Applied Data Science Capstone

from IBM Coursera

Capstone Project – The Battle of Neighborhoods

Project Report

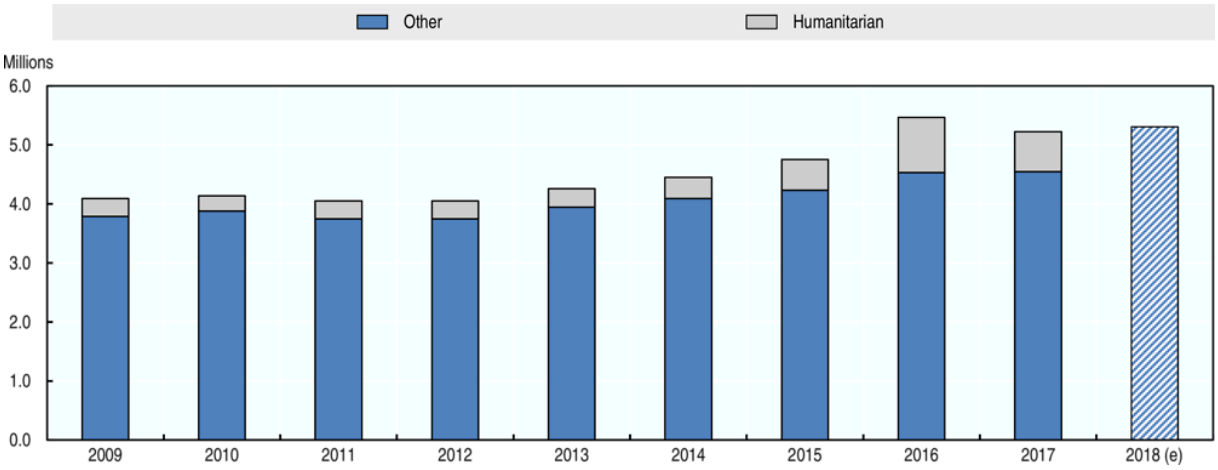
Tobias Kelle-Chong

April 2020

Index

Introduction

International migration is becoming a more and more relevant topic. According to a study be the OECD, the OECD member countries received about 5.3 million new permanent migrants in 2018, a 2% increase on 2017, according to preliminary data. Since 2015, European OECD countries have collectively received more permanent migrants than the United States. Nevertheless, the United States remains the largest single destination country for migrants, followed by Germany.



Permanent migration flows to OECD countries, 2009-18, *Source*: *OECD International Migration Database,* https://doi.org/10.1787/data-00342-en.

More than 4.9 million temporary labour migrants entered OECD countries in 2017, an 11% increase over 2016.

The problem this project aims to tackle is the search for a new neighborhood in case people have to move to a foreign city. To be more specific, this project serves to help people who have to move to a foreign city to find a new neighbourhood that is as similar as possible to the old neighbourhood they have to leave. On top of that, this project tries to identify the most suitable neighborhood for a given set of preferences.

Due to very different reasons a lot of people have to move homes. The decision process of where to buy a new house or rent an appartement is quite complex. A lot of features have to be taken into consideration, e.g. socioeconomic factors like unemployment or the income level of the neighborhood, but also crime rates, housing prices, reputation of public schools for the children, shops, malls, theatres, hospitals et cetera.

Data

Data Description:

1. Foursquare API

As requested by the assignment, this project will heavily use Four-square API as its prime data gathering source. With the API I will perform location search, location sharing and details about a business. Due to limitations of the API requests possible, the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

To find similarities between neighborhoods in New York and Chicago, we need to gather data about different kind of venues to find the characteristics. Hence, I will use Foursquare data for this task.

To determine the similarities of both cities, I will segment and cluster the neighborhoods to find similar places. In order to do that, a k-means clustering algorithm will be utilized, basen on location data provided by the Foursquare API.

2. Selected Socioeconomic Indicators in Chicago

The city of Chicago released a dataset of socioeconomic data to the Chicago City Portal.

A detailed description of the dataset can be found on the city of Chicago's website, but to summarize, the dataset has the following variables:

- Community Area Number (ca): Used to uniquely identify each row of the dataset
- Community Area Name (community_area_name): The name of the region in the city of Chicago
- Percent of Housing Crowded (percent_of_housing_crowded): Percent of occupied housing units with more than one person per room
- Percent Households Below Poverty (percent_households_below_poverty): Percent of households living below the federal poverty line
- Percent Aged 16+ Unemployed (percent_aged_16_unemployed): Percent of persons over the age of 16 years that are unemployed
- Percent Aged 25+ without High School Diploma (percent_aged_25_without_high_school_diploma): Percent of persons over the age of 25 years without a high school education
- Percent Aged Under 18 or Over 64:Percent of population under 18 or over 64 years of age (percent_aged_under_18_or_over_64): (ie. dependents)
- Per Capita Income (per_capita_income_): Community Area per capita income is estimated as the sum of tract-level aggragate incomes divided by the total population
- Hardship Index (hardship_index): Score that incorporates each of the six selected socioeconomic indicators

This dataset contains a selection of six socioeconomic indicators of public health significance and a “hardship index,” for each Chicago community area, for the years 2008 – 2012. Scores on the hardship index can range from 1 to 100, with a higher index number representing a greater level of hardship. I will use the Hardship Index only as this score includes each of the indicators.

I acknowledge that the time series ends in the year of 2012, but for this assignment I will ignore this fact and assume that the data is up to date and still valid for this decision process.

3. Chicago Crime Data

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

This dataset is quite large - over 1.5GB in size with over 7 million rows. A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

4. Home values - Zillow Home Value Index (ZHVI)

The Zillow Home Value Index (ZHVI)is a smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type.

Zillow publishes top-tier ZHVI (USD, typical value for homes within the 65th to 95th percentile range for a given region) and bottom-tier ZHVI (USD, typical value for homes that fall within the 5th to 35th percentile range for a given region).

Zillow also publishes ZHVI for all single-family residences (USD, typical value for all single-family homes in a given region), for condo/coops (USD), for all homes with 1, 2, 3, 4 and 5+ bedrooms (USD), and the ZHVI per square foot (USD, typical value of all homes per square foot calculated by taking the estimated home value for each home in a given region and dividing it by the home's square footage).

Check out https://www.zillow.com/research/data/ for an overview of ZHVI and a deep-dive into its methodology.

Methodology

The methodolgy of this project can be described by the following steps:

A: Define the characteristics of the neighborhoods

1. Retrieve the names of neighborhoods for both cities, we will label the current residence as "old" city or neighborhood and the new destination as "new" city or neighborhood.
2. Put the names of the neighborhoods in a dataframe and add the latitude and longitude data.
3. Use the Foursquare API to get location data of the venues in all the neighborhoods.
4. Clustering of the neighborhoods with a clustering algorithm to find similar neighborhoods in both cities.
5. Define each cluster by checking the main characteristics based on venues data.
6. Pick the cluster that your old neighborhood is in and select the neighborhoods from the new city.

B: Scoring of neighborhoods within the new city to make a recommendation of the best places

7. Score each of the potential new neighborhood based on socioeconomic data and define the top 10 list.
8. Score remaining neighborhood based on crime data and we'll get the top 5 list.
9. Last step is to check housing prices for the top 3 list and decide based on financial resources available.

Geocoding will be used to generate the coordinates of the neighborhoods. Geocoding is the computational process of transforming a physical address description to a location on the Earth's surface (spatial representation in numerical coordinates). We use Nominatim Geocoding service, which is built on top of OpenStreetMap data.

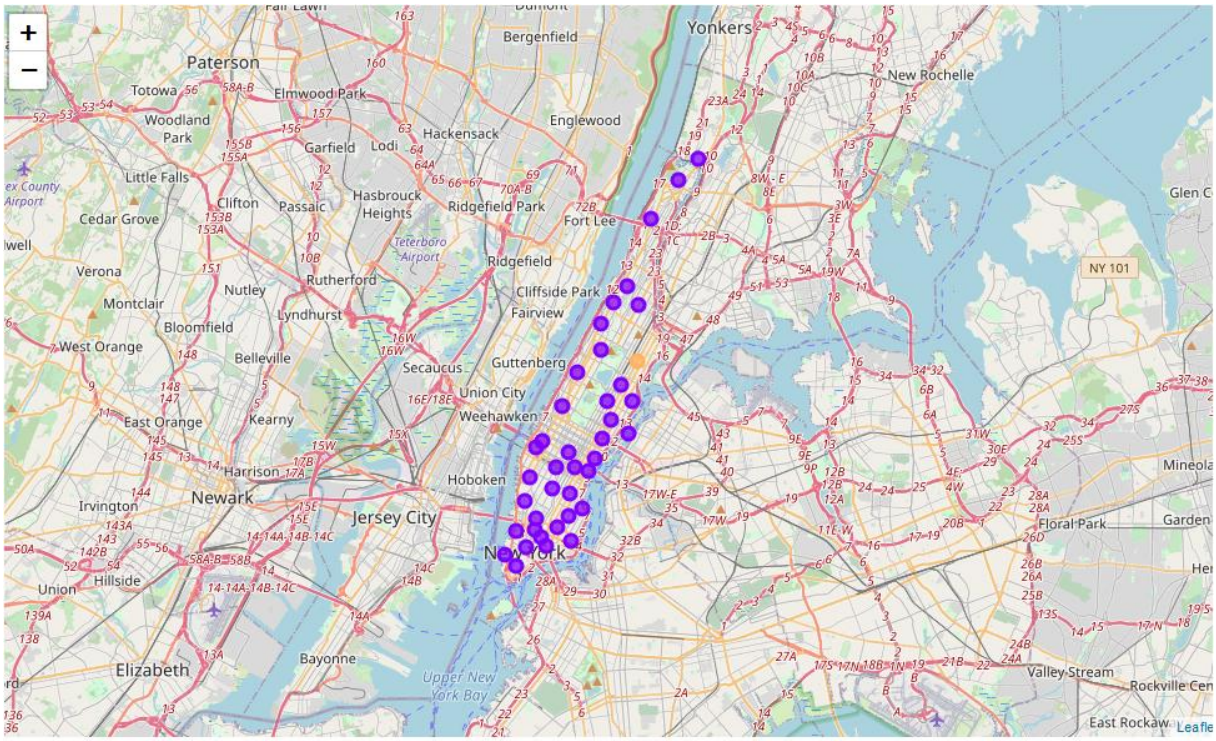
Results

As an example, for this project, I will be using following two cities:

New York is set to be the "old" home and serves as a starting point for the analysis

Chicago will be the "new" home and we will dive into a lot of data to find the best fit

There are 40 distinct neighborhoods in Manhattan and 75 distinct neighborhoods in Chicago that will be analysized during the project.

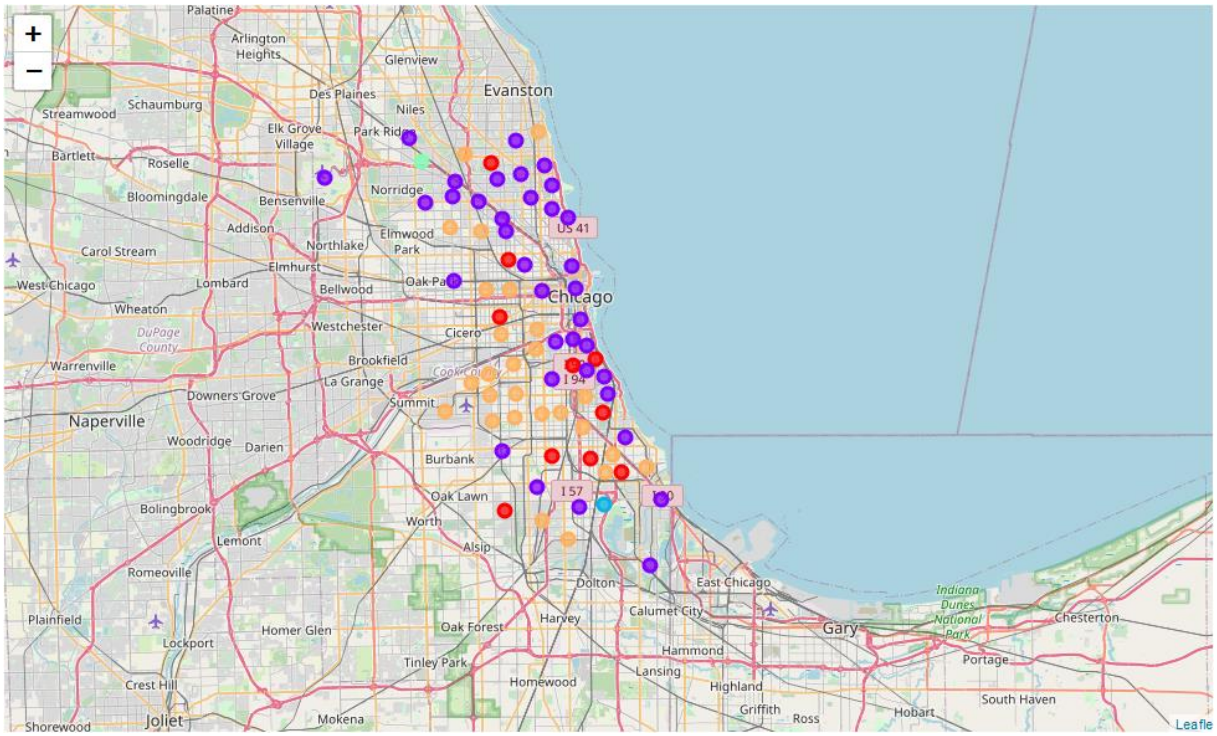


Map with neighborhoods in Manhattan, New York

Interestingly, when we put the neighborhoods in Manhattan together with the ones from Chicago, we find that they are quite similar. Almost all of them belong to Cluster 1. East

Harlem is the only neighborhood that was put into Cluster 4 and no single one was assigned to Cluter 2 and 3.

In contrast to Manhattan, the neighborhoods in Chicago seem to differ quite a lot. Here, we found a higher number of Cluster 0, 1, and 4. Whereas Cluster 2 hat a single neighborhood, Pullman and 3 is assigned to Norwood Park only.



Map with neighborhoods in Chicago

The clustering of the neighborhoods leads to the following results:

- Cluster 0 includes a lot of parks and restaurant, but apart from that it is quite hard to find a high similarity.
- Cluster 1 seems to have a very diverse mix of parks, restaurants and other interesting places, but it is also hard to find the one main topic that have all neighborhoods in common.
- Cluster 2 and 3 have only one neighborhood each. Interestingly, both of these neighborhoods look quite similar.
- Cluster 4 is interesting for our analysis since we have only one neighborhood in Manhattan, East Harlem, and multiples new neighborhoods to choose from.

Cluster 0 details:

	Neighborhood	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
52	North Park, Chicago	Bus Station	Gymnastics Gym	Park	Nature Preserve	Exhibit	Duty-free Shop	Eastern European Restaurant	Electronics Store	Empanada Restaurant	English Restaurant
61	Humboldt park, Chicago	Park	Food Truck	Baseball Field	Museum	Lake	Beach	History Museum	Café	Soccer Field	Yoga Studio
67	North Lawndale, Chicago	Seafood Restaurant	Train Station	Park	Construction & Landscaping	Coworking Space	Exhibit	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Electronics Store
74	Oakland, Chicago	Park	Boutique	Public Art	Track	Discount Store	Event Space	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Electronics Store
75	Fuller Park, Chicago	Fast Food Restaurant	Park	Sandwich Place	Yoga Studio	Ethiopian Restaurant	Dry Cleaner	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Electronics Store
80	Woodlawn, Chicago	Park	Coffee Shop	Yoga Studio	Exhibit	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Electronics Store	Empanada Restaurant	English Restaurant
82	Chatham, Chicago	Park	Boutique	Fast Food Restaurant	Bus Station	Donut Shop	Ice Cream Shop	Creperie	Eye Doctor	Eastern European Restaurant	Electronics Store
86	Calumet Heights, Chicago	Gym / Fitness Center	Bus Station	Park	Deli / Bodega	Financial or Legal Service	Filipino Restaurant	Dumpling Restaurant	Duty-free Shop	Fish Market	Eastern European Restaurant
109	Auburn Gresham, Chicago	Pool	Park	Discount Store	Basketball Court	Dry Cleaner	Duty-free Shop	Eastern European Restaurant	Electronics Store	Empanada Restaurant	English Restaurant
111	Mount Greenwood, Chicago	Cosmetics Shop	Park	Vineyard	Yoga Studio	Exhibit	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Electronics Store	Empanada Restaurant

Cluster 1 details: Excerp only

	Neighborhood	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
0	Marble Hill, New York City	Sandwich Place	Coffee Shop	Gym	Pharmacy	Deli / Bodega	Department Store	Diner	Discount Store	Kids Store	Donut Shop
1	Chinatown, New York City	Chinese Restaurant	Bakery	Cocktail Bar	American Restaurant	Coffee Shop	Spa	Salon / Barbershop	Optical Shop	Shanghai Restaurant	Asian Restaurant
2	Washington Heights, New York City	Café	Bakery	Grocery Store	Mexican Restaurant	Chinese Restaurant	Mobile Phone Shop	New American Restaurant	Spanish Restaurant	Coffee Shop	Latin American Restaurant
3	Inwood, New York City	Mexican Restaurant	Café	Bakery	Pizza Place	Lounge	Restaurant	Wine Bar	Frozen Yogurt Shop	Park	Deli / Bodega
4	Hamilton Heights, New York City	Pizza Place	Coffee Shop	Café	Mexican Restaurant	Deli / Bodega	Yoga Studio	Park	Caribbean Restaurant	School	Chinese Restaurant
5	Manhattanville, New York City	Coffee Shop	Seafood Restaurant	Deli / Bodega	Park	Mexican Restaurant	Italian Restaurant	Food & Drink Shop	Farmers Market	Lounge	Bike Trail
6	Central Harlem, New York City	Gym / Fitness Center	Chinese Restaurant	Seafood Restaurant	African Restaurant	Deli / Bodega	American Restaurant	Bar	French Restaurant	Fried Chicken Joint	Gym
8	Upper East Side, New York City	Italian Restaurant	Gym / Fitness Center	Coffee Shop	Exhibit	Bakery	Yoga Studio	Pizza Place	French Restaurant	Juice Bar	Spa
9	Yorkville, New York City	Italian Restaurant	Coffee Shop	Gym	Bar	Deli / Bodega	Wine Shop	Japanese Restaurant	Mexican Restaurant	Diner	Pizza Place
10	Lenox Hill, New York City	Italian Restaurant	Coffee Shop	Sushi Restaurant	Pizza Place	Cocktail Bar	Café	Burger Joint	Gym / Fitness Center	Gym	Salon / Barbershop
11	Roosevelt Island, New York City	Park	Farmers Market	Residential Building (Apartment / Condo)	Liquor Store	School	Scenic Lookout	Sandwich Place	Dog Run	Bridge	Kosher Restaurant
12	Upper West Side, New York City	Italian Restaurant	Bar	Bakery	Coffee Shop	Pizza Place	Ice Cream Shop	Bagel Shop	Thai Restaurant	Bookstore	Mediterranean Restaurant

Cluster 2 details:

	Neighborhood	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
88	Pullman, Chicago	History Museum	Yoga Studio	Exhibit	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Electronics Store	Empanada Restaurant	English Restaurant	Ethiopian Restaurant

Cluster 3 details:

	Neighborhood	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
49	Norwood Park, Chicago	Park	Yoga Studio	Exhibit	Dumpling Restaurant	Duty-free Shop	Eastern European Restaurant	Electronics Store	Empanada Restaurant	English Restaurant	Ethiopian Restaurant

So, we will define Cluster 4 as our example for the remainder of the project and the goal is to find the best fit for a person moving from East Harlem, NY to Chicago.

Now, we add socioeconomic data from the City of Chicago website to find the places which might be a not so good environment. This dataset contains a selection of six socioeconomic indicators of public health significance and a “hardship index,” for each Chicago community area, for the years 2008 – 2012. Scores on the hardship index can range from 1 to 100, with a higher index number representing a greater level of hardship.

I will use the Hardship Index only as this score includes each of the indicators.

	Neighborhood	Community Area Number	Hardship Index
0	Forest Glen	12.0	11.0
1	Morgan Park	75.0	30.0
2	Garfield Ridge	56.0	32.0
3	Rogers Park	1.0	39.0
4	Avalon Park	45.0	41.0
5	West Lawn	65.0	56.0
6	McKinley Park	59.0	61.0
7	West Pullman	53.0	62.0
8	Archer Heights	57.0	67.0
9	West Elsdon	62.0	69.0

Now we have reduced our list of potential new neighborhoods down to 10 and can go on to analyze further relevant data. From here, it looks like Forest Glen is quite a good spot to live.

Next, we add crime data into the analysis.

	Neighborhood	Community Area Number	Hardship Index	Count	Count per Year
0	Forest Glen	12.0	11.0	2584	608.000000
1	Archer Heights	57.0	67.0	4470	1051.764706
2	McKinley Park	59.0	61.0	5032	1184.000000
3	West Elsdon	62.0	69.0	5223	1228.941176
4	Avalon Park	45.0	41.0	6670	1569.411765

Again, Forest Glen (608 crime incidents per year over a five-year period) has the best score in relation to the crime statistics which should come at no surprise. Interestingly, however, is that neighborhoods with a high Hardship Index seem to have relatively low crime rates.

Last step is to check housing prices for the top 5 list and decide based on financial resources available.

	Neighborhood	Hardship Index	Crimes per Year	Current avg. Housing Prices
0	Forest Glen	11.0	608.000000	363921
1	Archer Heights	67.0	1051.764706	213038
2	McKinley Park	61.0	1184.000000	235814
3	West Elsdon	69.0	1228.941176	203498
4	Avalon Park	41.0	1569.411765	122019

Finally, the last step in my analysis shows a high relationship between socioeconomic data, crime data and the average housing prices for Forest Glen.

Discussion & Conclusion

The final decision for a new neighborhood can now be based on the financial resources available. If a budget of more than USD 360k is not a problem, than Forest Glen would be the best fit for someone moving from East Harlem, NY to Chicago.

In my project I tried to show a data-based solution to tackle the problem of choosing a new home neighborhood when people have to move to a new place. It takes a variety of data into account to find a neighborhood as similar as the current one or any given neighborhood that can serve as a starting point.