A Personalised Neighbourhood Recommender Model

By Hugo Herrera, June 2020

1. Introduction

1.1. Background

In these days, it is more and more common to have to leave your country due to a job offer abroad. One of the most challenging issues in a situation like this is to find a good place to live, and good can be interpreted as a place located in an environment as similar as possible to the current homeplace.

The following example helps to illustrate why finding similar environments it is relevant in situations like this. Let's imagine for a second a person who cannot live without coffee, as many of us do, so every day before going to work, this person walks to her favourite coffee shop that is just around the corner. This person also has the habit of having a spicy tuna roll on Wednesday's evening in the Japanese Restaurant that is two blocks from her home, lastly, jogging on Saturday's morning in the neighbourhood's park is a most for her.

1.2. Problem Description

Although it is possible to spend some hours on the internet searching for tips and recommendations on where to live, blogs and articles could consider some aspects in their analysis that are not necessarily aligned with the specific preferences of the person who is looking for recommendations. For instance, a blogger could prioritise neighbourhoods that gave access to subway stations or that have a vivid nightlife, whereas what our person used as example wants is coffee shops and Japanese restaurant options nearby, notwithstanding a rent in a range of her possibilities.

In order to provide more suitable recommendations, the proposed model takes into account the specific preferences of a person facing a situation like this, it considers the type of venues the user would like to have nearby, as well as the rent amount willing to pay, and location of the offices to be relocated.

2. Data

Data requirements are classified in two blocks as follows:

2.1. **Data provided by the user** (A user is a person who is looking for a neighbourhood to live)

- Apartment monthly rent willing to pay
- Percentage of variation of rent to pay
- List of top 5 venue categories ranked from 1 to 5
- The office where the user will be relocated. It could be the office description, address, or location

2.2. **External data –** Not provided by the user

- The average amount of rent per neighbourhood

        - The average rent cost per neighbourhood in Mexico City was obtained from the website propiedades.com (2020)
    - Venues data
        - The list of venues and their category per neighbourhood was gathered from Foursquare through the Foursquare API
    - Neighbourhoods geolocation data
        - Latitude and longitude of all neighbourhoods considered were obtained from a data repository provided by the Mexico City government (CDMX 2020)
    - List of the top five boroughs to live in Mexico City
        - The top five list was set based on the report titled "Informe de Desarrollo Humano Municipal 2010-2015. Transformando México desde lo local", which was carried out by the United Nations Development Program Mexico (UNDP 2019)

3. Methodology

3.1. User's data

As mentioned in the previous section, to catch and use the user's input is vital to come up with a model tailored according to user's preferences. Having said that, the model was developed based on a set of data defined as if a real user provided it. This initial set of simulated user's data is summarised on the following tables:

User's preferred venue categories		
Ranking	Venue Category	Rating
1	Coffee Shop	10
2	Japanese Restaurant	8
3	Gym / Fitness Center	6
4	Mexican Restaurant	4
5	Park	2

Table 1: User's venue categories preferences

Table 1 above shows the list of venue categories that were selected to test the model. As it could be appreciated, along with the ranked list of venue categories a rating score was added having values from 10 to two, where one was given to the top one user's preference and a rate of two given to the 5th venue category preference. These ratings were used later in the process to create a weighted matrix of venue categories. A pandas datraframe was created with the same structure as Table 1 to utilise this information in the model.

It is worth noting that for this exercise, it was assumed that the user is looking specifically for an apartment to rent. Based on this assumption, a monthly rent of $15,000 Mexican Pesos was considered as well as an allowed variation of 20% from the baseline budget (Table 2). With these pieces of information, an upper limit of $18,000 and a lower limit of $12,000 were calculated and used to delimit the list of neighbourhoods to recommend. All data containing in Table 2 was stored as variables to be used in the model.

Data provided by the user

Monthly rent	$15,000
Variation of rent	20%
Office in Mexico City	IBM, Mexico City

Table 2: Test data, simulating a user's input

3.2. External data

Mexico City is well-known for its size, and this means that there are a lot of neighbourhoods dispersed in a bit less than 1,500 km^2. Having this in mind, the search for neighbourhoods to recommend was limited to the top 5 municipalities of Mexico City, that according to the research performed by the UNPD (2019) have the higher human development index (HDI). This index indicates that the municipalities with higher values presumably provide the best conditions to live, since the HDI is a measure that synthesises the level of improvement of countries, states, and municipalities in three basic dimensions of human development: Health, Education, and Income. (UNPD 2010).

The following table shows the top 5 municipalities that were considered as a filter in the model due to their high HDI:

Municipality	HDI
Benito Juárez	0.9440
Miguel Hidalgo	0.9171
Coyoacán	0.8827
Cuauhtémoc	0.8784
Iztacalco	0.8612

Table 3: HDI by municipality

Having the above's list of Municipalities, the list of neighbourhoods that belong to each of these Municipalities along with their geolocation data was gathered from an open data repository provided by the Mexico City government (CDMX 2020). This result in a list of 404 neighbourhoods.

The next step was to get the average rent for each of the 404 neighbourhoods; the data required was obtained from the website porpiedades.com (2020). The search settings were: Rent, Apartments, DF/CDMX, Neighbourhood. The information collected was the average price.

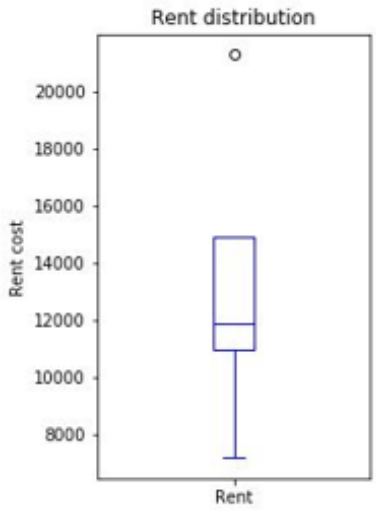
A csv file was developed, read, and transformed into a pandas dataframe containing the filtered list of neighbourhoods, their municipality, latitude, longitude, and average monthly rent.

3.3. Exploratory data analysis

Based on the dataframe created, the basic statistics were computed by using the describe method. The statistics were complemented with a boxplot, built by using the Matplotlib library to analyse the data visually.

	Rent
count	5.000000
mean	13245.046871
std	5274.817103
min	7182.519231
25%	10953.774648
50%	11895.841270
75%	14903.111111
max	21289.988095

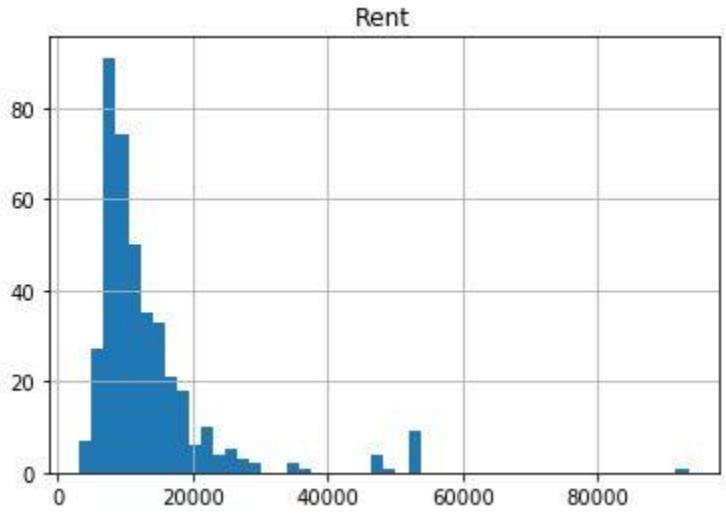
Table 4: Summary statistics



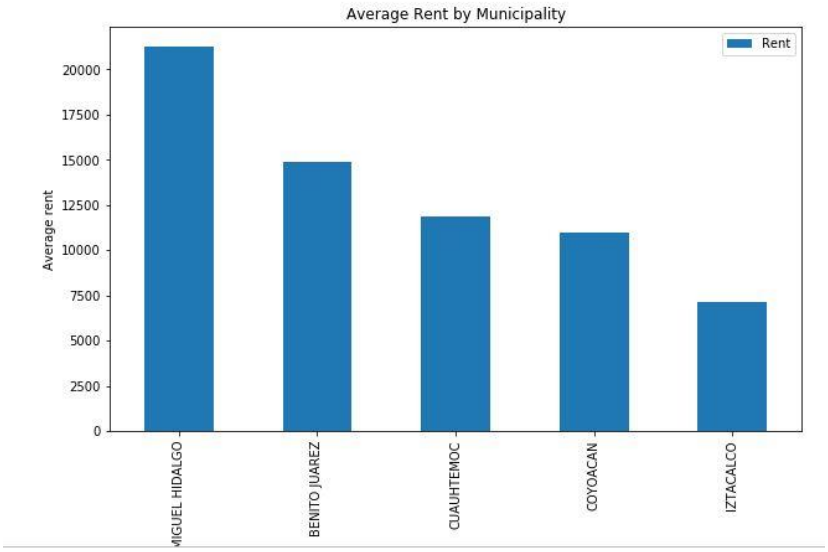
Graph 1: Rent boxplot

Analysing the summary statistics table and the boxplot, it can be noticed that the rent median is around $12,000 MNX (Mexican pesos), whereas the interquartile range goes from approximately $11,000 MXN to nearly $15,000 MXN. Lastly, in the summary statistics, it can be noted that the maximum rent to pay is a bit above $21,000 MXN, while the average is around $13,000 MXN.

Going deeper into the analysis, the Matplotlib histogram shown below was also created. Here it is noticed that the average rent variable has a right-skewed distribution, where the mean is slightly higher than the median due to the few large values that drive the mean upward.

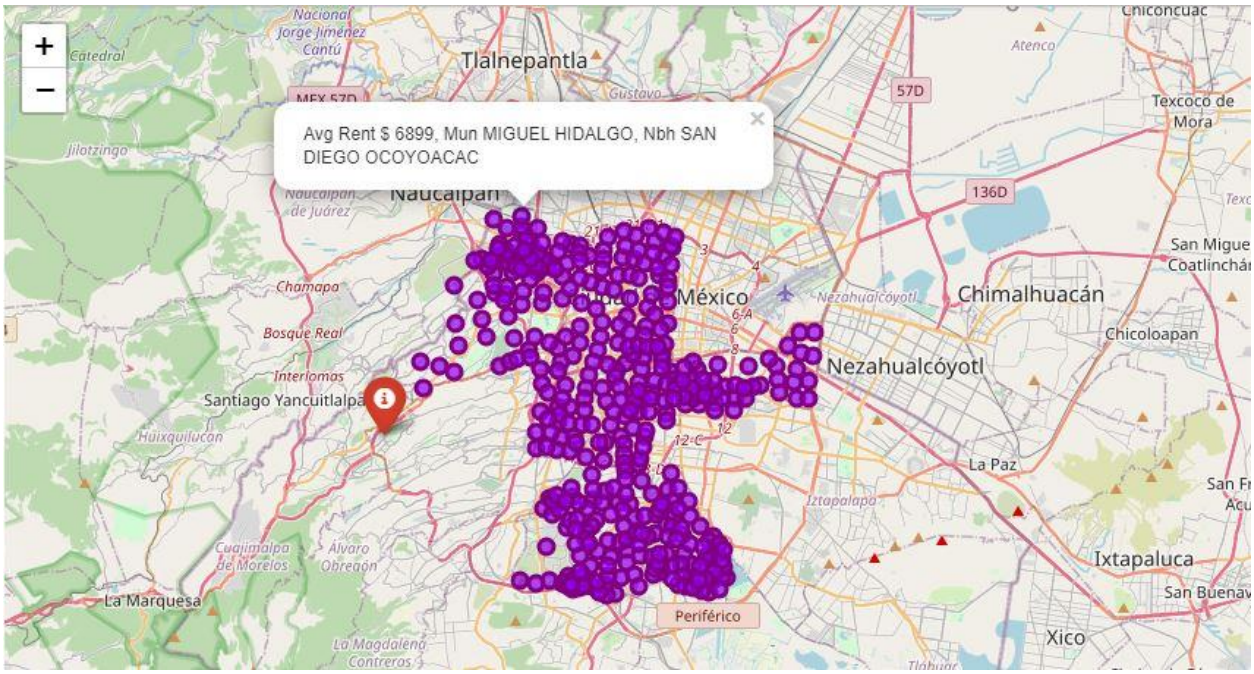


The bar graph below also built thanks to the Matplotlib, compares the five municipalities considered in this exercise. The graph clearly shows that the Miguel Hidalgo municipality has an average rent that surpasses the $15,000 budget. In contrast, the Iztacalco municipality is under the budget on average. Based on these findings, it can be inferred that the three remaining municipalities, which are Benito Juárez, Cuauhtémoc and Coyoacán, have more probabilities to have neighbourhoods within the budget range.



Graph 2: Average rent by municipality

Considering that the office in Mexico City, where our simulated user is going to be relocated was defined as "IBM, Mexico City", its coordinates, as well as the Mexico City coordinates, were retrieved using the Nominatim service. Then the first map could be built displaying the office location, the point in red with the information sign, along with all neighbourhoods of the five municipalities previously selected. (Map 1)



Map 1: Office location and all neighbourhoods of the five municipalities selected.

After visualising the large number of neighbourhoods, 404 to be precise, portrayed as purple dots on Map 1, it became clear that some additional variables were required to achieve the goal of ending up with a simple list of the best ten neighbourhoods to recommend.

3.4. Data preparation

The first variable that helped to reduce the list of neighbourhoods was the rent range previously calculated. After applying the filter, the list was reduced to exactly 100 neighbourhoods.

In order to get the next piece of information required to narrow down the list to the final ten neighbourhoods to recommend to the user, the Foursquare API was called selecting the explore endpoint. This call retrieved a list of all venues per each neighbourhood in a radius of 500 meters. It's relevant to inform that a limit of a hundred venues per neighbourhood was established. Also, a function was used to get the venue category, and another function to normalise the json file into a pandas dataframe.

The next step carried out was to apply the pandas function get_dummies, this function performs a technique known as one-hot encoding which converts a categorical variable, in our case the venue category, into numerical values of 0 or 1. This process paved the way to apply to the recommender algorithm to be detailed in the next section.

3.5. Recommender system

There are basically two types of recommender systems, content-based and collaborative filtering. For this scenario, the content-based system was selected with the aim of getting a list of the ten best neighbourhoods to recommend to the user considering the venue categories that the user likes the most.

Two data inputs are required to build the system: 1) Input user ratings, and 2) A matrix containing the things to be recommended to the user.

In our scenario, the input user rating was stored in a dataframe with the values previously shown on Table 1, where venue categories were ranked and rated by the user.

User's preferred venue categories		
Ranking	Venue Category	Rating
1	Coffee Shop	10
2	Japanese Restaurant	8
3	Gym / Fitness Center	6
4	Mexican Restaurant	4
5	Park	2

Table 1: User's venue categories preferences

On the other hand, the dataframe resulted after applying the one-hot encoding, was filtered by the venue categories selected by the user, then summed by venue category and transposed to be used as the second data input required. For our scenario, the outcome was a matrix containing the total number of venues for each combination of neighbourhood and venue category. A preview of this matrix is shown on Table 5 below.

	Venue Category	ACTIPAN	ALGARIN	ALTILLO COND ALTILLO UNIVERSIDAD	ANAHUAC DOS LAGOS	ANAHUAC II	ANAHUAC LAGO NORTE	ANAHUAC LAGO SUR	ANAHUAC MARIANO ESCOBEDO	ATLANTIDA	...
0	Coffee Shop	8	1	1	1	5	1	0	3	1	...
1	Gym / Fitness Center	4	1	0	1	1	0	0	1	2	...
2	Japanese Restaurant	1	0	1	0	1	0	0	1	0	...
3	Mexican Restaurant	4	5	0	5	15	3	2	4	9	...
4	Park	0	0	1	0	0	0	0	0	0	...

Table 5: Neighbourhoods and venue categories matrix

4. Results

4.1. Top ten neighbourhoods

Once having the two pieces required, the neighbourhoods and venue categories matrix was multiplied by the user's ratings to have a weighted version of the previous matrix, which then was summed and filtered by taking the first ten records to obtain the top ten neighbourhoods finally.

	Overall Score
ACTIPAN	128.0
VILLA COYOACAN	124.0
ANAHUAC II	124.0
CENTRO VIII	122.0
NARVARTE V	118.0
NARVARTE II	104.0
SANTA CATARINA BARR	102.0
PIEDAD NARVARTE	102.0
MERCED GOMEZ	102.0
INTEGRACION LATINOAMERICANA U HAB	96.0

Table 6: Top ten neighbourhoods

This final list of ten neighbourhoods can be interpreted as the list of neighbourhoods that have the largest quantity of venues that belong to the categories that the user likes the most. Notwithstanding that these neighbourhoods are also within the user's budget.

To know precisely the number of venues per venue category and neighbourhood, the dataframe shown below as Table 7 was built. On that dataframe, it could be appreciated that the top one neighbourhood Actipan, has eight venues categorised as Coffee Shops, four different Gym / Fitness Centers, four Mexican Restaurant options, and even one Japanese Restaurant. These numbers were the cause of having the highest score, converting it to the most recommendable neighbourhood for our user according to her venue preferences.

	Venue Category	ACTIPAN	VILLA COYOACAN	CENTRO VIII	ANAHUAC II	NARVARTE V	NARVARTE II	SANTA CATARINA BARR	PIEDAD NARVARTE	MERCED GOMEZ	INTEGRACION LATINOAMERICANA U HAB
0	Coffee Shop	8	8	4	5	6	7	7	7	6	7
1	Gym / Fitness Center	4	0	1	1	1	1	1	2	3	0
2	Japanese Restaurant	1	1	1	1	1	0	0	0	0	0
3	Mexican Restaurant	4	9	17	14	11	7	6	5	5	6
4	Park	0	0	0	0	0	0	1	0	2	1

Table 7: Number of venues per neighbourhood and venue category

4.2. Most common venues per neighbourhood

The top ten neighbourhoods list was complemented with the description of the five most common venue categories. The procedure to get this additional information started taking the same dataframe that resulted after applying the one-hot encoding, it was grouped by neighbourhood, and the mean for each one was computed. The outcome of this step can be translated as finding the frequency of each venue category within the neighbourhood. After this, a function was applied to sort descending, and lastly, the top five venue categories were stored in a dataframe shown in Table 7 below.

Neighborhood	Municipality	Overall Score	Latitude	Longitude	Rent	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
ACTIPAN	BENITO JUAREZ	128	19.3699	-99.1782	17707	Coffee Shop	Cosmetics Shop	Gym / Fitness Center	Mexican Restaurant	Health & Beauty Service
VILLA COYOACAN	COYOACAN	124	19.3474	-99.1631	16023	Mexican Restaurant	Coffee Shop	Ice Cream Shop	Plaza	Café
CENTRO VIII	CUAUHTEMOC	122	19.4315	-99.1464	12100	Mexican Restaurant	Taco Place	Deli / Bodega	Bar	Coffee Shop
ANAHUAC II	MIGUEL HIDALGO	120	19.4412	-99.1775	17707	Mexican Restaurant	Taco Place	Coffee Shop	Gym	Restaurant
NARVARTE V	BENITO JUAREZ	118	19.3861	-99.1572	16086	Mexican Restaurant	Taco Place	Coffee Shop	Spa	Bakery
NARVARTE II	BENITO JUAREZ	104	19.3971	-99.1530	16086	Taco Place	Coffee Shop	Mexican Restaurant	Seafood Restaurant	Ice Cream Shop
SANTA CATARINA BARR	COYOACAN	102	19.3483	-99.1736	16232	Coffee Shop	Mexican Restaurant	Ice Cream Shop	Bakery	Shopping Mall
PIEDAD NARVARTE	BENITO JUAREZ	102	19.4027	-99.1562	15248	Clothing Store	Coffee Shop	Cosmetics Shop	Boutique	Mexican Restaurant
MERCED GOMEZ	BENITO JUAREZ	102	19.3656	-99.1899	15802	Taco Place	Coffee Shop	Mexican Restaurant	Ice Cream Shop	Bakery
INTEGRACION LATINOAMERICANA U HAB	COYOACAN	96	19.3383	-99.1793	12789	Pizza Place	Coffee Shop	Taco Place	Mexican Restaurant	Wings Joint

Table 8: Top ten neighbourhoods plus top five most common category venues

On Table 7 shown above, it can be noted for instance that, the neighbourhood Actipan besides having a large quantity of category venues liked by the user, it also has Cosmetic shops and Health & Beauty Service as its second and fifth most common venue category correspondingly.

4.3. Distance to the workplace

Mexico City is a place widely known for its dense traffic, due to this fact, the distance from home to the workplace is critical. Here is where the geolocation of each neighbourhood comes into play. The distance from all of the top ten neighbourhoods to the workplace, "IBM, Mexico City", was calculated by using the respective latitudes and longitudes, with them the Euclidian distance was computed by means of applying the Numpy linalg.norm. Although the linalg.norm strictly speaking returns the Frobenius

norm (Numpy 2020) and not the Euclidean norm, for our scenario they can be considered as equivalents since the outcome gives the square root of the sum of squares of the distances in each dimension or $E = \sqrt{\Sigma_i (x_i - y_i)^2}$. The calculated distance was added to the previous dataframe ending with the structure of Table 8 shown below.

Neighborhood	Municipality	Overall Score	Distance	Rent	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
ACTIPAN	BENITO JUAREZ	128	8.12	17707	Coffee Shop	Cosmetics Shop	Gym / Fitness Center	Mexican Restaurant	Health & Beauty Service
VILLA COYOACAN	COYOACAN	124	10.35	16023	Mexican Restaurant	Coffee Shop	Ice Cream Shop	Plaza	Café
CENTRO VIII	CUAUHTEMOC	122	12.55	12100	Mexican Restaurant	Taco Place	Deli / Bodega	Bar	Coffee Shop
ANAHUAC II	MIGUEL HIDALGO	120	10.6	17707	Mexican Restaurant	Taco Place	Coffee Shop	Gym	Restaurant
NARVARTE V	BENITO JUAREZ	118	10.28	16086	Mexican Restaurant	Taco Place	Coffee Shop	Spa	Bakery
NARVARTE II	BENITO JUAREZ	104	11.47	16086	Taco Place	Coffee Shop	Mexican Restaurant	Seafood Restaurant	Ice Cream Shop
SANTA CATARINA BARR	COYOACAN	102	9.38	16232	Coffee Shop	Mexican Restaurant	Ice Cream Shop	Bakery	Shopping Mall
PIEDAD NARVARTE	BENITO JUAREZ	102	10.5	15248	Clothing Store	Coffee Shop	Cosmetics Shop	Boutique	Mexican Restaurant
MERCED GOMEZ	BENITO JUAREZ	102	7.12	15802	Taco Place	Coffee Shop	Mexican Restaurant	Ice Cream Shop	Bakery
INTEGRACION LATINOAMERICANA U HAB	COYOACAN	96	8.72	12789	Pizza Place	Coffee Shop	Taco Place	Mexican Restaurant	Wings Joint

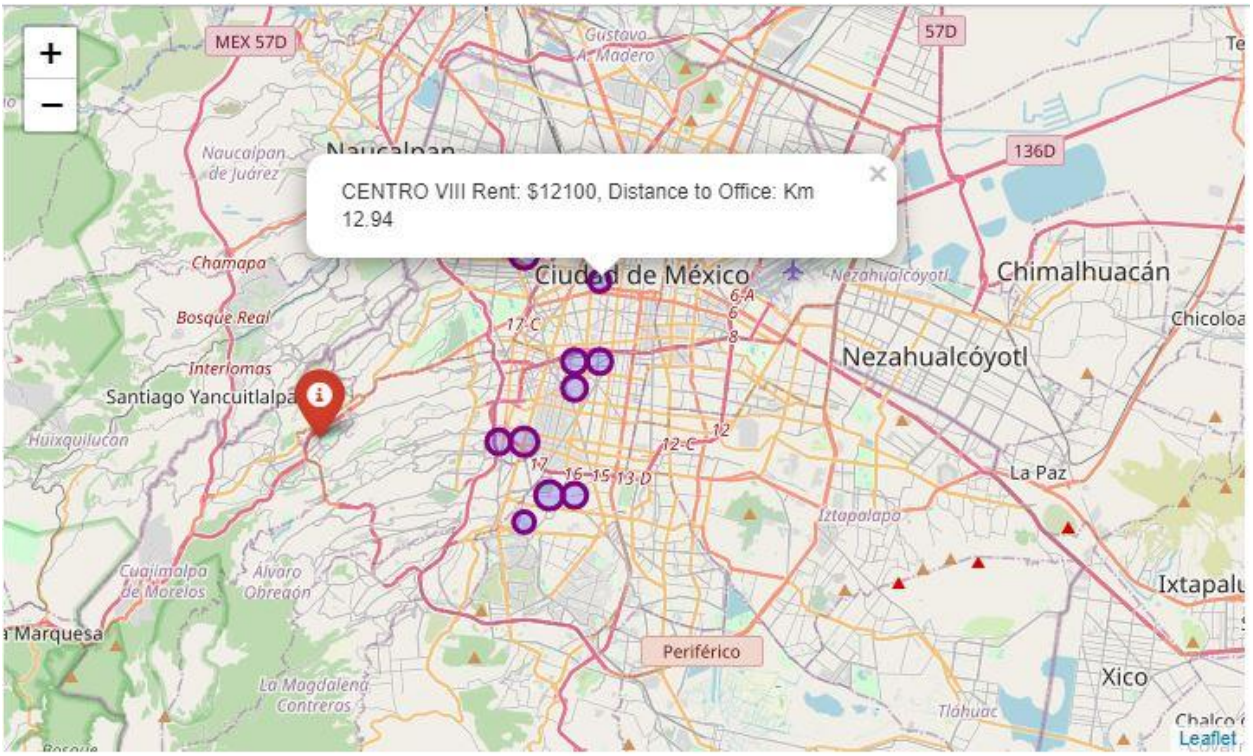
Table 8: Final dataframe

The ultimate purpose of the final dataframe shown above was to become the first product to be given to the user since it contains all relevant information collected through the process classified as relevant due to it helps the user to make a more informed decision.

1.1. Maps

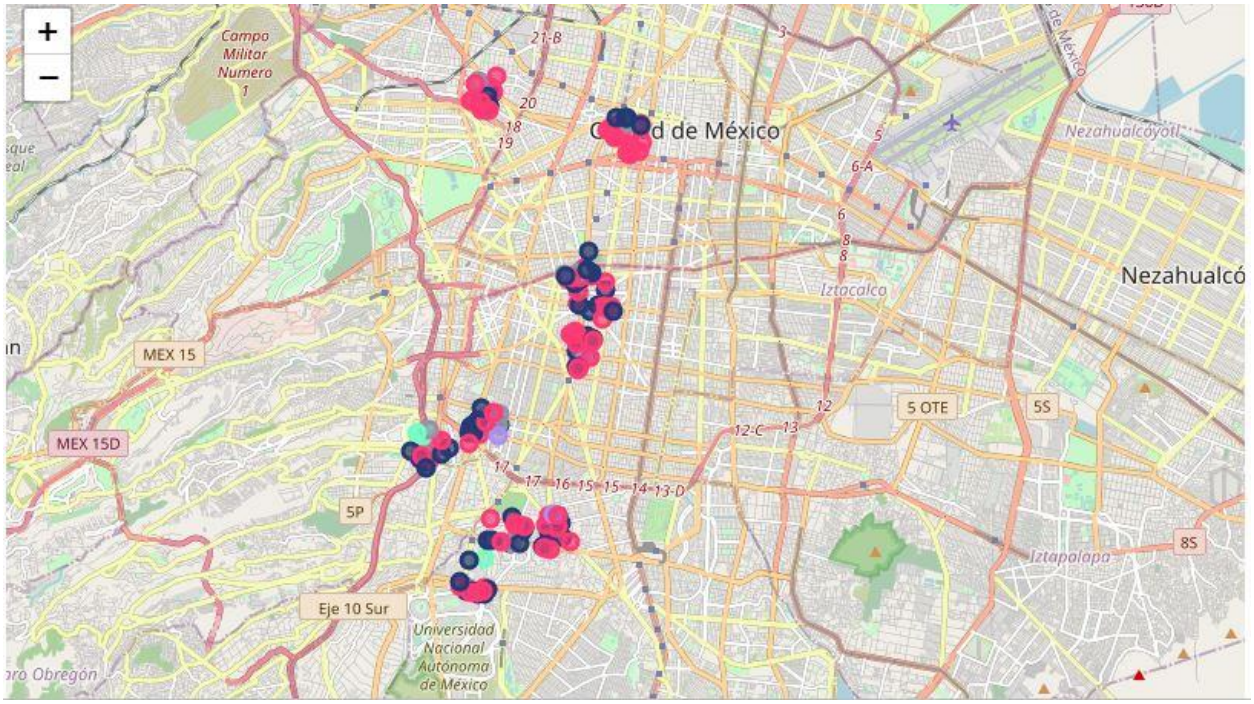
To complement the final dataframe, a couple of maps were also created and added as products to be given to the user.

The first map shown below as Map 1, illustrates where the office is located as well as each of the top 10 neighbourhoods recommended. The size of circles indicates the average rent cost, by clicking on a neighbourhood, a popup appears with the name of the neighbourhood, average rent cost, as well as the distance to the office.

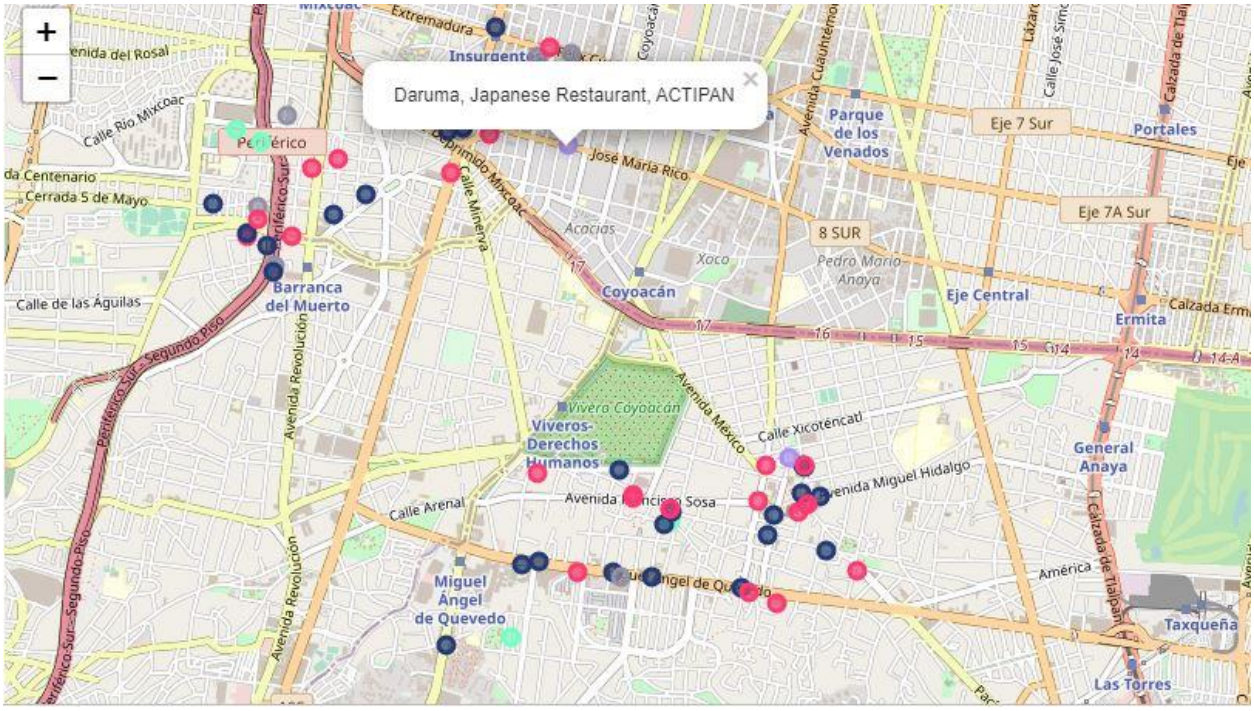


Map 2: Office location and top ten neighbourhoods

Last but not least, the second map shown below as Map 3, displays all venues that belong to the venue categories preferred by the user for each of the top ten neighbourhoods. To facilitate the read, venue categories were colour coded. Map 4 shows a zoomed-in version of the same map where colour differences can be better appreciated as well as the popup information which includes the name of the venue and its category.



Map 3: All preferred venues by each of the top ten neighbourhoods



Map 4: Map 3 zoomed in

5. Discussion

The results presented demonstrate once again that the more data, the more personalised user experience. In the specific case presented in this article, the venue categories preferred by the user were taken as the cornerstone to build a recommender system which produced a list of ten neighbourhoods that were the closest to user's preferences.

In addition to the user's preferences in terms of venue categories, the model considered the user's budget and a selection of municipalities with the highest HDI as filters. These filters helped to delimit the list of possible neighbourhoods to recommend, reducing the amount of data to handle. Hence, improving performance.

Distance from workplace to each of the top ten neighbourhoods recommended was not considered as a filter; instead, it was included as part of the set of information given to the user to be considered in her final decision.

6. Conclusion

As it was mentioned at the beginning of this article, leaving your current home chasing a good job opportunity abroad, it is not an easy decision. Moreover, if it is considered that implies searching for a place to live. This searching task could turn the decision not only difficult but highly stressing.

Although it is true that even having the most advanced recommendation system, relocation will not cease to be complicated. Nevertheless, recommender systems with some degree of personalisation as it is the model presented could help to lighten the burden and contribute to making a more informed decision, or at least to make a decision more aligned with users' preferences.

The recommender system presented is a Content-based kind of system due to the model searched for neighbourhoods with similar characteristics to the user's preferences. Although the results were satisfactory, like almost anything in this life, there is room for future work and improvement. In the particular case of this model, ratings and ranking information could also be gathered from Foursquare. Having this, the user could have the chance to find neighbourhoods that not only have many Coffee Shops but identify the neighbourhoods where the best Coffee Shops are located.

The previous is just an example of how adding relevant variables could produce more personalised user experiences. For the moment, our hypothetical user is at least better armed with information to decide where to live in Mexico City. Whichever her decision is, the model helped her to ensure having a Coffee Shop nearby to have her morning caffeine dose, a park to continue jogging on Saturdays, and a Japanese Restaurant to have that spicy tuna roll as dinner every Wednesday night. However, here the model is completely incompetent to foresee if the Mexican definition of spicy matches the user's definition.

7. References

- Propiedades.com., 2020. Mean prices and real state statistics by zone [online]. México: Propiedades.com. Available from: https://propiedades.com/valores [Accessed 12 June 2020].
- CDMX., 2020. Mexico City Open data [online]. México: Mexico City Government. Available from: https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/export/ [Accessed 11 June 2020].
- UNDP, 2019. Municipal human development report 2010-2015. “Transforming Mexico from the local” [online]. México: United Nations Development Program. Available from: https://www.mx.undp.org/content/mexico/es/home/library/poverty/informe-de-desarrollo-humano-municipal-2010-2015--transformando-.html [Accessed 16 June 2020].
- Numpy., 2020. numpy.linalg.norm [online]. USA: The SciPy community. Available from: https://numpy.org/doc/stable/reference/generated/numpy.linalg.norm.html [Accessed 18 June 2020].