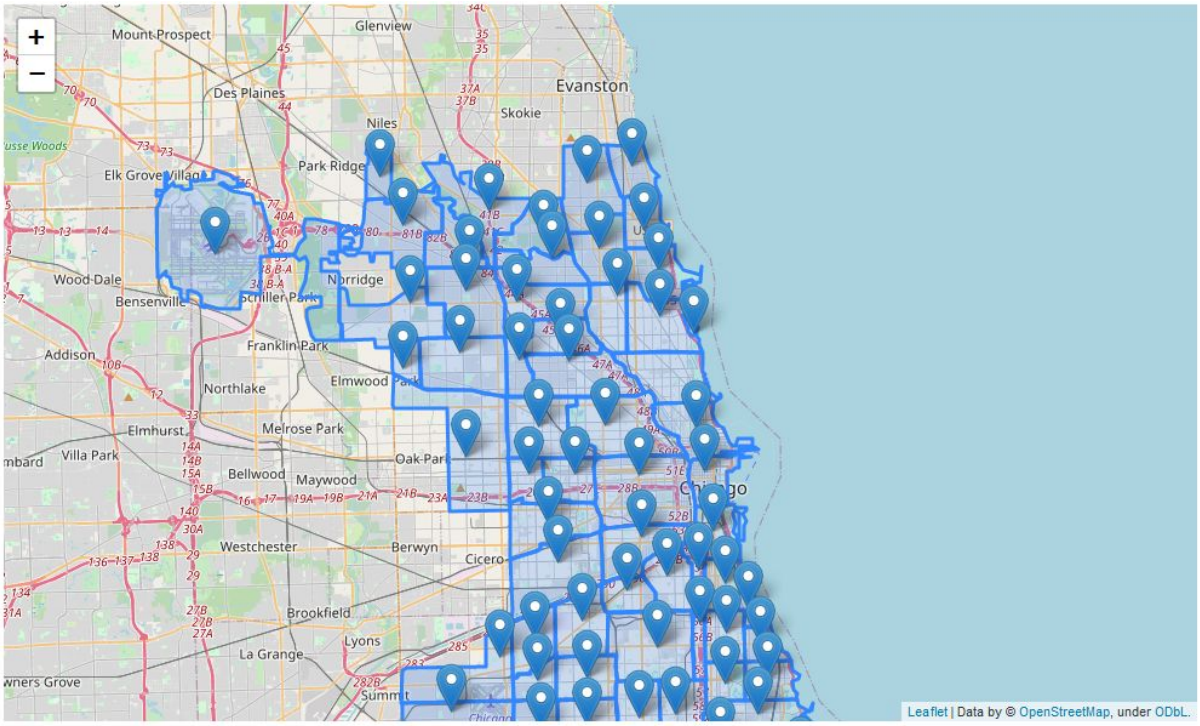
Crime Type
HOMICIDE
HUMAN TRAFFICKING
INTERFERENCE WITH PUBLIC
INTIMIDATION
KIDNAPPING

Community Name AUBURN GRESHAM
KIDNAPPING rate 3.49 % of crimes

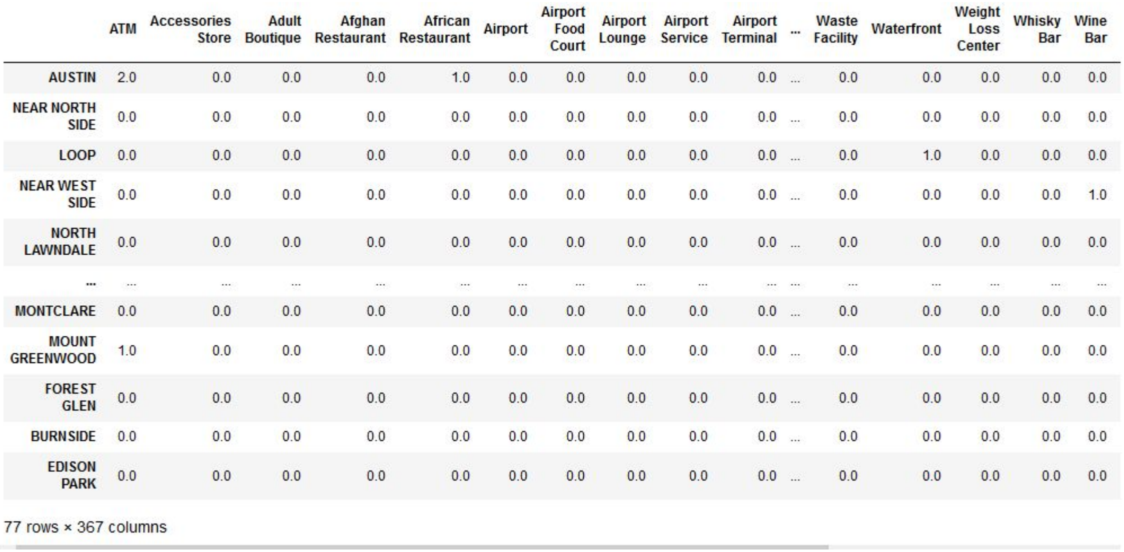
2.1 Amenity related data:

Now that the data about crime have been treated, it is expected to get the amenity rate from each neighbourhood. To do so , it is first needed to get the center point of each community area, using Nominatim.

A quick check on the map, allows to see if the locations are correct or totally wrong:



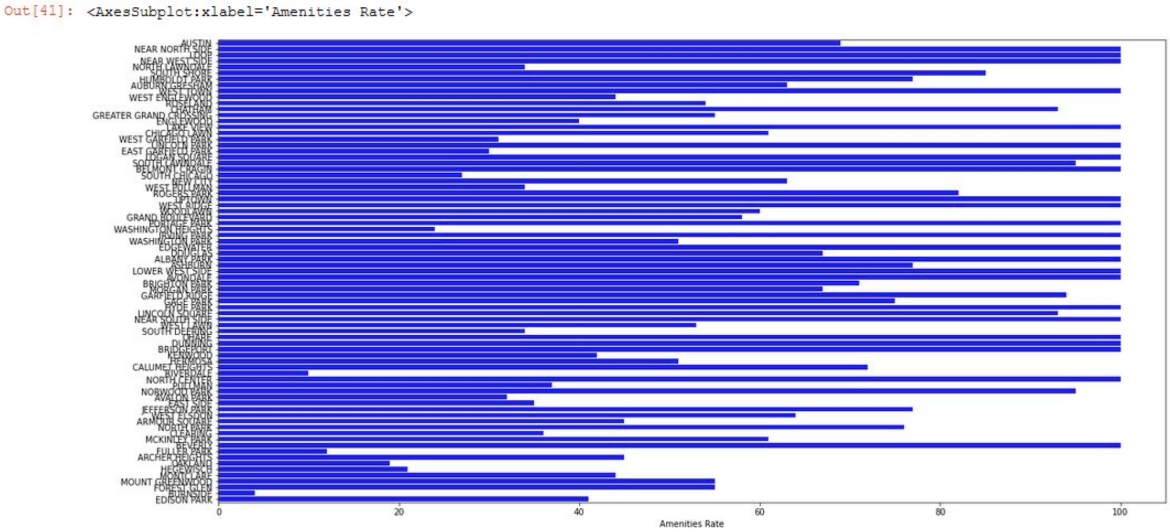
Once used, the community area location is inserted into an algorithm creating one dataframe row, containing a maximum of 100 amenities near the center point of a community area. The radius used is calculated based on the size of each CA. A loop is applied to process every community area. The resulting data frame is the following:



	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	...	Waste Facility	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar
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	0.0	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	0.0	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	1.0	0.0	0.0	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	0.0	1.0
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	0.0	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	0.0	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	0.0	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	0.0	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	0.0	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	0.0	0.0

77 rows × 367 columns

Once this dataframe is obtained, the sum of the column is added as a last column, and a bar chart shows the number of amenities within the radius the function calculated .



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

The main interest of this bar chart is to spot anomalies. Typically, a 0 value for an area would be most likely related to an error of algorithm.

The method to get the number of amenities could be criticized. The ideal would have been to search within the real boundaries of each community area. However the Foursquare API does not allow this kind of research, as it is a research per radius and center point. This is good for a relatively square and round neighbourhood, but can result in false results for long geometry (such as long and thin rectangle) because having a radius on this kind of shape has little sense.

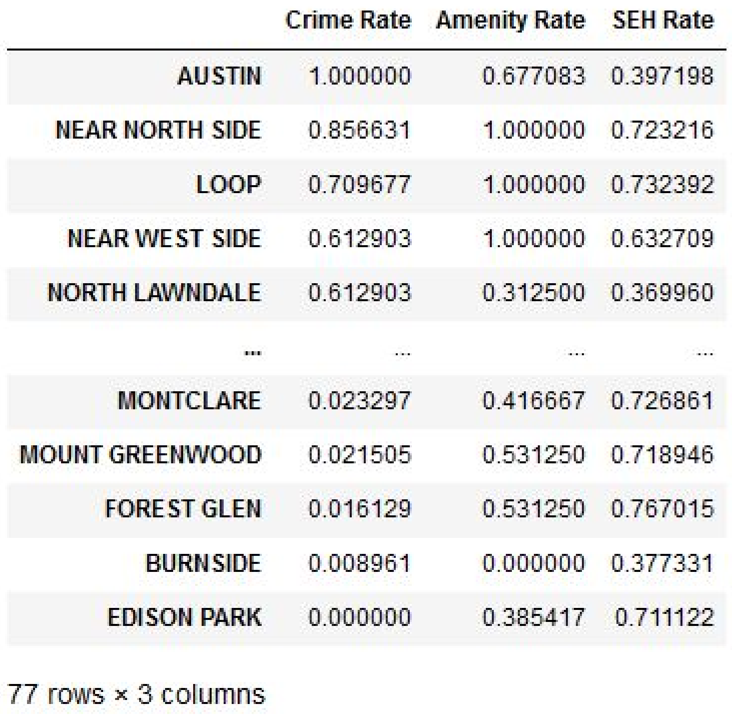
2.3 SEH related data:

For the SEH data, the analysis is most straight forward. Once the data has been preprocessed and standardized, a SEH rate has been directly added to the dataframe, calculated by summing the standardized values.

3. K Means Algorithm

The main purpose of this paper is to provide a map with the clustered community areas of Chicago according to the XYZ rating system.

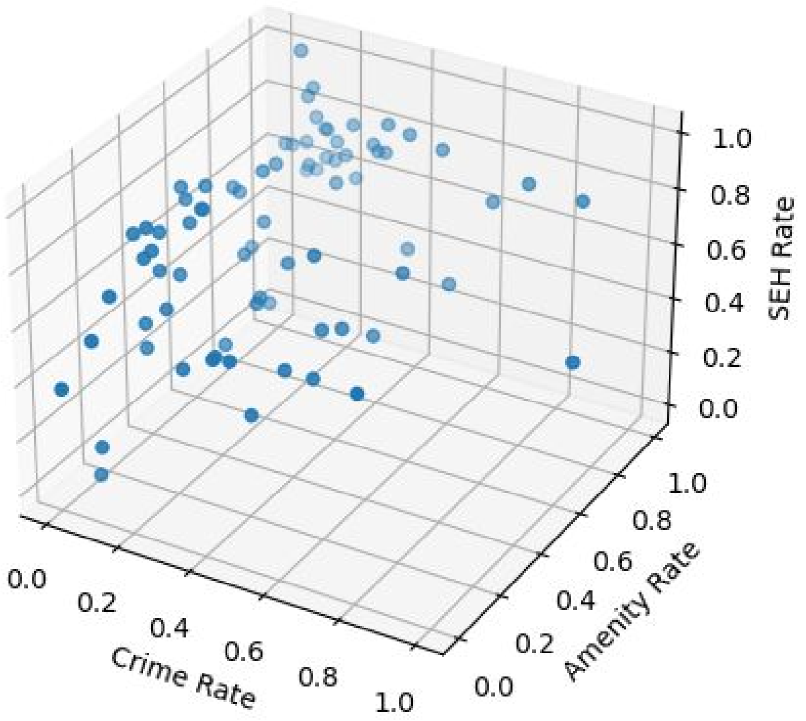
The standardized data frame comprises the Crime rate (0 to 1, 1 being the worst), Amenity rate (0 to 1, 1 being the best), SEH Rate (0 to 1, 1 being the best).



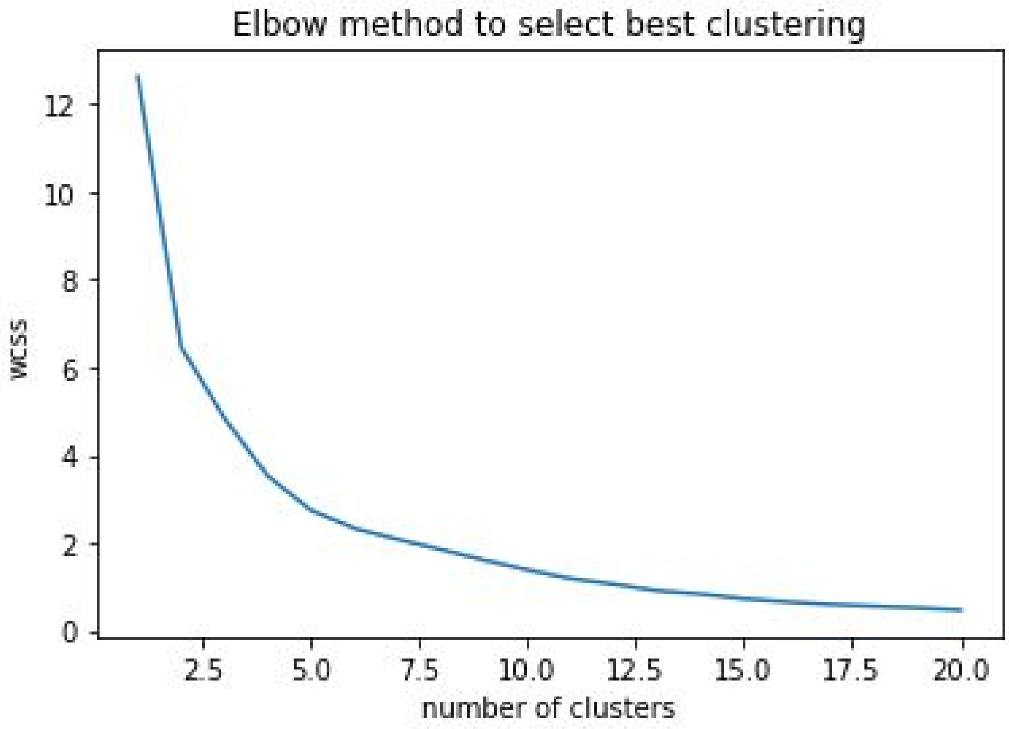
	Crime Rate	Amenity Rate	SEH Rate
AUSTIN	1.000000	0.677083	0.397198
NEAR NORTH SIDE	0.856631	1.000000	0.723216
LOOP	0.709677	1.000000	0.732392
NEAR WEST SIDE	0.612903	1.000000	0.632709
NORTH LAWNDALE	0.612903	0.312500	0.369960
...	...	...	...
MONTCLARE	0.023297	0.416667	0.726861
MOUNT GREENWOOD	0.021505	0.531250	0.718946
FOREST GLEN	0.016129	0.531250	0.767015
BURNSIDE	0.008961	0.000000	0.377331
EDISON PARK	0.000000	0.385417	0.711122

77 rows × 3 columns

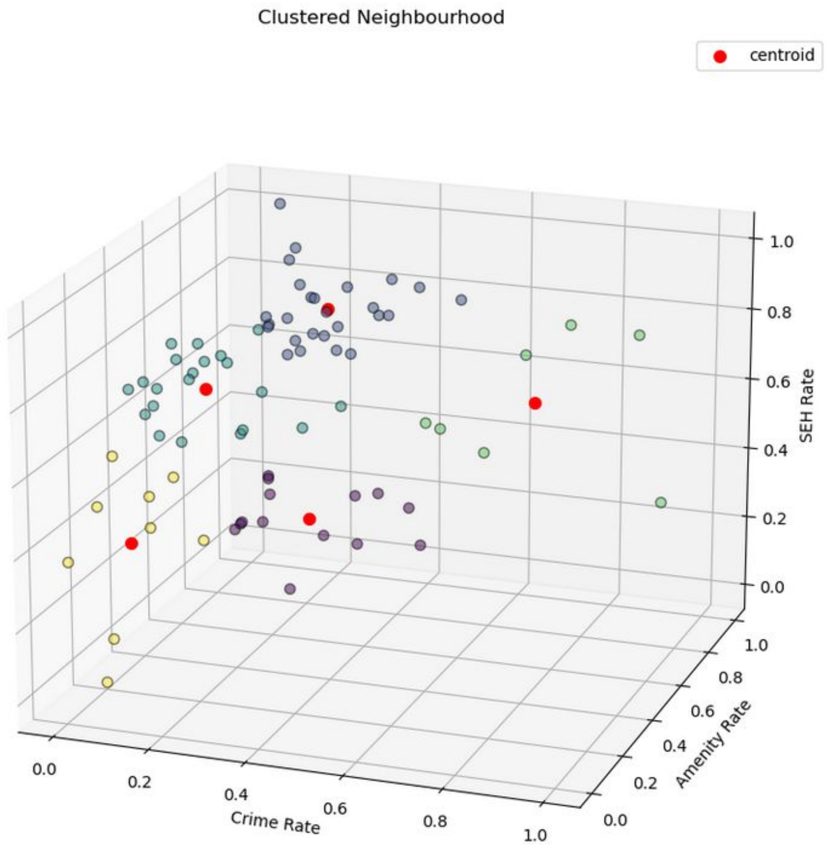
A 3D data visualization before K means clustering is provided here:



With this data, the elbow method is used to select the best k possible:



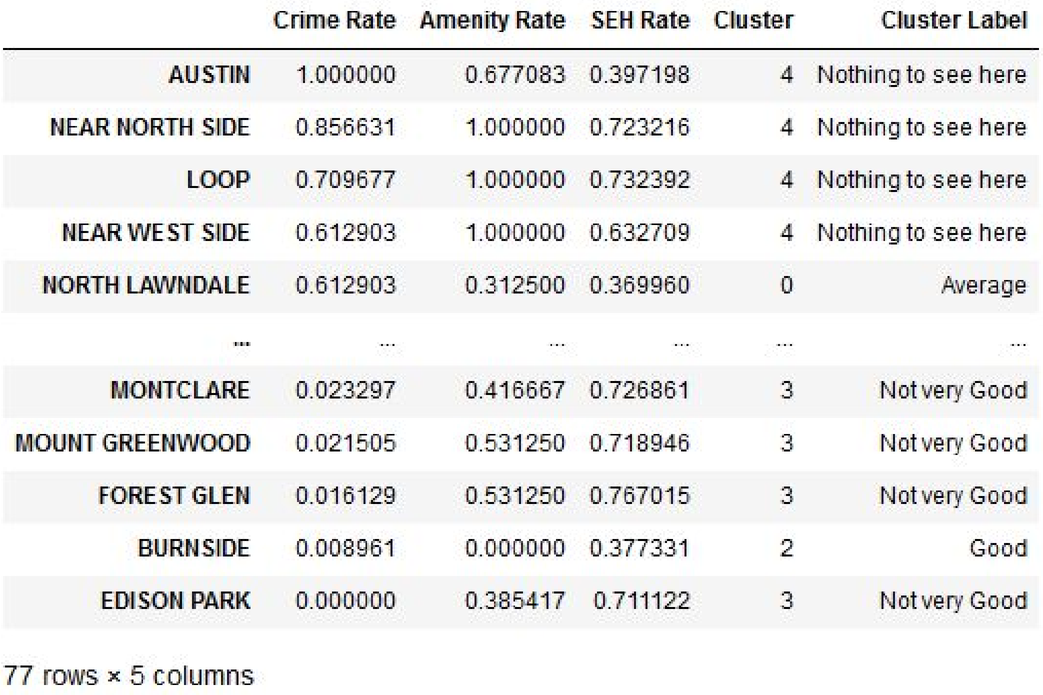
Here the best K could be 2 or 5. As 2 is far from being detailed enough, 5 is selected to run the clustering operation. A 3D plot of the clustered points with the centroid is displayed below:



For a better understanding on the choropleth map, each of the clusters has been given a name:

- 0 is Average
- 1 is Very Good
- 2 is Good
- 3 is Not very Good
- 4 is Nothing to see here

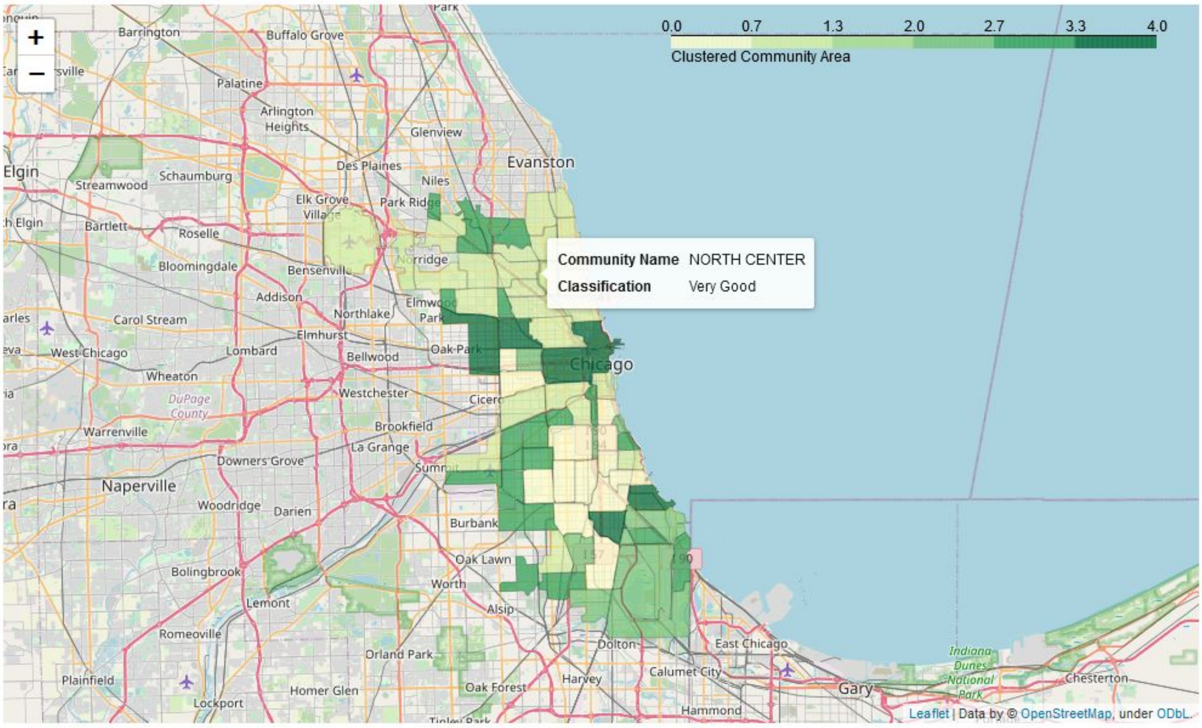
Which results in the following dataframe:



	Crime Rate	Amenity Rate	SEH Rate	Cluster	Cluster Label
AUSTIN	1.000000	0.677083	0.397198	4	Nothing to see here
NEAR NORTH SIDE	0.856631	1.000000	0.723216	4	Nothing to see here
LOOP	0.709677	1.000000	0.732392	4	Nothing to see here
NEAR WEST SIDE	0.612903	1.000000	0.632709	4	Nothing to see here
NORTH LAWNDALE	0.612903	0.312500	0.369960	0	Average
...	...	...	...	...	...
MONTCLARE	0.023297	0.416667	0.726861	3	Not very Good
MOUNT GREENWOOD	0.021505	0.531250	0.718946	3	Not very Good
FOREST GLEN	0.016129	0.531250	0.767015	3	Not very Good
BURNSIDE	0.008961	0.000000	0.377331	2	Good
EDISON PARK	0.000000	0.385417	0.711122	3	Not very Good

77 rows × 5 columns

Now, plotting a choropleth map with these data, it is possible to display the different clusters on a map, with hovering information.



Community Name NORTH CENTER
Classification Very Good

The best neighbourhoods possible are now depicted in this choropleth map.

CONCLUSION

For the XYZ company, a data analysis of different neighbourhoods of Chicago has been made. This analysis was splitted in three main parts. The first one was the data acquisition and preprocessing. It has been made regarding the company criteria (Crime, SEH and Amenity rate) and demand. Once the data have been preprocessed, The data understanding has been improved by using data visualization, to show the different data in a comprehensive way. Then, to end with the data have been processed with K-means clustering algorithm to cluster the community areas. Once done, XYZ has an interactive data map available, allowing them to select the neighbourhood they must invest in, and the ones they should stay away from.