PYTHON DATA SCIENCE

CAPSTONE PROJECT REPORT

Battle of the Neighbourhoods – Finding the right neighbourhood in Chicago

Author: Dayli Steinhoff

22th April 2020

Chapter Overview

1. Introduction: From problem to solution

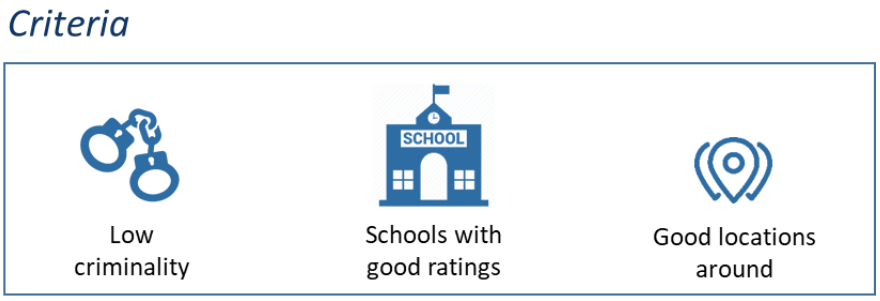
1.1. Starting point and target group

Chicago is considered one of the most popular cities in the US and with around 2.7 mill citizens the third-most-populous city in the United States. For new families moving into this city it is hard to find the proper place to live because of the large number of neighbourhoods and also because of lack of (customized) information for the target group.

This capstone project explores Chicago ant its community areas and the result is a preselection of neighbourhoods as well as valuable insights for users who are not familiar with the city. This project targets to support „young families" willing to move to Chicago and are struggling to find a suitable neighbourhood.

The target audience of this case study are young families looking for a safe environment to grow up their children. The target group is looking for a place close to good public schools as well as amenities (e.g. sport, restaurants, museums etc.)

The project is divided into 3 analysis according to the 3 criteria points, which are assumed to be the most relevant criteria when looking for a place to live.



1.2. The goal

The goal of this project is to generate a pre-selection of potential neighbourhoods based on the criteria mentioned above. Disadvantages (like criminality) as well as advantages (top public schools with ratings and amenities) will be then shown to enable the stakeholders of this project to make the best choice.

Detailed Results:

- On a first step the target audience will receive full transparency about the crime situation in Chicago including not only trends, type of crimes but also hot-spots with high criminality.

- As parents pursue to offer their children not only a save environment but also a good education the second section consists on a ranking of the top schools based on given metrics. Hereby schools with similar performance metrics will be classified into groups in order to enable choosing the right type of public schools. Combined with the information about criminality (top 10 filtered list), the user will then have the chance to start picking the areas of interest.

- After the second level filter (exclude high criminality and areas with top public schools), the user will have the chance to explore the remaining neighbourhoods by amenities e.g. restaurants, fitness,

entertainment etc. around the neighbourhoods. Hereby the user will obtain clusters of neighbourhoods which have similar venues.

2. Data Sources used for the analysis

2.1. Data Acquisition – General Information

The data acquired for this project is a combination of three different sources aimed to answer questions related to the 3 criteria mentioned in the previous chapter:

- Crime Situation in Chicago extracted from Chicago Data Portal
- Public Schools in Chicago Report extracted from Chicago Data Portal
- Venues information based on Foursquare

2.2. Crime Situation in Chicago – Data acquisition and data preparation

The first data source used for the project is "Crimes_-_2001_to_present.csv" available at https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

Due to the large size of the original file (over 7 million rows) only a part of the original document can be downloaded from the official website for free. The reduced dataset (more than 350k rows and 22 columns) is however sufficient for plotting trends. Additionally, I decided to download the full set of information available for the year 2020 for describing the current situation and plotting hot-spots of criminality. With 56.522 delicts over the first 4 months in the year 2020 and 22 columns, the user can have up-to-date transparency about the most common type of delicts as well as the locations with the higher criminality occurrences.

The available columns are: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location.

For the analysis the following columns were used and accordingly formatted for analysis/visualization purposes (Source: https://data.cityofchicago.org).

- ID: Unique identifier for the record.
- BLOCK: The partially redacted address where the incident occurred, placing it on the same block as the actual address.
- PRIMARY TYPE: The primary description of the IUCR code. Classification of the type of crimes.
- COMMUNITY AREA: Indicates the community area where the incident occurred. Chicago has 77 community areas.
- LOCATION DESCRIPTION: Description of the location where the incident occurred.
- YEAR: Year the incident occurred
- LATITUDE: The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
- LONGITUDE: The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block
- LOCATION: The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

For the crime data, the information for 2020 (Crimes_-_2020.csv) was appended using the append function to the data frame generated from the csv file: "Crimes_-_2001_to_present.csv". The document was enhanced by the "Name of the community area" and "ZIP Code" from the second data

source (Chicago public schools) using the merge function on “Community Area” (number) as a common identifier. I have ensured to remove duplicates.

The type of crimes were pivoted with the years to provide not only a yearly development of the crimes, but also to identify the major drivers (e.g. which type of crimes are having an influence on the overall trend).

For the map visualization, the data frame was adjusted to allow a faster calculation which might not be possible with the large existing number of combinations for latitudes and longitudes. Hereby I have used the location column and grouped the occurrences using the count function and later enhancing all relevant information (Latitude, Longitude etc.).

2.3. Crime Public Schools – Data acquisition and data preparation

The dataset is a csv file available at *https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t* and consists of 79 columns and 566 rows, whereby each row contains the information for one school. Besides location information, e.g. **ZIP code, Community area and coordinates**, also different metrics are included in the dataset. For this project I made use of the **following metrics**:

- SAFETY SCORE: Student Perception/Safety score from 5 Essentials survey
- PARENT ENVIRONMENT SCORE: Parent Perception/Environment score from parent survey. These scores range from 30 to 70.
- ENVIRONMENT SCORE: Supportive Environment score from 5 Essentials survey
- INSTRUCTION SCORE: Ambitious Instruction score from 5 Essentials survey
- PARENT ENGAGEMENT SCORE: Parent Perception/Engagement score from parent survey
- AVERAGE STUDENT ATTENDANCE: Average daily student attendance
- RATE OF MISCONDUCTS: # of misconducts per 100 students

There were more metrics, but I have decided to reduce the scope into the most relevant ones. **Other relevant metrics e.g. graduation, CPS Performance and college enrolment rate etc. are not available for the majority of the schools and hence I excluded them from the analysis to avoid biased conclusions.**

Since during this project, I will use the k-means algorithm with the public school data, there are some adjustments required. The k-means method however isn't directly applicable to categorical variables because Euclidean distance function is not meaningful for discrete variables. Hence the datatype needed to be transformed into floating so that I can start processing the data. 513 schools were derived out of 566 rows having a full set of minimum information for the analysis.

2.4. Foursquare – Venue information in Chicago - Data acquisition and data preparation

First I took the neighbourhood’s names, latitudes and longitudes values from the previously created zip file (slicing the available information from Chicago public schools data frame). These coordinates and neighbourhood’s names are the basis for the format function within the Foursquare API. The URL uses the explore (venue) functionality to obtain the venues in the neighbourhood. I limited to 100 venues in a ratio of 500m. I sent a Get request and obtained a Json file and transformed it into a pandas data frame. The information that I will use from Foursquare is the 10 most common venues in a neighbourhood. This will be assigned to the results from the previous sections so that at the end the user will have some preselected neighbourhoods clustered into the typical venues.

One of the most important data processing steps in this part is one hot encoding. One hot encoding is a process by which categorical variables are converted into a form that enables Machine Learning algorithms to work since they cannot operate on label data directly. All input variables and output

variables need to be numeric. In this project, hot encoding was used to transform the venues categories before running the k-mean algorithm.

3. Methodology – General description

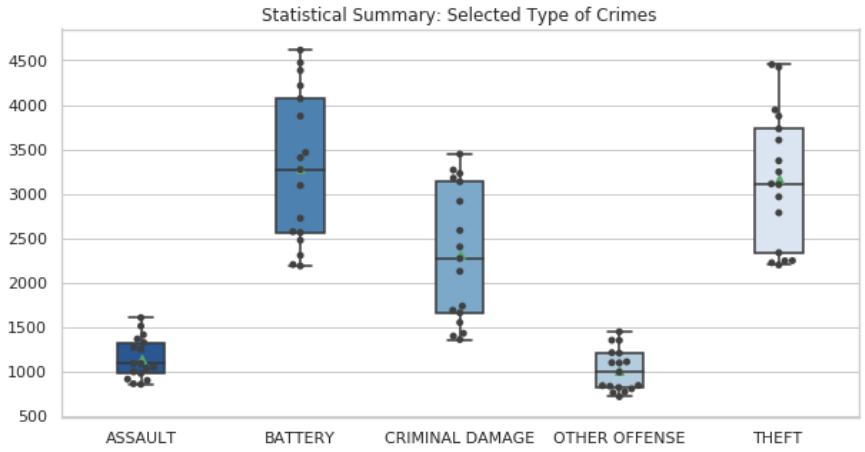
In this section I will introduce the analysis I conducted on my data briefly. For more details about how the method was used and the corresponding results, please go to chapter 4Results.

3.1. Explanatory Data Analysis:

For the first part of the project I have used explanatory data analysis to provide transparency about the current crime situation in Chicago and identify places with high(er) criminality and.

For this purposes, the data was reshaped, enhanced and pivoted in additional tables. In addition, statistical summaries were created for example to describe the average occurrence of the delicts per year based on a 17-years-period from 2002 to 2019 using the "describe()" function. This function returns the mean, standard deviation, minimum and maximum value as well as the quartiles.

For visualization I have shortened the view by the top 5 identified type of crimes and used a boxplot. A boxplot summarizes the distribution of numerical variable (in this case the number of delict per year) for one or several groups (in this case type of crimes). This shows the underlying distribution and the number of points of each group. With 3286 incidents per year, battery shows the highest number of delicts average per year and reached a maximum of 4616 in 2003.



1Statistical Summary of the type of crimes frequency per year

3.2. Visualization:

I made strong use of visualization not only to provide the final results to the target audience, but also to get some insight about the data. The most common visualization were line charts for trends and enhanced by pivot charts (to compare trends in different categories simultaneously), different type of rankings and folium maps making use of the cluster functionality. For more details, please go to the results chapter (4Results).

3.3. Modelling:

Within the scope of this project, k-Means clustering was used for:

- Clustering schools based on evaluation criteria (performance metrics e.g. scores and ratios)
- Clustering neighbourhoods with similar venues so that the families can have at glance a short list of the areas based on the amenities located around.

k-means is an unsupervised machine learning algorithm that cluster the data based on a predefined cluster size. The reason I used this algorithm for this project because it perfectly serves the purpose of my project.

Clustering schools based on evaluation criteria:

For the clustering of the public schools in Chicago I chose the following evaluation criteria: Safety Score, Environment Score, Instruction Score, Parent Engagement Score, Parent Environment Score, Average Student Attendance and Rate of Misconducts (per 100 students). For this project, I have used 7 clusters for splitting 513 schools.

For the k-Means clustering the dataset was normalized first. Normalization is a statistical method that helps mathematical-based algorithms interpret features with different magnitudes and distributions equally. I used the StandardScaler() to normalize the selected features in the dataset and KMeans () from sklearn.cluster to derive the following schools clusters:

For interpretation and use of the table above please go to the Results Section.

Clustering neighbourhoods with similar venues: The 3rd and last focus on finding similar neighbourhoods according to amenities. For this purpose I used the Foursquare API to explore neighbourhoods in Chicago and requested the venue categories. These categories are then used as the features for grouping the neighbourhoods into clusters. In summary here some of the most important steps before applying the k-means clustering:

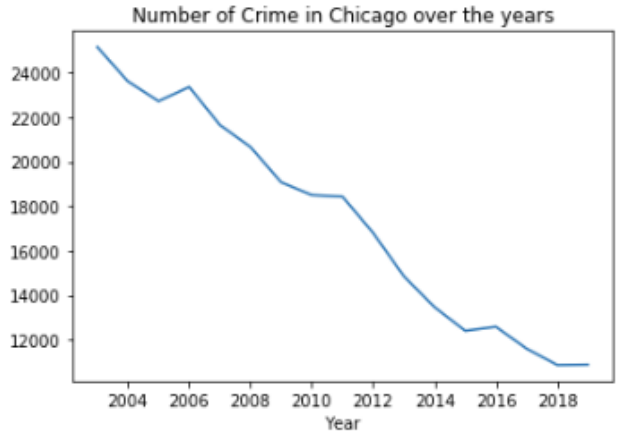
- The rows are grouped in neighbourhoods and the mean of the frequency for each category is calculated
- Venues are sorted from most common in a descending order
- Data frame is created with the most common venues as columns. This data frame is the one used for fitting the K-mean function

For this project, I split 77 the neighbourhoods into 5 clusters. **For the final result the data frame however was further pre-filtered in order to exclude the neighbourhoods with high criminality and select only the areas with top-rated schools and exclude the other observations.**

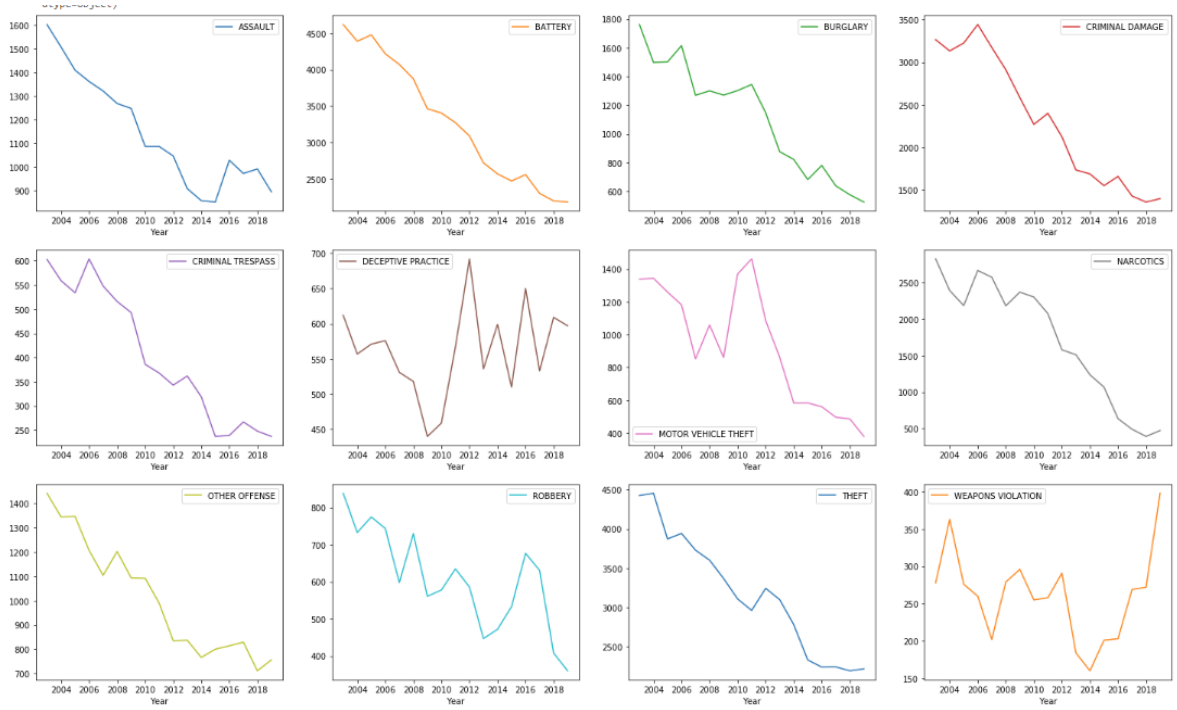
4. Results

4.1. Chicago Crime – Analysis and results

For the trend analysis, the time period selected was from 2002 onwards since the number of observations for 2001 is low due to data gaps. From the trend below, we can see that the frequency of crimes in Chicago has been continuously decreasing.

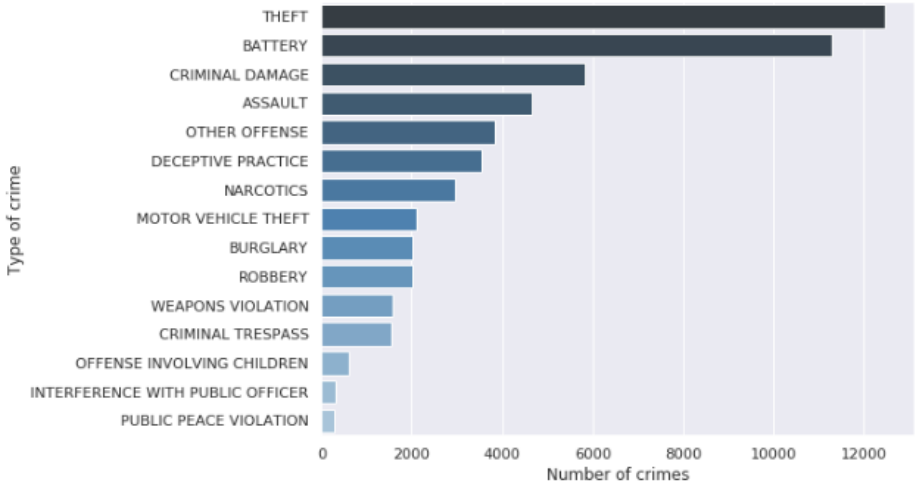


When taking a deeper look at the different types of crime, we can observe that this positive declining trend is benefiting from especially lower number of delicts related to theft, narcotics, battery and criminal damage among others:



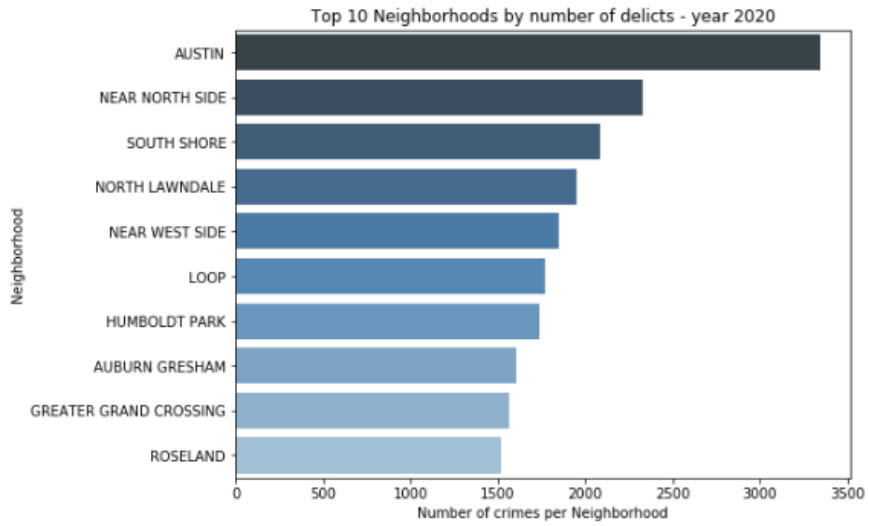
After an increase in the year 2016, improved technology, more involvement from federal authorities in gun crimes and efforts to rebuild community led to a trend improvement. *(See: Chicago Tribune//https://www.chicagotribune.com/news/breaking/ct-met-chicago-crime-stats-end-of-year-20181228-story.html)*

Despite the positive improvements, theft, battery and criminal damage are still ranking as the most common crimes for the year 2020 (01/2020 – mid of 04/2020).



Comparing the 10 neighbourhoods with the highest crime rates it is visible that Austin seems to be the less safe area in Chicago. **This will be used as the first level filter when selecting the right**

neighbourhood (Out of 77 neighbourhoods, 10 of them will be excluded from the final recommendation).



2. Top 10 neighbourhoods by number of delicts in Chicago

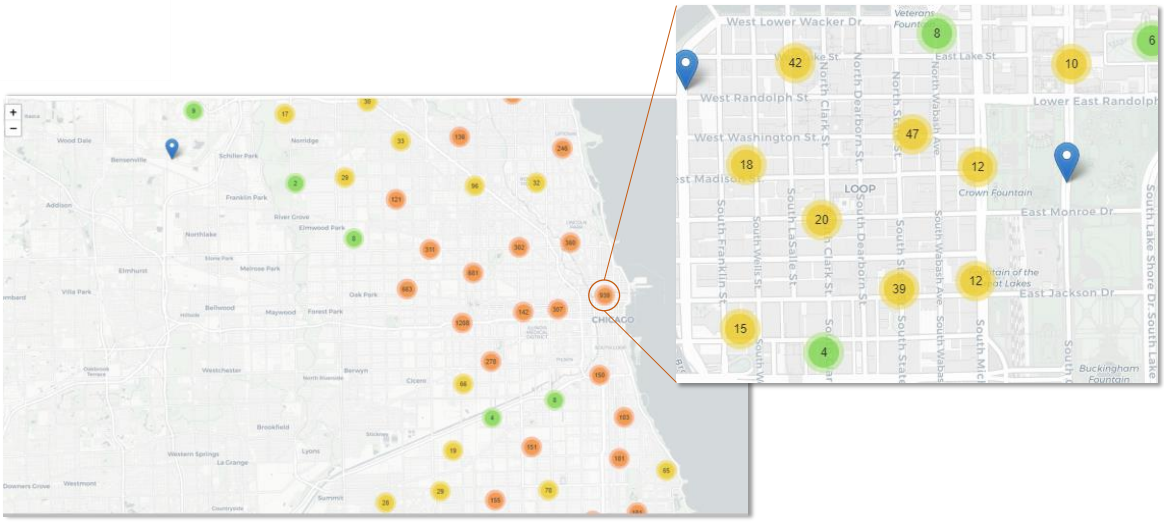
In case that the user would like to take a risk and move to a considered risky area, I have also provided more granular information on block level. By taking a closer look at the coordinates we can identified that some blocks within Loop are ranked with a higher number of cases for the year 2020 (from January to April). This view is however limited to a shorter number of observation.

Location	ID_x	Block	Beat	District	Ward_x	Community Area	Year	Latitude	Longitude	ZIP Code	Community Area Name
(41.883500187, -87.627876698)	112	001XX N STATE ST	111	1	42.0	32.0	2020	41.883500	-87.627877	60605	LOOP
(41.754592961, -87.741528537)	67	076XX S CICERO AVE	833	8	18.0	65.0	2020	41.754593	-87.741529	60629	WEST LAWN
(41.976290414, -87.905227221)	63	100XX W OHARE ST	1651	16	41.0	76.0	2020	41.976290	-87.905227	60656	OHARE
(41.742710224, -87.634088181)	53	083XX S STEWART AVE	622	6	21.0	44.0	2020	41.742710	-87.634088	60619	CHATHAM
(41.868541914, -87.639235361)	51	011XX S CANAL ST	124	1	25.0	28.0	2020	41.868542	-87.639235	60608	NEAR WEST SIDE
(41.88171846, -87.627760426)	50	0000X S STATE ST	112	1	42.0	32.0	2020	41.881718	-87.627760	60605	LOOP
(41.884650262, -87.627915459)	45	001XX N STATE ST	111	1	42.0	32.0	2020	41.884650	-87.627915	60605	LOOP
(41.963070794, -87.655984213)	41	044XX N BROADWAY	1913	19	46.0	3.0	2020	41.963071	-87.655984	60640	UPTOWN
(41.897895128, -87.624096605)	40	008XX N MICHIGAN AVE	1833	18	2.0	8.0	2020	41.897895	-87.624097	60610	NEAR NORTH SIDE
(41.750940757, -87.625185222)	39	0000X W 79TH ST	623	6	6.0	44.0	2020	41.750941	-87.625185	60620	CHATHAM
(41.891694878, -87.626155832)	37	0000X E GRAND AVE	1834	18	42.0	8.0	2020	41.891695	-87.626156	60610	NEAR NORTH SIDE
(41.883475491, -87.627876969)	37	001XX N STATE ST	111	1	42.0	32.0	2020	41.883475	-87.627877	60605	LOOP

3. List of location by blocks with high frequency of crimes in less than 4 months (2020).

In order to have a more comprehensive picture I have created a folium map making use of the cluster functionality with 10000 coordinates sorted by the number of cases. The users can use the map and zoom into the zones until finding the affected blocks and hence make a decision about the potential places to move since the transparency is provided on a very granular level.

Hereby the “hot spots” in Chicago for crimes:



4. Hot spots of criminality in Chicago – Information from January to April 2020

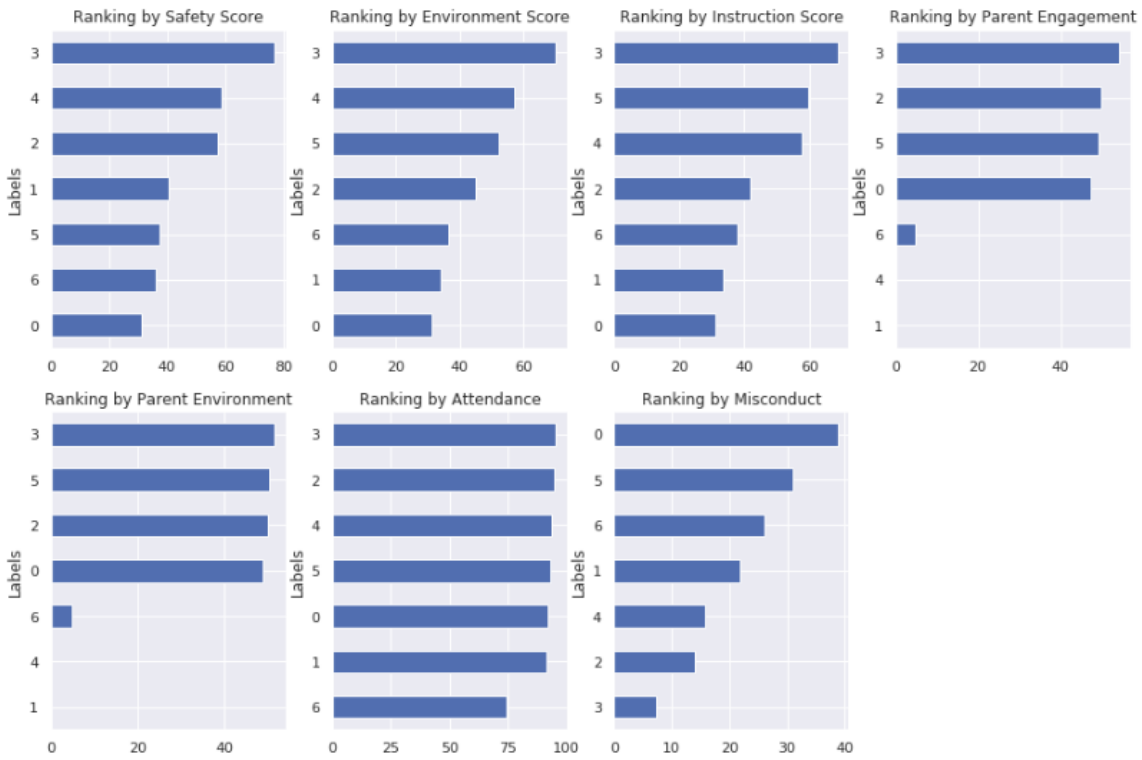
4.2. Chicago Public Schools – A closer look

According to the information provided I have filtered to 513 public schools I have created an overview of the different scores and some values like means and distribution.

	Safety Score	Environment Score	Instruction Score	Parent Engagement Score	Parent Environment Score	Average Student Attendance	Rate of Misconducts (per 100 students)
count	513.000000	513.000000	513.000000	513.000000	513.000000	513.000000	513.000000
mean	49.504873	47.766082	48.288499	38.136452	38.109162	92.512281	22.050682
std	20.110837	16.215584	17.417176	21.738163	21.685333	5.561582	28.264920
min	1.000000	1.000000	1.000000	0.000000	0.000000	57.900000	0.000000
25%	35.000000	37.000000	37.000000	41.000000	41.000000	91.900000	5.300000
50%	48.000000	47.000000	47.000000	48.000000	48.000000	94.400000	12.800000
75%	61.000000	58.000000	59.000000	52.000000	52.000000	95.500000	28.100000
max	99.000000	99.000000	99.000000	69.000000	70.000000	98.400000	251.600000

5. Average ratings by metric of Chicago Public Schools

After running the K-means clustering method, it is possible to access each cluster. Before going into details please find the following ranking of the schools by score, where it is visible that the best school cluster is the number 3 which I will explore closerly.



6. Top Cluster of public Schools in Chicago by rated metrics

The user can now take a deeper look into the different clusters. In order to have all details for the user I have added an (automatized) cell with written information comparing the average of the cluster against the overall average.

Cluster 3

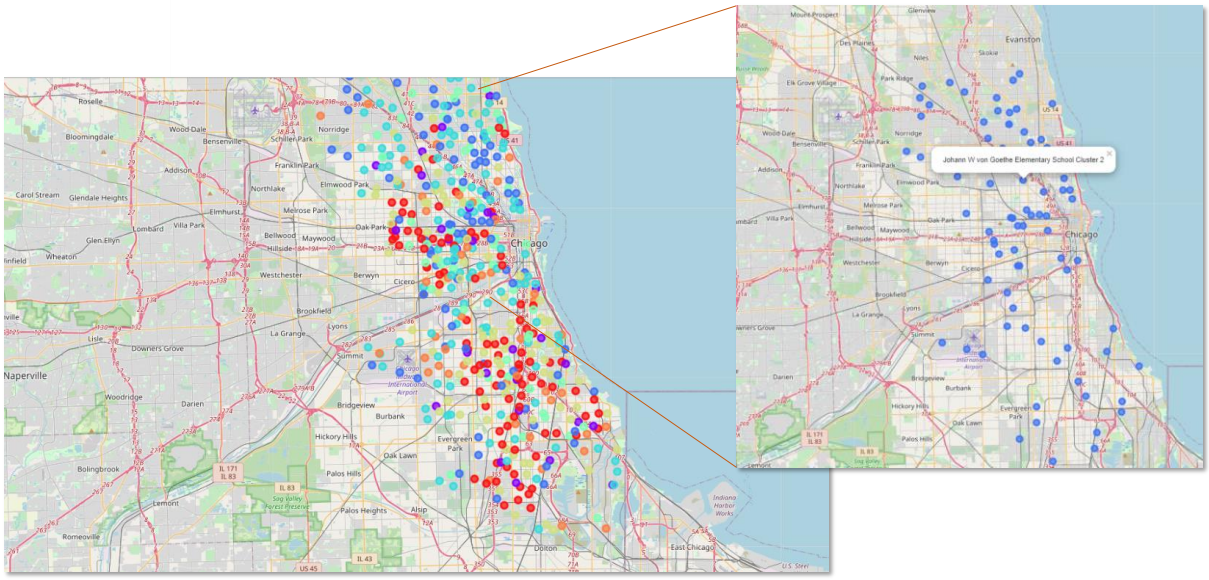
```
df3.iloc[3,:]
```

```
Safety Classification                    better than average
Environment Classification               better than average
Instruction Classification               better than average
Parent Engagement Classification         better than average
Parent Environmental Classification      better than average
Average Student Attendance Classification  better than average
Rate of Misconducts Classification       better than average
Name: 3, dtype: object
```

```
public_schools_df.loc[public_schools_df['Labels'] == 3, public_schools_df.columns[[1] + list(range(4, public_schools_df.shape[1]))]]
```

	Name of School	Safety Score	Environment Score	Instruction Score	Parent Engagement Score	Parent Environment Score	Average Student Attendance	Rate of Misconducts (per 100 students)	Latitude	Longitude	Community Area	Neighborhood
6	Lenart Elementary Regional Gifted Center	79.0	51.0	67.0	52.0	52.0	97.4	0.3	41.747150	-87.628002	44	CHATHAM
11	Abraham Lincoln Elementary School	99.0	74.0	66.0	56.0	47.0	96.0	2.0	41.924497	-87.644522	7	LINCOLN PARK
12	William Penn Elementary School	78.0	99.0	99.0	52.0	53.0	91.8	5.7	41.858370	-87.721336	29	NORTH LAWNDALE
17	Northside College Preparatory High School	99.0	99.0	88.0	57.0	62.0	95.7	2.8	41.981352	-87.708672	13	NORTH PARK
24	Albany Park Multicultural Academy	66.0	66.0	71.0	46.0	51.0	97.0	2.3	41.971143	-87.709627	14	ALBANY PARK

When displaying the clusters in a folium map the user can find the places close to schools having the best rankings. After filtering with the cluster 3 (best cluster) it is more easy to focus on the locations with the best ranking schools.



Within the cluster 3 (top schools), there are 85 top public schools which are located in 45 different neighbourhoods whereby we can find already 5 top rated schools in Lake view, West Town etc. Since some of the locations below show a high number of delicts, they need to be filtered out.

Neighborhood	
LAKE VIEW	5
WEST TOWN	5
NORTH CENTER	4
ALBANY PARK	3
WEST RIDGE	3

Hereby the new cluster with top schools excluding the areas with high criminality. Now the user have a choice over 73 public schools in 39 neighborhoods.

	Name of School	Safety Score	Environment Score	Instruction Score	Parent Engagement Score	Parent Environment Score	Average Student Attendance	Rate of Misconducts (per 100 students)	Latitude	Longitude	Community Area	Neighborhood
0	Lenart Elementary Regional Gifted Center	79.0	51.0	67.0	52.0	52.0	97.4	0.3	41.747150	-87.628002	44.0	CHATHAM
1	James E McDade Elementary Classical School	99.0	57.0	52.0	61.0	52.0	96.2	0.0	41.734514	-87.619177	44.0	CHATHAM
2	Abraham Lincoln Elementary School	99.0	74.0	66.0	56.0	47.0	96.0	2.0	41.924497	-87.644522	7.0	LINCOLN PARK
3	Walter L Newberry Math & Science Academy Eleme...	71.0	52.0	66.0	54.0	47.0	95.1	2.1	41.913974	-87.646015	7.0	LINCOLN PARK
4	LaSalle Elementary Language Academy	99.0	62.0	52.0	53.0	48.0	96.8	7.0	41.913882	-87.637601	7.0	LINCOLN PARK

7. List of selected top public schools in Chicago located in areas with low criminality

After the second analysis we have 39 neighborhoods instead of the original 77 to select for example Lake View, West Town, Lincoln Park etc. In the next section, I will cluster this neighborhoods into 5 groups of locations according to the ammenities around.

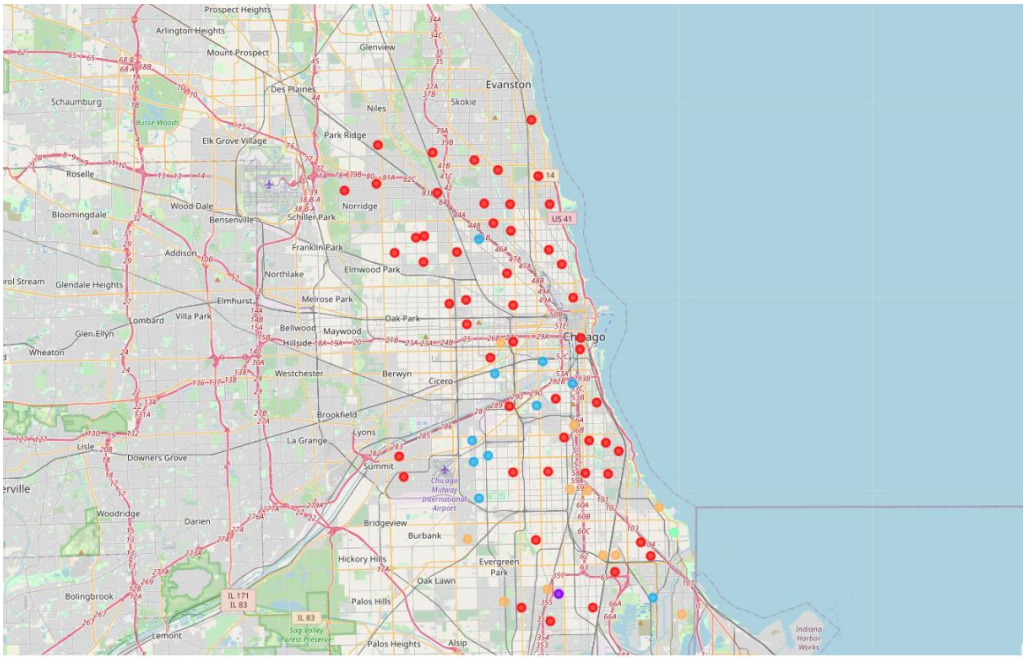
4.3. Final step: Clustering the pre-filtered neighbourhoods according to similar locations

Based on the scrape zip file (containing neighbourhoods and coordinates) I retrieved a json file from Foursquare showing 1044 venues with 200 unique categories to start with the analysis. After data

processing and running the k-mean algorithm (with 5 clusters) I labelled the neighbourhoods with clusters labels:

ZIP Code	Neighborhood	Latitude	Longitude	Location	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
60625	ALBANY PARK	41.964855	-87.714831	(41.96485522, -87.71483051)	0.0	Mexican Restaurant	Sandwich Place	Bakery	Coffee Shop	Fast Food Restaurant
60632	ARCHER HEIGHTS	41.805523	-87.725819	(41.80552325, -87.72581893)	2.0	Mexican Restaurant	Bakery	Nightclub	Pharmacy	Seafood Restaurant
60616	ARMOUR SQUARE	41.843952	-87.635318	(41.84395162, -87.63531815)	2.0	Chinese Restaurant	Cosmetics Shop	Mobile Phone Shop	Park	Mexican Restaurant
60652	ASHBURN	41.738864	-87.729919	(41.73886352, -87.72991927)	4.0	Park	Nightclub	Event Service	Yoga Studio	Flower Shop
60620	AUBURN GRESHAM	41.738426	-87.668136	(41.7384262, -87.66813597)	0.0	Fast Food Restaurant	Snack Place	Bar	Pizza Place	Park

The information was merged and displayed in the following type of clusters.



8. Clusters of neighborhoods based on similar venues

The 5 clusters were furter filtered excluding neighborhoods with high criminality and including only areas with top rated schools:

Cluster 0: neighbourhoods with restaurants, coffee shops and stores. Fast food like hot dogs and donuts are very common in this area. Before filter: 53 neighbourhoods, after filter 28 potential neighbourhoods (e.g. **Lake View, West Town, North Center** and **Albany Park**).

.

	ZIP Code	Neighborhood	Latitude	Longitude	Location	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Name of School
0	60625	ALBANY PARK	41.964855	-87.714831	(41.96485522, -87.71483051)	0	Mexican Restaurant	Sandwich Place	Bakery	Coffee Shop	Fast Food Restaurant	Liquor Store	Mobile Phone Shop	Donut Shop	Fried Chicken Joint	Chinese Restaurant	Albany Park Multicultural Academy,John Palmer ...
2	60639	BELMONT CRAGIN	41.925596	-87.769843	(41.92559567, -87.76984337)	0	Chinese Restaurant	Sandwich Place	Diner	Fast Food Restaurant	Currency Exchange	Supermarket	Donut Shop	Restaurant	Automotive Shop	Pharmacy	Belmont-Cragin Elementary School
4	60608	BRIDGEPORT	41.833882	-87.650619	(41.8338821, -87.65061898)	0	Bar	Mexican Restaurant	Chinese Restaurant	Pizza Place	Art Gallery	Grocery Store	Coffee Shop	Burger Joint	Shipping Store	Sandwich Place	Mark Sheridan Elementary Math & Science Academy
5	60632	BRIGHTON PARK	41.828863	-87.692493	(41.8288631, -87.69249318)	0	Soccer Field	Hot Dog Joint	Convenience Store	Gym	Grocery Store	Basketball Court	Dry Cleaner	Flower Shop	Filipino Restaurant	Fast Food Restaurant	Columbia Explorers Elementary Academy,Calmeca ...
6	60617	CALUMET HEIGHTS	41.727821	-87.564393	(41.72782129, -87.56439326)	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Robert A Black Magnet Elementary School

Cluster 1: After removing the locations with criminality and excluding the areas not included in the top ten, this cluster remains empty. Before the filters, this cluster had a basketball court which might be interesting for basketball fans.

	ZIP Code	Neighborhood	Latitude	Longitude	Location	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
199	60643	WASHINGTON HEIGHTS	41.701967	-87.64769	(41.70196704, -87.64768959)	1	Basketball Court	Yoga Studio	Eastern European Restaurant	Food	Flower Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Event Space	Event Service

Cluster 2: mainly typical for the Mexican and Latin American restaurants but also parks and banks. Before filter: 10 neighbourhoods, after filter 6 potential neighbourhoods (e.g. **South Lawndale** and **Gage Park**).

	ZIP Code	Neighborhood	Latitude	Longitude	Location	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Name of School
1	60618	AVONDALE	41.941122	-87.719531	(41.94112155, -87.71953058)	2	Mexican Restaurant	Bar	Bus Station	Boat or Ferry	Karaoke Bar	Concert Hall	Donut Shop	Convenience Store	Tattoo Parlor	Bank	Carl von Linne Elementary School
15	60632	GAGE PARK	41.795181	-87.711094	(41.79518118, -87.71109447)	2	Mexican Restaurant	Café	Elementary School	Grocery Store	Park	Dry Cleaner	Flower Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Eric Solorio Academy High School,Talman Elemen...
23	60608	LOWER WEST SIDE	41.858876	-87.662225	(41.85887606, -87.6622254)	2	Mexican Restaurant	Bar	Breakfast Spot	Latin American Restaurant	Thrift / Vintage Store	Rock Club	Cocktail Bar	Coffee Shop	Taco Place	Pizza Place	John A Walsh Elementary School
30	60617	SOUTH DEERING	41.699740	-87.562121	(41.69973969, -87.56212104)	2	Mexican Restaurant	Child Care Service	Bakery	Deli / Bodega	Eastern European Restaurant	Food	Flower Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Burnham Elementary Inclusive Academy
31	60623	SOUTH LAWNDALE	41.850776	-87.705075	(41.85077575, -87.70507547)	2	Mexican Restaurant	Park	Seafood Restaurant	Bank	Chinese Restaurant	Donut Shop	Diner	Discount Store	Filipino Restaurant	Fast Food Restaurant	John Spry Elementary Community School,Spry Com...
34	60629	WEST LAWN	41.766758	-87.719508	(41.7667583, -87.71950804)	2	Mexican Restaurant	Fast Food Restaurant	Ice Cream Shop	Chinese Restaurant	Automotive Shop	Flower Shop	Miscellaneous Shop	Seafood Restaurant	Bowling Alley	Donut Shop	Mariano Azuela Elementary School

Cluster 3: After removing the locations with criminality and excluding the areas not included in the top ten, this cluster remains empty. Before the filters, this cluster show Yoga studios and event locations in South Chicago.

ZIP Code	Neighborhood	Latitude	Longitude	Location	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
60617	SOUTH CHICAGO	41.74383	-87.542891	(41.7438302, -87.542891)	3	Food	Yoga Studio	Dry Cleaner	Flower Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Event Space	Event Service	Ethiopian Restaurant

Cluster 4: peculiarity of farmers markets but also Yoga studios. Before and after filter 5 potential neighbourhoods (e.g. **Chattam and Beverly**).

	ZIP Code	Neighborhood	Latitude	Longitude	Location	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Name of School
3	60643	BEVERLY	41.705140	-87.658116	(41.70514024, -87.65811642)	4	American Restaurant	Cosmetics Shop	Park	Caribbean Restaurant	Salon / Barbershop	Yoga Studio	Dry Cleaner	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Kate S Kellogg Elementary School,Elizabeth H S...
7	60619	CHATHAM	41.728508	-87.607471	(41.72850759, -87.60747108)	4	Fast Food Restaurant	Train Station	Yoga Studio	Donut Shop	Flower Shop	Filipino Restaurant	Farmers Market	Event Space	Event Service	Ethiopian Restaurant	Lenart Elementary Regional Gifted Center,James...
10	60612	EAST GARFIELD PARK	41.870912	-87.699887	(41.87091163, -87.69988652)	4	Park	Hot Dog Joint	American Restaurant	Train Station	Donut Shop	Flower Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Event Space	Willa Cather Elementary School,Mary Mapes Dodg...
13	60621	ENGLEWOOD	41.772754	-87.637295	(41.77275433, -87.63729464)	4	Business Service	Yoga Studio	Eastern European Restaurant	Food	Flower Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Event Space	Event Service	Joshua D Kershaw Elementary School
24	60655	MOUNT GREENWOOD	41.697198	-87.697264	(41.69719792, -87.6972638)	4	Park	Home Service	Yoga Studio	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Event Space	Event Service	Ethiopian Restaurant	Elementary School	Annie Keller Elementary Gifted Magnet School

5. Discussion

Despite of the declining number of delicts in Chicago in the last years, I will recommend young families to avoid the areas detected during the analysis like **Austin, Near North Side**, **South Shore** and **North Lawndale** (Please consult :Top 10 neighbourhoods by number of delicts in Chicago). In case that the users are still considering moving into one of this neighbourhoods, please go into a more granular level and look the location by blocks (See 4. Hot spots of criminality in Chicago – Information from January to April 2020).

After analysing Chicago Public schools, a list of top schools with all metrics showing performances higher than the average. Places like **Lake View, West Town, North Center** and **Albany Park** were having even more than one top-rated public school and seem very attractive for parents looking for

high quality public schools. These neighbourhoods were allocated to **Cluster 0** having a focus on restaurants, coffee shops and stores. Fast food like hot dogs and donuts is very common in this area.

All information is also available (unfiltered) in case the the user would like to take a riskier choice. Enough granularity was provided so that blocks with higher number of delicts can be avoided.

6. Conclusion

This project achieved its target to provide more transparency to young families about potential neighborhoods to move in. By excluding neighborhoods with higher criminality occurrence and selecting areas with public schools , the user had a preselection reducing the number of neighborhoods in Chicago from 77 to 39 potential places. In addition, based on a classification into clusters the target group can now have a closer selection based on her/his respective interests.

This project can be enhanced in the future including cost of living (rent) but also finding the neighborhood with the optimal distance to the future job location.