PYTHON DATA SCIENCE

CAPSTONE PROJECT REPORT

Battle of the Neighbourhoods – Finding the right neighbourhood in Chicago

(Week 4 only)

Author: Dayli Steinhoff

22th April 2020

1. Introduction: From problem to solution

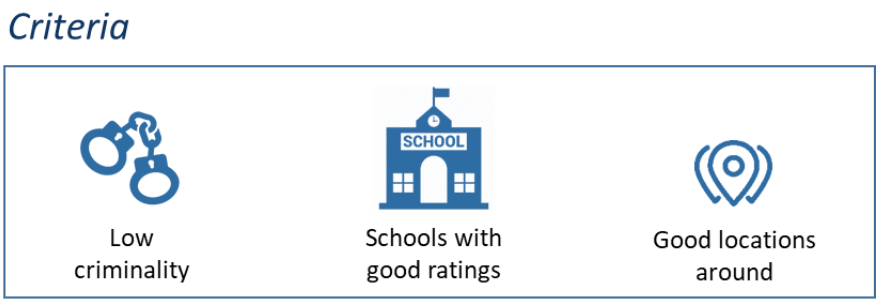
1.1. Starting point

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.

This capstone project explores Chicago ant its community areas and give 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 NOT TO assign the place where the families are going to live, but rather providing them with full transparency to make their own decision and setting some filters based on criminality (avoiding places with high number of delicts) and neighbourhoods with top-rated schools.

Potential results:

- On a first step the target audience will receive full transparency about the crime situation in Chicago (e.g. trend, type of crimes and 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 (See next subchapter chapter) using the merge function on "Community Area" (number) as a common identifier. I have ensured to remove the 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 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

Other metrics e.g. graduation 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.

The target of this section is to cluster the schools (partition) into groups of institutions that have similar characteristics using the k-means algorithm. This method however isn't directly applicable to categorical variables because Euclidean distance function is not meaningful for discrete variables. Hence unnecessary columns should be removed and the datatype needs to the 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.

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.