

Expatriation – choosing the best neighborhood to move in to Houston with ML. A practical case.

Table of Contents

1 Introduction.....	2
1.1 Background.....	2
1.2 Problem.....	2
1.3 Interest.....	3
1.4 Scope of this work.....	3
2 Data.....	5
2.1 Origin of data.....	5
2.2 About data used in this work.....	5
2.2.1 venues information from Foursquare API.....	5
2.2.2 Coordinate from ‘Search Nearby’ from Google Maps.....	5
2.2.3 Criminal records.....	6
2.2.4 Housing pricing from Redfin.....	7
3 Methodology.....	9
3.1 Clustering - finding regions with similitude between origin and destination.....	9
3.1.1 Splitting destination in sub regions (cells).....	9
3.1.2 Getting venue information.....	10
3.1.3 Houston cells analysis according to venues nearby.....	10
3.1.4 Origin region cluster determination.....	11
3.2 Refining according to House prices.....	12
3.2.1 Classify by prices ranges.....	13
3.3 Refine according to Criminal data.....	14
3.3.1 Assign offense to cell then calculate cell total score.....	17
4 Get the final top cell.....	17
5 Results.....	18
6 Discussion and Conclusion.....	21
7 References.....	22

1 Introduction

1.1 Background

Everyday, people move out of their home country to another one. The United Nations estimated the number of people living outside their home country to be around 232 millions^[1].

There is several motivations of why one could decides or be forced to expatriate to a foreign country. To list a few, this could be for:

- Business purpose; an employee is sent to another country by its company
- To flee conflict region or persecution
- For economical purpose; be able to find a job or simply for better salary conditions
- etc.

Whatever the reason is, expatriates will encounter a common challenge: leaving their home, there region they know well to another home, another region yet to be known.

1.2 Problem

Choosing the right spot isn't easy. Each particular situation could narrow down the choice, to a continent, a country or a town. But even when the destination town has been defined, the expatriate still face a large panel of choices. E.g., the choice could lays anywhere from one hour drive radius from the future work location. As diverse as the world can be, different choices can lead to completely different experiences.

Choosing a neighborhood that correspond to one aspiration is not an easy task. The success in choosing generally come with a solid knowledge of what a location can offer. This comes after several trials, errors and is helped by friends discussions, feedback and recommendations.

Expatriates will typically have limited knowledge about the destination prior to the expatriation as they don't live there yet. Also, as for friend recommendations, they will have few acquaintance to counsel prior or short after their arrival.

Still, it is possible to get help and recommendations through numerous local associations, books or simply over internet. This said, it is hard to find tailor made recommendations as one aspirations are by nature individuals.

It worth noting that expatriation candidate have a limited time to dedicate in finding the right spot to move in as there are many challenges encountered during the expatriation process. In fact, the choice of the neighborhood to move in is probably one of the last thing an expatriate candidate will focus in. Indeed, what would be the point to put so much effort finding the good neighborhood while the Visa application is still ongoing and the outcome uncertain? Is it really relevant while the location of the job

is not yet known? Expatriate with family must take care also that their child have are accepted by a proper school.

Even though the location is tackled at the end of the process, it deserve much attention as the expatriate will live there for a significant period of their life. In most case, renting means to commit for at least about 1 year. Buying, for even longer.

1.3 Interest

In the author opinion, being able to find the right neighborhood is a major factor of success of an expatriation. If there is other obvious factor of success as Visa/legal, work, etc. A bad geographical situation will brings problem that will come in addition to the other problems an expatriate have to face.

The destination location is important, for the expatriate that seek to enjoy is life as anyone else and this start with where he lives. What quality of life can be expected from a bad neighborhood?

The cost for companies expatriating employees are not negligible. The cost for employee should not be forgotten has well. We can list for instances:

- Visa/work permit and legal fee
- Moving cost
- Employee compensation
- Spouse quitting his job
- etc.

A bad experience will obviously leads to an early renunciation/stop of the expatriation experience or poorer employee morale. Without question, this cause a negative impact on the expected return of investment for both the company, the employee and his family.

1.4 Scope of this work

In this work, we develop an algorithm that aims to help choosing the best location for an expatriate based

- on his current location or a location he knows and would appreciate to live in
- a target town where he aims to settle

The purpose is to help the subject to choose the best neighborhood possible of a town that is unknown to him by clustering the destination town neighborhood according to their markers and find which ones are the closest from a neighborhood the expatriate know well and likes to live in.

In this work, we will focus on one practical case. The case chosen here is the case of an expatriate leaving in his hometown Liege, a town of 200,000 inhabitants of Belgium to move in to Houston, a town of 2,300,000 inhabitants of U.S.A. ^{[2][3]}

Having a current locations that he likes for its venues and surroundings, the goal set here is to determine, using machine learning algorithms, which neighborhood of the destination town he should choose.

Best neighborhood will be chosen under the following criteria:

1. Destination town neighborhood alike to home neighborhood in term of nearby venues
2. Safest places in term of criminality for a given housing budget.

The first part of this work will be done by splitting the destination town in same size square. For each subdivision, nearby venues data available from the API Foursquare will be listed, namely: Name, category and location. A clustering algorithm will be applied in order to group the subdivision that are alike. Then, the same model will be applied to the origin home neighborhood to determine which sub-region of the destination town are most alike.

For the second part, criminal record from HPD will be compiled and a criminality score will be attributed to each sub-region according of the offense type (more point awarded for a homicide than for a simple felony). In parallel, house pricing per ft² from Redfin will be compiled and attributed to each destination town sub-division. Then, a multi-objective analysis will be conducted to sort out the safest neighborhood by house price range.

Other features than the one cited above will not be taken into account in the analysis. For instances, following feature are excluded of the analysis:

- housing quality,
- distance from work locations
- weather difference (e.g. flood area)
- road conditions / public transportation
- Population density

At the end, a map generate using Folium library will be displayed with the best candidate per price ranges so the expatriate can choose it's new neighborhood according to his financial abilities.

2 Data

2.1 Origin of data

In this work, following data have been used.

1. Data on venues are coming from Foursquare API (<https://developer.foursquare.com>)
2. Coordinate of specific address have been obtained with the functionality 'Search Nearby' from Google Maps (<https://www.google.be/maps/>)
3. Criminal record data for Houston in 2019 have been taken from the NIBRS (https://www.houstontx.gov/police/cs/Monthly_Crime_Data_by_Street_and_Police_Beat.htm)
4. Housing pricing per zipcode have been provided by [Redfin](https://www.redfin.com/blog/data-center/), a national real estate brokerage. (<https://www.redfin.com/blog/data-center/>)

2.2 About data used in this work

2.2.1 venues information from Foursquare API

Foursquare has built a massive dataset of accurate location data all across the world. In this work, we are using the Foursquare API in order to get information about venues in a given range of a location.

Data of Foursquare API has been used in order to assess what kind of venues where available across the destination and origin town. Then, we have process the venues information in order to cluster the different subdivision of the destination town in several categories and compared them to the origin town.

Bellow are some extract of the Foursquare API used in this work.

	Cell_index	Cell_Latitude	Cell_Longitude	Venue	Venue_Category	Venue_Latitude	Venue_Longitude
0	0	29.601804	-95.561416	Walgreens	Pharmacy	29.600119	-95.563614
1	0	29.601804	-95.561416	Starbucks	Coffee Shop	29.599570	-95.563985
2	0	29.601804	-95.561416	Tornado Burger	Burger Joint	29.611505	-95.564204
3	0	29.601804	-95.561416	El Vaquero	Mexican Restaurant	29.589083	-95.564442
4	0	29.601804	-95.561416	CVS pharmacy	Pharmacy	29.600460	-95.565012

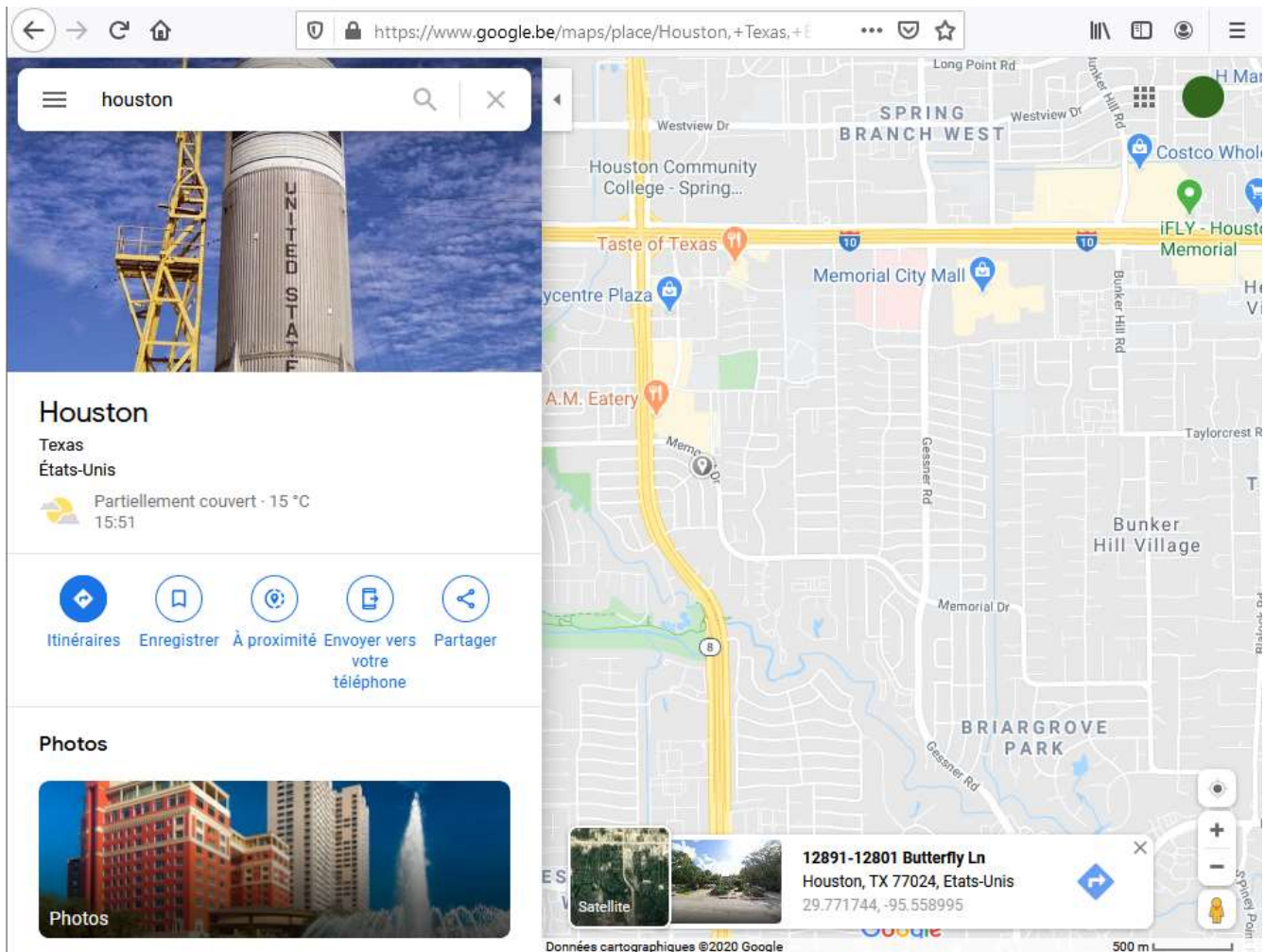
2.2.2 Coordinate from 'Search Nearby' from Google Maps

Maps is a service provided by Google allowing user to navigate on a world map and get information on what venues there is nearby.

The services has been using in this works for

- preliminary location explorations;
- and to get coordinates of the center of the destination town and of the selected origin address

The figure below show a snapshot of Google Maps centered on Houston:



2.2.3 Criminal records

The dataset used contains a breakdown of Group "A" Offenses for which HPD wrote police reports. The data is broken down by police districts and beats, and displayed by street name and block range.

Those data have been used first assigning each entries to a subdivision of the destination town, then a criminality score have been determined by summing each entries. A weight have been given according to the offense type:

1. no physical damage (theft, fraud) = 1 point
2. simple assault = 2 points
3. aggravated assault = 3 points

4. Rape = 5 point

5. Homicide = 10 point

An extract of the original data is shown here under:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Incident	Occurrence Date	Occurrence Hour	NIBRS Class	NIBRS Description	Offense Count	Beat	Premise	Block Range	Street Name	Street Type	Suffix	ZIP Code
2	5619	01/01/2019	0	290	Destruction, damage, vandalism		1'9C30	Residence, Home (Includes Apartment)	9622	SAN CARLOS			77013
3	17319	01/01/2019	0	35A	Drug, narcotic violations		1'7C10	Highway, Road, Street, Alley	0	EAST	FWY		77020
4	18119	01/01/2019	0	290	Destruction, damage, vandalism		1'16E40	Residence, Home (Includes Apartment)	16718	LONE QUAIL	CT		77489
5	19019	01/01/2019	0	520	Weapon law violations		1	Residence, Home (Includes Apartment)	1909	MELBOURNE			77026-000
6	20519	01/01/2019	0	13A	Aggravated Assault		1'15E30	Residence, Home (Includes Apartment)	4034	OSBY	DR		77025
7	20519	01/01/2019	0	23H	All other larceny		1'15E30	Residence, Home (Includes Apartment)	4034	OSBY	DR		77025
8	20519	01/01/2019	0	290	Destruction, damage, vandalism		1'15E30	Residence, Home (Includes Apartment)	4034	OSBY	DR		77025
9	20519	01/01/2019	0	35A	Drug, narcotic violations		1'15E30	Residence, Home (Includes Apartment)	4034	OSBY	DR		77025
10	31119	01/01/2019	0	13B	Simple assault		1'19G20	Residence, Home (Includes Apartment)	4700	KIRKWOOD	RD	S	77072
11	31119	01/01/2019	0	290	Destruction, damage, vandalism		1'19G20	Residence, Home (Includes Apartment)	4700	KIRKWOOD	RD	S	77072
12	31119	01/01/2019	0	13A	Aggravated Assault		2'19G20	Residence, Home (Includes Apartment)	4700	KIRKWOOD	RD	S	77072
13	34419	01/01/2019	0	13B	Simple assault		1'5F10	Residence, Home (Includes Apartment)	7429	LONG POINT	RD		77055
14	34819	01/01/2019	0	290	Destruction, damage, vandalism		1'15E30	Residence, Home (Includes Apartment)	4065	SILVERWOOD	DR		77025
15	34819	01/01/2019	0	13B	Simple assault		2'15E30	Residence, Home (Includes Apartment)	4065	SILVERWOOD	DR		77025
16	39919	01/01/2019	0	13B	Simple assault		1'20G10	Residence, Home (Includes Apartment)	3500	WOODCHASE	DR		77042
17	40219	01/01/2019	0	13A	Aggravated Assault		1'13D20	Residence, Home (Includes Apartment)	6502	ROXBURY	RD		77087
18	40519	01/01/2019	0	13B	Simple assault		1'19G20	Residence, Home (Includes Apartment)	6819	COOK	RD		77072
19	41419	01/01/2019	0	290	Destruction, damage, vandalism		1'7C20	Residence, Home (Includes Apartment)	4521	KASHMERE	ST		77026
20	43019	01/01/2019	0	35A	Drug, narcotic violations		1'10H40	Highway, Road, Street, Alley	1160	GRAY	ST		77002
21	43019	01/01/2019	0	35B	Drug equipment violations		1'10H40	Highway, Road, Street, Alley	1160	GRAY	ST		77002
22	43719	01/01/2019	0	13B	Simple assault		2'11H40	Residence, Home (Includes Apartment)	1114	CHOATE	CIR		77017
23	44419	01/01/2019	0	13B	Simple assault		2'10H40	Bar, Nightclub	2600	TRAVIS	ST		77006
24	45919	01/01/2019	0	120	Robbery		1'1A30	Highway, Road, Street, Alley	2200	SOUTHWEST	FWY		77098

2.2.4 Housing pricing from Redfin

Housing pricing in the destination Town have been extracted from Redfin database, a US national real estate brokerage.

Price per sq foot for renting and buying have been used for each zipcode have been used. Prices have been assigned for each destination town subdivision.

For the subdivision of interest, the prices have been processed with the criminality score calculated using HPD dataset in order to determine the best subdivision: for each category of subdivision determined using Foursquare, what was the safer subdivision for a price range.

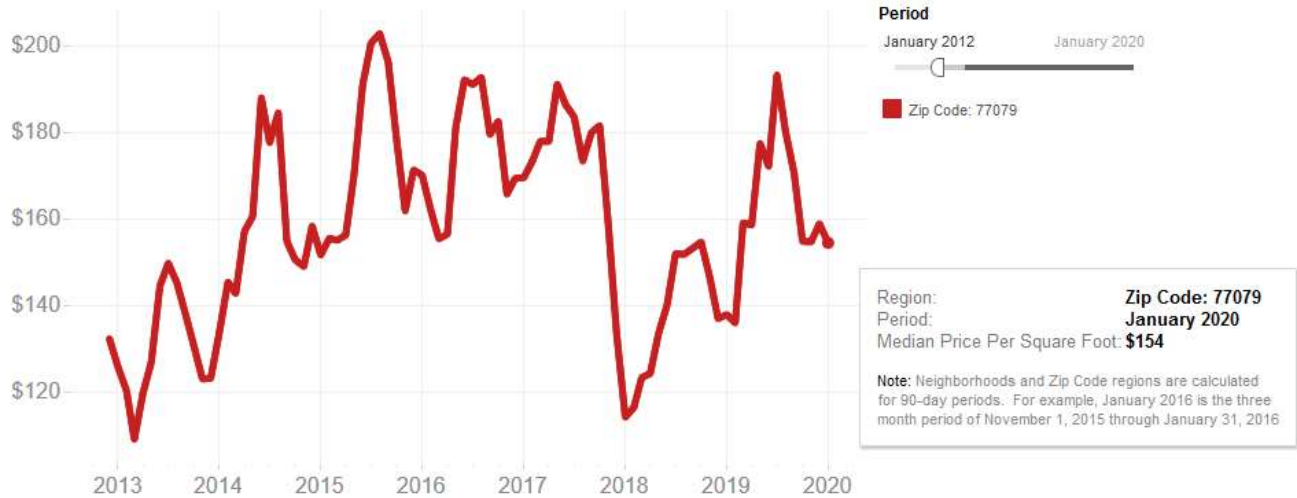
Here under is an example of information provided by Redfin

Home Prices, Sales & Inventory

Active Listings Months Supply Days on Market Price Per Square Foot Sale-to-List Sold Above List Price Drops Of

State: TX Region Type: Zip Code Region: Zip Code: 77079 Property Type: All Residential Show Values As: Value Seasonally Adjusted: False

Median Sale Price Per Square Foot



REDFIN

+ a b l e a u

Navigation icons: back, forward, search, etc.

3 Methodology

The methodology in finding the best location is divided in two phases:

1. Finding location in destination town alike to the 'origin' address (assumed to be liked by the person seeking to move)
2. Inside those locations alike, find the best one in term of housing prices (per squared feet) and criminality (score given according to offense)

The subject used in this work is

- destination town : Houston, Texas, USA
- origin address : Boulevard Piercot, Liege, Belgium

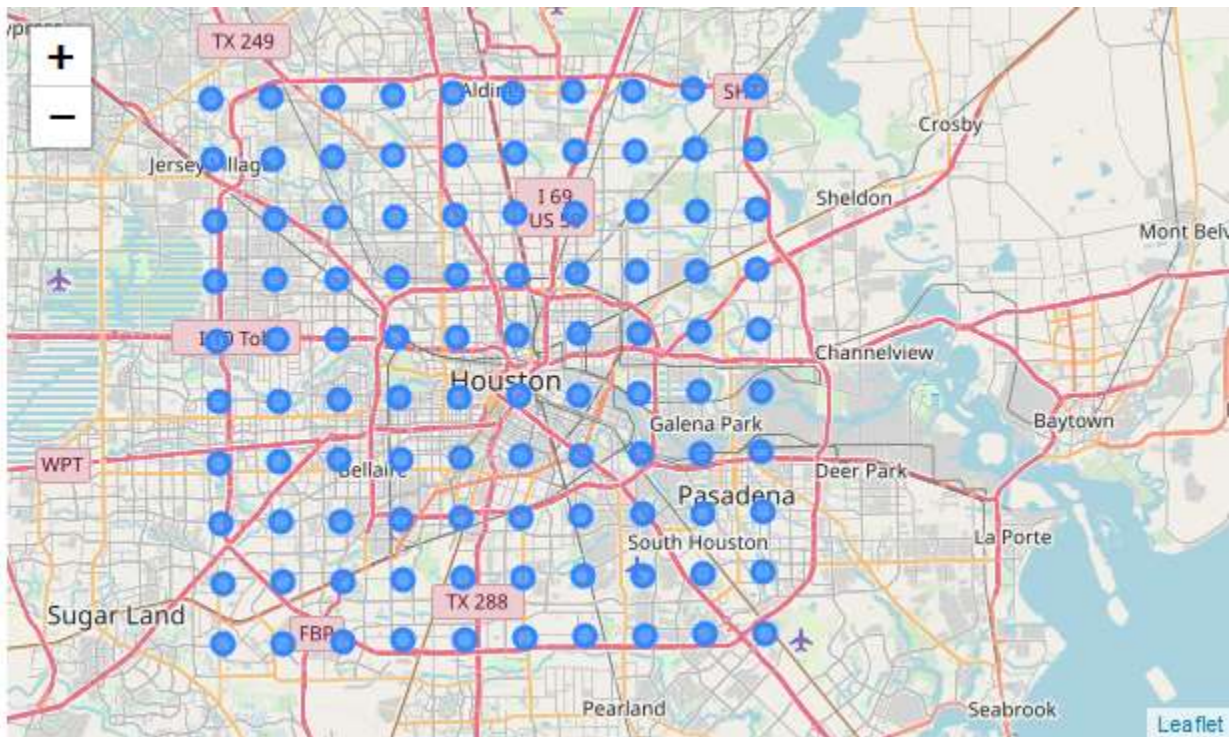
3.1 Clustering - finding regions with similitude between origin and destination

3.1.1 Splitting destination in sub regions (cells)

Houston, the destination town chosen, is split in 4x4 kilometer square. Each split will be called Cell in this work. There is a total of 100 cells considered.

We start from the center coordinate of Houston and get the coordinate for each cell by applying an offset of 4km in X and Y UTM projection. For doing so, we used the UTM package that allows to transform coordinate to/from UTM projection from/to coordinate. Use of UTM projection is needed because we cannot directly translate coordinate by a given 'flat' distance.

Center has been slightly shifted so each cell are inside the belt 8, use as informal city limits by some residents. The image below shows the distribution of the cells on the map with Folium package.



3.1.2 Getting venue information

Using Foursquare API, we got all venue in a given range for each cell and put that in a dataframe. Venues were collected in a 1500m radius to each cell. Ideally, we should have selected a larger radius or more cell, unfortunately, this work has been conducted using a free license of Foursquare API and consequently the number of requests were limited. The image below represents a slice of the dataset obtained.

```
df_Houston_Cell_Venue.head(5)
```

	Cell_index	Cell_Latitude	Cell_Longitude	Venue	Venue_Category	Venue_Latitude	Venue_Longitude
0	0	29.601804	-95.561416	Walgreens	Pharmacy	29.600119	-95.563614
1	0	29.601804	-95.561416	Starbucks	Coffee Shop	29.599570	-95.563985
2	0	29.601804	-95.561416	Tornado Burger	Burger Joint	29.611505	-95.564204
3	0	29.601804	-95.561416	El Vaquero	Mexican Restaurant	29.589083	-95.564442
4	0	29.601804	-95.561416	CVS pharmacy	Pharmacy	29.600460	-95.565012

3.1.3 Houston cells analysis according to venues nearby

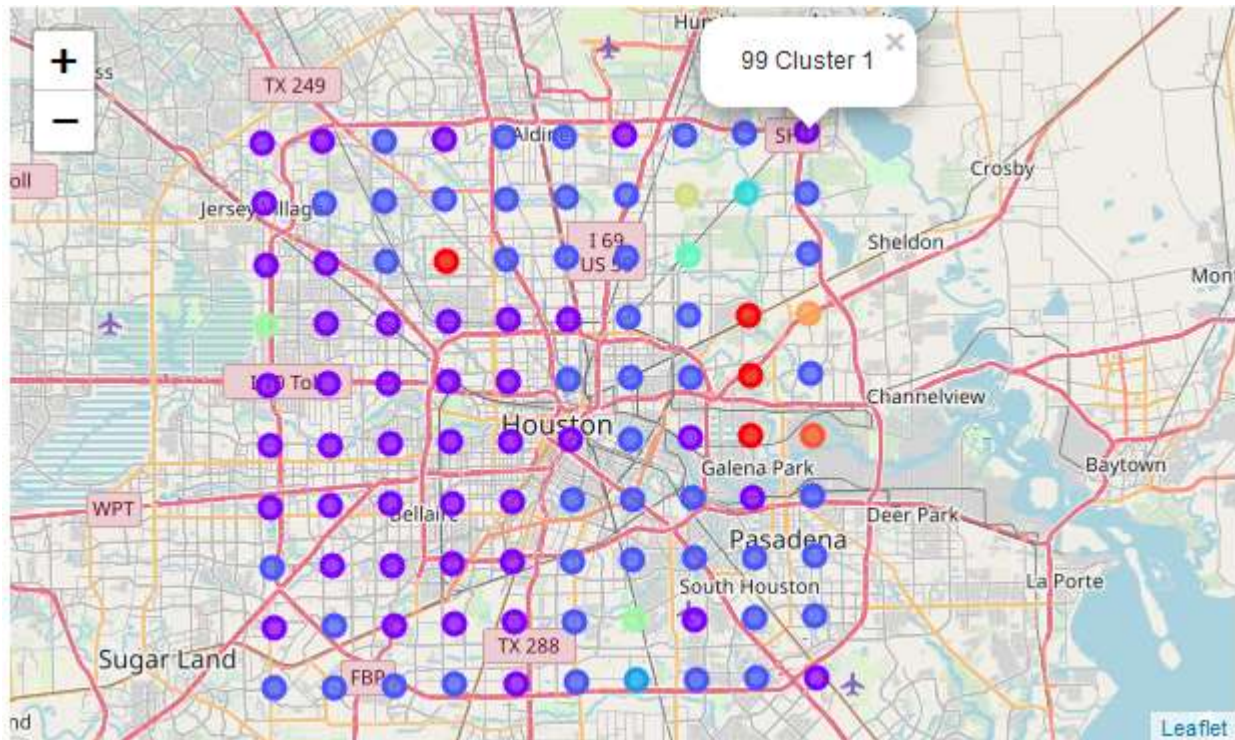
For each cell, the repartition of the nearby venue has been determined. Then a K-means clustering has been applied. The final K number selected is 10.

This number has been chosen after several trial in order to meet an optimum:

- Too small, the cut is not efficient. There is cluster containing only 1 or 2 cells that are very different from the other
- Too high, there is only small cluster

The goal here was to not segregate too much the cells because the cells will be further classify according to price range and criminality score.

The image bellow shows the map (using Folium) of the cells divided in ten cluster:



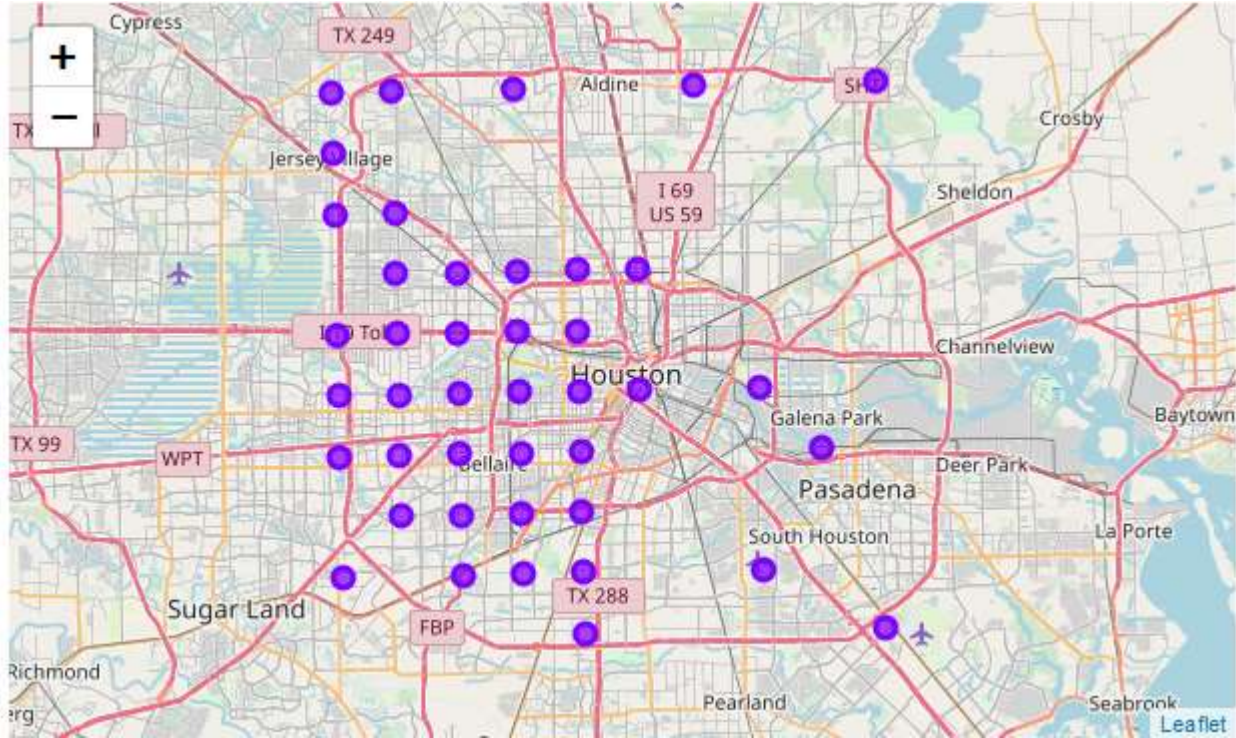
As it can be see, sum cluster only contains one or two cells, then there is two main cluster containing most of the cells are:

- Cluster 1 in purple
- Cluster 2 in blue

Further segregation could have been done based on venues nearby. Instead, it has been decided to continue the segregation based on cell prices and criminality.

3.1.4 Origin region cluster determination

Using the same K-means model, the origin address has been clustered. It has been determined that the closest cluster is the number 1. The following map shows the destination cells alike to the origin address:



Each square is analyzed according to available foursquare data on nearby venue. Then, each square will be clusterized using K-means clustering technics Then, the origin town neighborhood will be assigned a cluster using the same model. A map showing location having the most similitude with the origin will be shown using geopy***

3.2 Refining according to House prices

Prices data from RedFin dataset have been used. RedFin provided numerous information regarding to house selling, especially in USA.

Among others, average selling price per square feet is provided by zip code. These data have been manually compiled in a csv files and further worked using Pandas DataFrame.









	zipcode	price_sqft
0	77459	111
1	77477	100
2	77099	89
3	77072	89
4	77042	109

Then, the zip code of each cell is determined using geopy library. This library allows to get the address, zip code included, from the coordinate of the given cells. Finally, cells dataset are merged with price per squared feet.

Cell_index	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	zipcode	price_sqft
0	2	Mexican Restaurant	Pharmacy	Coffee Shop	Storage Facility	Fast Food Restaurant	Burger Joint	Camera Store	Gas Station	Sandwich Place	Juice Bar	77459	111.0
1	1	Mexican Restaurant	Sandwich Place	Seafood Restaurant	Deli / Bodega	Video Store	Bar	Comfort Food Restaurant	Pharmacy	Fried Chicken Joint	Food	77477	100.0
2	2	Fast Food Restaurant	Mobile Phone Shop	Mexican Restaurant	Gas Station	Video Store	Asian Restaurant	Storage Facility	Other Repair Shop	Fried Chicken Joint	Motel	77099	89.0
3	1	Chinese Restaurant	Asian Restaurant	Bubble Tea Shop	Vietnamese Restaurant	Korean Restaurant	Bakery	Hotel	Ice Cream Shop	Juice Bar	Fast Food Restaurant	77072	89.0
4	1	Hotel	Sandwich Place	Grocery Store	Bakery	Chinese Restaurant	Mexican Restaurant	Pizza Place	Discount Store	Japanese Restaurant	Breakfast Spot	77042	109.0

3.2.1 Classify by prices ranges

Inside the cluster 1, the same of the origin address, the price goes from 87\$ to 360\$ per ft² to buy a house. Cell cluster 1 has been further subdivided in ten more price range and are shown in the following map with:

- Purple  : cat 0 : 87.0 - 114.3\$
- Blue  : cat 1 : 114.3 - 141.6\$
- Blue2  : cat 2 : 141.6 - 168.9\$
- Blue3  : cat 3 : 168.9 - 196.2\$
- Green  : cat 4 : 196.2 - 223.5\$
- Green2  : cat 5 : 223.5 - 250.8\$
- Yellow  : cat 6 : 250.8 - 278.1\$
- (None in that cat): cat 7 : 278.1 - 305.4\$
- (None in that cat): cat 8 : 305.4 - 332.7\$
- Red  : cat 9 : 332.7 - 361.0\$

Incident	Occurrence Date	Occurrence Hour	NIBRS Class	NIBRSDescription	Offense Count	Beat	Premise	Block Range	StreetName	Street Type	Suffix	ZIP Code	
0	8220	01/01/2020	0	23G	Theft of motor vehicle parts or accessory	1	8C50	Residence, Home (Includes Apartment)	9311	BELLA PINE	CT	NaN	77078
1	23020	01/01/2020	0	290	Destruction, damage, vandalism	1	11H30	Residence, Home (Includes Apartment)	8064	LENORE	NaN	NaN	77017
2	24120	01/01/2020	0	13B	Simple assault	1	17E10	Residence, Home (Includes Apartment)	5930	DASHWOOD	DR	NaN	77081
3	27120	01/01/2020	0	290	Destruction, damage, vandalism	1	14D30	Residence, Home (Includes Apartment)	5218	KENILWOOD	DR	NaN	77033
4	29320	01/01/2020	0	290	Destruction, damage, vandalism	1	20G30	Other, Unknown	2851	WALLINGFORD	DR	NaN	77042

Description have been translated as a score using following rules:

General philosophy:

1. no physical damage (theft, fraud) = 1 point
2. simple assault = 2 points
3. aggravated assault = 3 points
4. Rape = 5 point
5. Homicide = 10 point

In details:

- 'Theft of motor vehicle parts or accessory' : 1,
- 'Destruction, damage, vandalism':1,
- 'Simple assault':2,
- 'Theft from motor vehicle':1,
- 'Drug, narcotic violations':1,
- 'Aggravated Assault':3,
- 'Intimidation':2,
- 'Robbery':1,
- 'Forcible rape':5,
- 'Motor vehicle theft':1,
- 'Credit card, ATM fraud':1,
- 'Identify theft':1,
- 'Drug equipment violations':1,
- 'Murder, non-negligent':10,

- 'All other larceny':1,
- 'Shoplifting':1,
- 'Arson':2,
- 'Weapon law violations':1,
- 'Pocket-picking':1,
- 'Burglary, Breaking and Entering':1,
- 'Forcible sodomy':5,
- 'Counterfeiting, forgery':1,
- 'Pornographs, obscene material':5,
- 'Theft from building':1,
- 'Forcible fondling':1,
- 'Purse-snatching':2,
- 'Bribery':1,
- 'Extortion, Blackmail':2,
- 'Embezzlement':1,
- 'False pretenses, swindle':1,
- 'Human Trafficking/Commercial Sex Act':5,
- 'Animal Cruelty':2,
- 'Kidnapping, abduction':5,
- 'From coin-operated machine or device':1,
- 'Statutory rape':5,
- 'Stolen property offenses':1,
- 'Prostitution':1,
- 'Impersonation':1,
- 'Wire fraud':1,
- 'Hacking/Computer Invasion':1,
- 'Purchasing prostitution':1,
- 'Incest':5,

'Assisting or promoting prostitution':1,

'Sexual assault with an object':5}

Then, each crime has been assigned to the closest cell and then summed in order to obtain a total score.

High score depict high criminality in the area.

3.3.1 Assign offense to cell then calculate cell total score

Assignment of a crime has been done using geopy to get the coordinate of each crime record. Then, closest cell for each record has been found using the euclidian distance to each cell coordinate :

$$\text{Min of SQRT} ((\text{Cell_lat}-\text{Crime_lat})^2 + (\text{Cell_lon}-\text{Crime_lon})^2)$$

Then, all score for each cell has been summed to get the total criminality score.

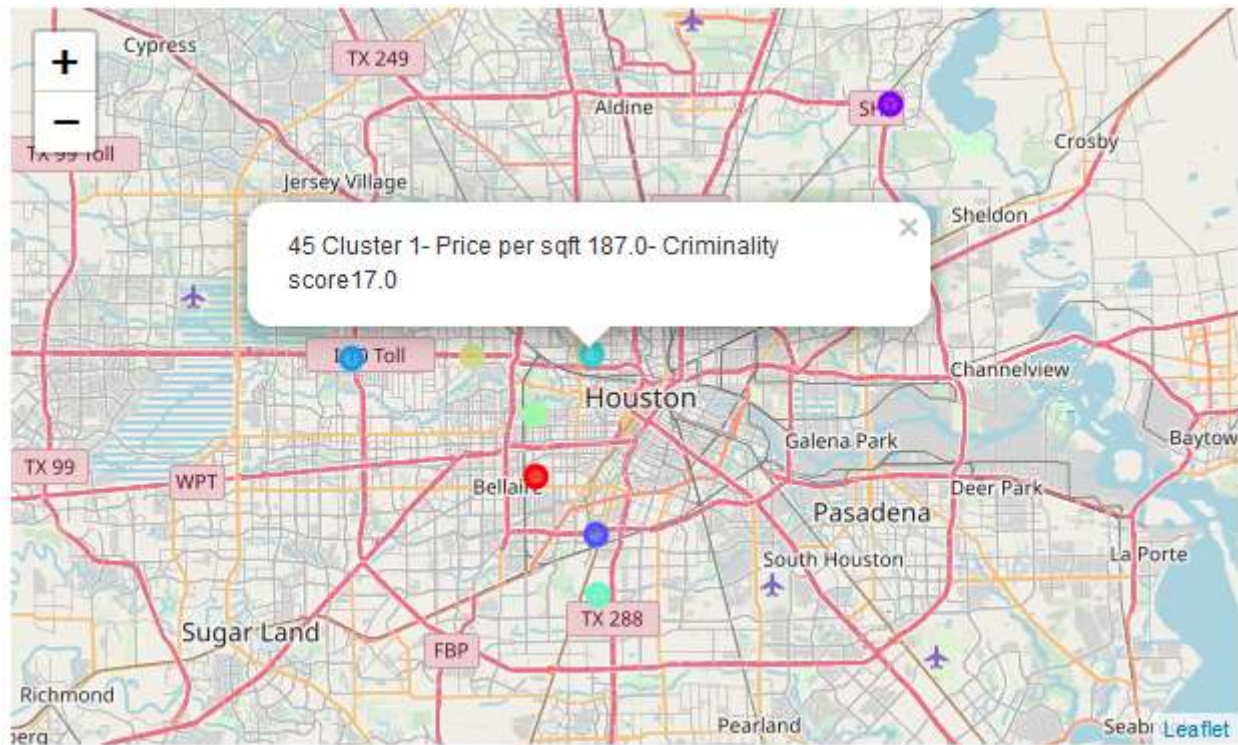
Finally, crime data have been merged with the dataset containing the cells, their venue and venue cluster, their price and price category.

	Cell_Latitude	Cell_Longitude	Cell_index	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	zipcode	price_sqft	price_cat	criminality
99	29.933116	-95.197018	99		Fast Food Restaurant	Asian Restaurant	Pizza Place	Department Store	Mexican Restaurant	Burger Joint	Pharmacy	Sushi Restaurant	Bank	Fried Chicken Joint	77044	98	0	0
5	29.782123	-95.565997	5		1 Restaurant	Mexican Restaurant	Burger Joint	Coffee Shop	American Restaurant	Furniture / Home Store	Cosmetics Shop	Sandwich Place	Seafood Restaurant	Steakhouse	77079	145	2	16
42	29.677027	-95.398041	42		1 BBQ Joint	Rental Car Location	Gas Station	General Entertainment	Bar	Pizza Place	Fast Food Restaurant	Moving Target	Smoke Shop	Theme Park Ride / Attraction	77054	135	1	4
25	29.783703	-95.48331	25		1 Spa	Donut Shop	Italian Restaurant	Sushi Restaurant	Liquor Store	Pizza Place	Video Store	Gas Station	Pet Store	Pizza Place	77024	255	6	2
45	29.785231	-95.400617	45		1 Mexican Restaurant	Creole Restaurant	Coffee Shop	Park	Spa	American Restaurant	Bar	Ice Cream Shop	Italian Restaurant	Juice Bar	77007	187	3	17
33	29.71234	-95.440216	33		1 Pizza Place	Gym	Mexican Restaurant	Bakery	American Restaurant	Sandwich Place	Coffee Shop	Italian Restaurant	Convenience Store	Gym / Fitness Center	77005	360	9	500
34	29.748407	-95.44109	34		1 Furniture / Home Store	Women's Store	Hotel	Coffee Shop	Clothing Store	Sandwich Place	Cosmetics Shop	Bank	French Restaurant	Shopping Mall	77019	229	5	3096
41	29.640958	-95.397185	41		1 Park	Candy Store	Golf Course	Museum	Athletics & Sports	Sandwich Place	Soccer Field	Rugby Stadium	Electronics Store	Donut Shop	77030	203	4	4

4 Get the final top cell

Then, for each category price, the lowest criminality cell has been selected and displayed in a map.

Those cells shown represent the most similar cell in term of venues to the origin address with the lowest criminality score per price range:



5 Results

The final map shows the selected address that a person should consider to move in. They answer to following constraints:

1. Be in a given location, here Houston
2. Be similar of a provided address of reference, here Bvd Piercot, in Liege, Belgium
3. Have the lowest criminality score per price category

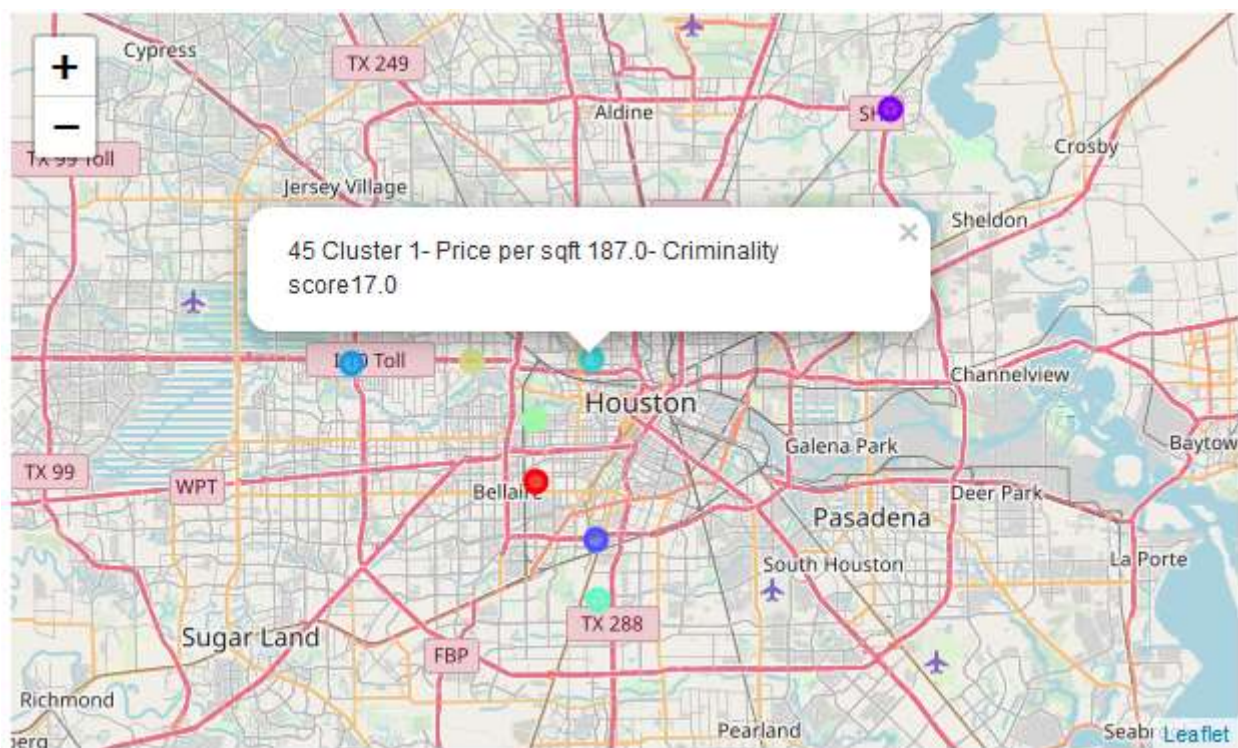


Table next table provide a summary of cell information for further considerations

Cell_Latitude	Cell_Longitude	Cell_index	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	zipcode	price_sqft	price_cat	criminality
29.933116	-95.197018	99		Fast Food Restaurant	Asian Restaurant	Pizza Place	Department Store	Mexican Restaurant	Burger Joint	Pharmacy	Sushi Restaurant	Bank	Fried Chicken Joint	77044	98	0	0
29.782123	-95.565997	5		1 Clothing Store	Mexican Restaurant	Burger Joint	Coffee Shop	American Restaurant	Furniture / Home Store	Cosmetics Shop	Sandwich Place	Seafood Restaurant	Steakhouse	77079	145	2	16
29.677027	-95.398041	42		1 BBQ Joint	Rental Car Location	Gas Station	General Entertainment	Bar	Pizza Place	Fast Food Restaurant	Moving Target	Smoke Shop	Theme Park Ride / Attraction	77054	135	1	4
29.783703	-95.48331	25		1 Spa	Donut Shop	Italian Restaurant	Sushi Restaurant	Liquor Store	Sandwich Place	Video Store	Gas Station	Pet Store	Pizza Place	77024	255	6	2
29.785231	-95.400617	45		1 Mexican Restaurant	Cajun / Creole Restaurant	Coffee Shop	Park	Spa	American Restaurant	Bar	Ice Cream Shop	Italian Restaurant	Juice Bar	77007	187	3	17
29.71234	-95.440216	33		1 Pizza Place	Gym	Mexican Restaurant	Bakery	American Restaurant	Sandwich Place	Coffee Shop	Italian Restaurant	Convenience Store	Gym / Fitness Center	77005	360	9	500
29.748407	-95.44109	34		1 Furniture / Home Store	Women's Store	Hotel	Coffee Shop	Clothing Store	Sandwich Place	Cosmetics Shop	Bank	French Restaurant	Shopping Mall	77019	229	5	3096
29.640958	-95.397185	41		1 Park	Candy Store	Golf Course	Museum	Athletics & Sports	Sandwich Place	Soccer Field	Rugby Stadium	Electronics Store	Donut Shop	77030	203	4	4

Further more, we can note that following venue where not found in the destination town. This is interesting to have this in mind. If it is important type of venue seeks by one, they will simply not found it. In this case, the following where not found in Houston.

- Brasserie
- Bridge Club
- Butcher
- Friterie
- Moroccan Restaurant
- Opera House

6 Discussion and Conclusion

It is a complex task to objectify and determine what an individual will like or not. It touch the very nature of our emotion and by definition are not quantifiable. However, there are some common features triggering common emotions. Given, enough data on these feature, data science can help us finding what makes one happy.

This work shows it is possible to help people finding a new home based on yet available data such as venues, prices, criminality. There is other feature worth being analyzed, we can list for example beauty of the landscape, local pollution, people in those area, culture, etc. There is of course other feature that are not yet described on data and some features that, today, could not be catch in dataset. There will always be a part of our emotions that will be indiscernible and remains mysterious. That why that one should not solely bases his choices on algorithm. But when time is lacking, those algorithm help narrow the search field and present unthought options.

Concerning the presented works, results could be greatly improved by

- Using more cell. This will grant a better definitions of the different neighborhood of choice. Of course, this need more raw power and more API privilege (number of request, power allotted)
- Extend venue radius for analysis (Foursquare limitation toward number of request)
- For venues, take into account the distance from a central point. The closer the better (or worse depending of type of venue)
- Criminality, use a larger time frame (geopy limitation on request, timeout)
- Criminality instead of assigning to closest cell, account for each crime committed in a certain radius
- Review criminality scoring

7 References

1. DAILY MAIL, "*More people than ever living outside their home country*". 12 September 2013. [viewed 5 March 2020].
Available from: <http://www.dailymail.co.uk/news/article-2418902/More-people-living-outside-home-country-Number-migrants-worldwide-hits-232-million.html>
2. WIKIPEDIA, "*Liege*" 1 March 2020. [viewed 5 March 2020].
Available from: <https://en.wikipedia.org/wiki/Liège>
3. WIKIPEDIA, "*Houston*" 1 March 2020. [viewed 5 March 2020].
Available from: <https://en.wikipedia.org/wiki/Houston>
4. <https://developer.foursquare.com>
5. [https://www.houstontx.gov/police/cs/Monthly Crime Data by Street and Police Beat.htm](https://www.houstontx.gov/police/cs/Monthly_Crime_Data_by_Street_and_Police_Beat.htm)
6. <https://www.google.be/maps/>
7. <https://www.redfin.com/blog/data-center/>