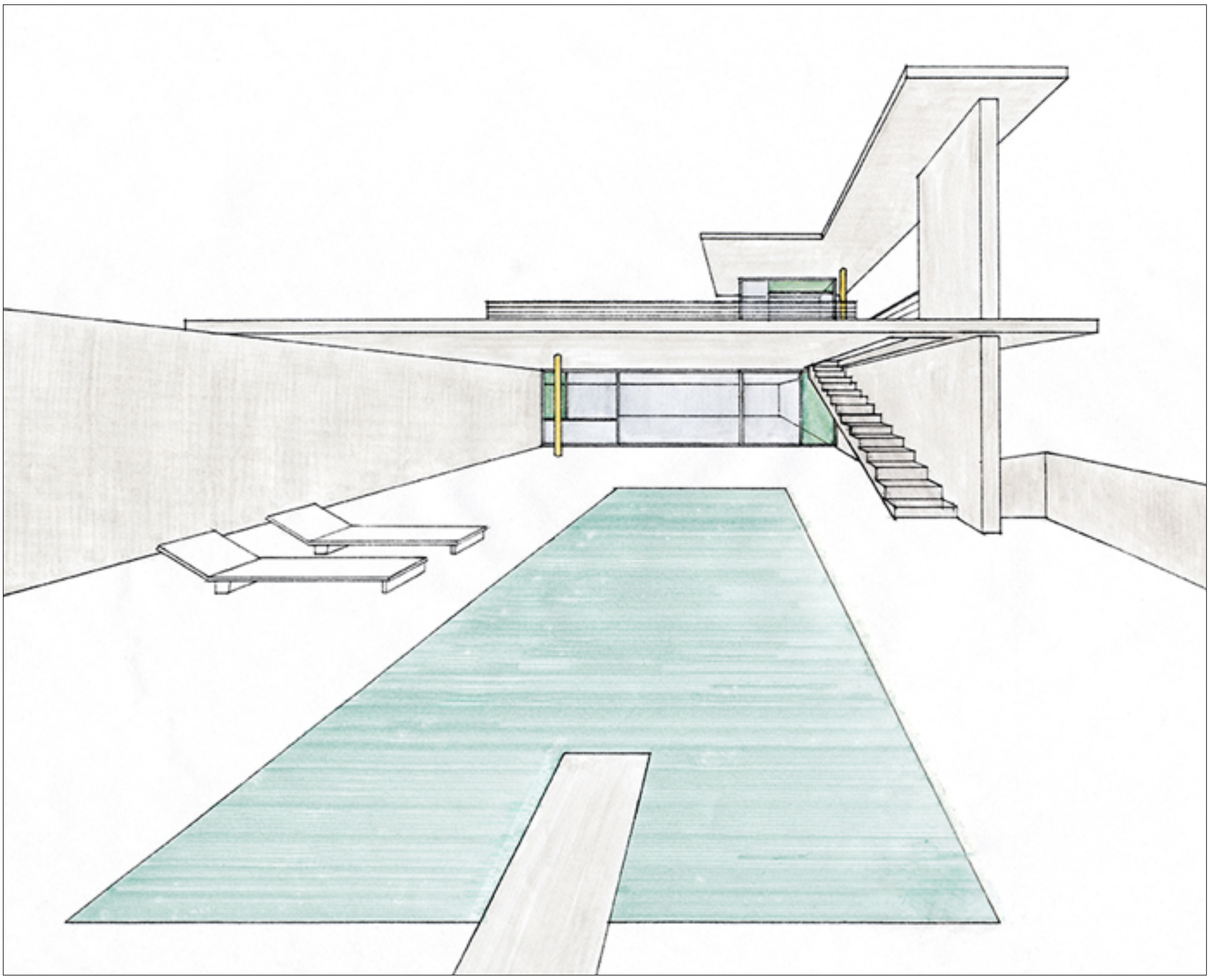
HOMEFINDER



Intelligent new home finder

Prepared for: coursa IBM data science professional certificate Capstone project

Prepared by: Caifang Cai

8 May 2020

INTRODUCTION

Objective

The main objective of this project is to provided an intelligent new home location finder for families relocated for a new job. This is based on the fact that the client family has already have a reference home and looking for a similar place to install the family. The criteria taken into account including:

1. Transport (bus, subway, local trains);
2. Closeness to commerce (bakery, supermarkets, etc.);
3. Nearby culture environment. Such as museums, galleries etc.
4. Public parks and sport centres.

The maximum distance should be limited in a radius of 12km to the new working place.

This major goal is to provide an intelligent tool to take into account of the above 4 major criteria to find a new home location similar to a reference home address. I give the top three recommendations locations for a new home location.

Stat of the art

Relocating home to unfamiliar region is indeed a painful task. It takes too much humain energy.

So far, when we need to relocate the home following a new job offre. The first step we need to do is to check on internets for days and days or to ask friends for all kinds of information. The second step is to process all these informations and make a 'not bad' choice barely based on humain juge. Both steps are not reliable which often leads to a 'not good' choice.

Interests

This project interests all people who need to relocate a family or to find a new home location in an unfamiliar region. The only condition to use is to have a reference home address. This reference home address can be your current home address that you are satisfied with or any address that you know well to be a good location for your home.

DATA ACQUISITION AND CLEANING

Table 1: office and reference home addresses used in homefinder

Office address	100 Avenue de Paris, 91344 Massy, France
Reference home address	102 Boulevard Richard Wallace, 92800 Puteaux

Dara source and demo example

We will use Foursquare data as our source data, an example in Paris region as our final result for demonstration. The reference home address and the new office address around which we are looking to install a family are given as follows in Table 1.

More specifically, we will use Foursquare explore end-point to explore the areas of concern.

Data cleaning

The raw data obtained from Foursquare is a json file which basically as following:

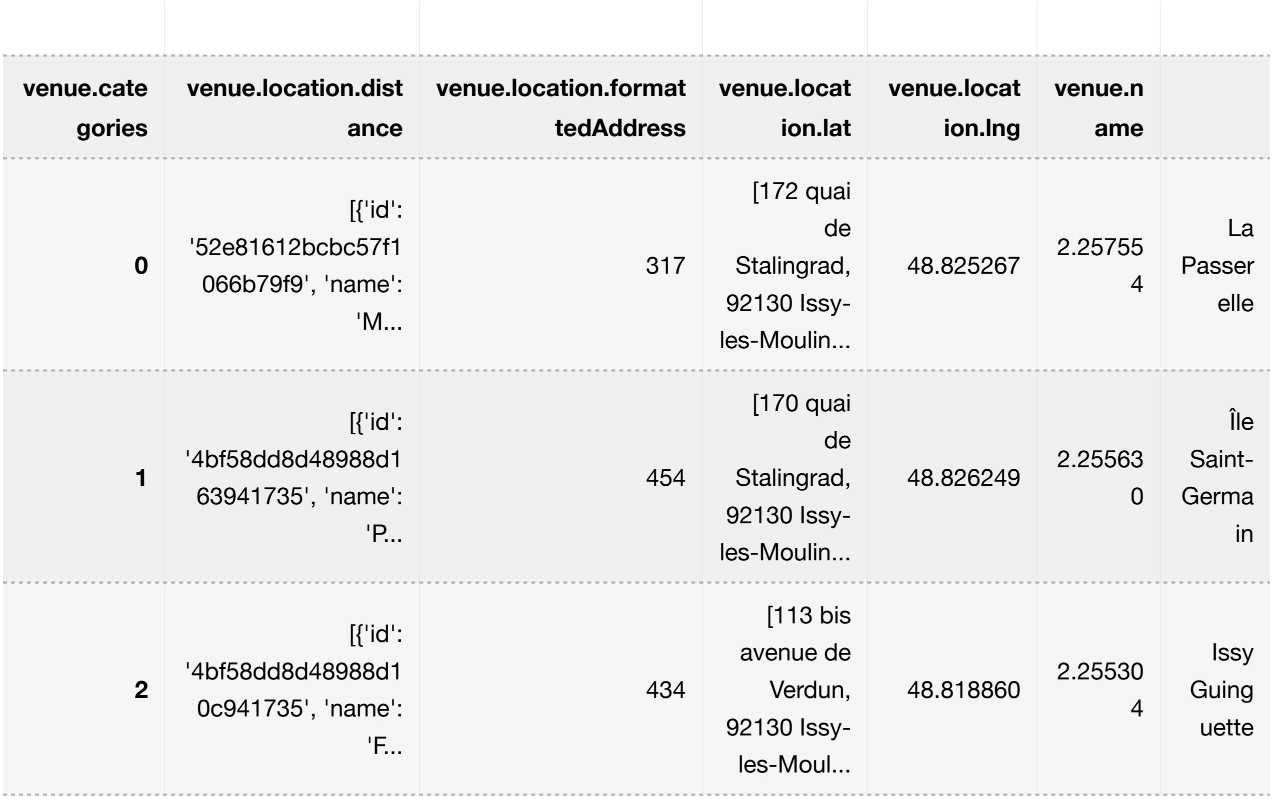
```
{'meta': {'code': 200, 'requestId': '5ef6780030567d545e80ec19'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Issy-les-Moulineaux',
  'headerFullLocation': 'Issy-les-Moulineaux, Paris',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 73,
  'suggestedBounds': {'ne': {'lat': 48.83141760900001,
    'lng': 2.271401296161224},
   'sw': {'lat': 48.81341759099999, 'lng': 2.244113103838776}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '541485c1498efe97bd3efa43',
       'name': 'La Passerelle',
       'location': {'address': '172 quai de Stalingrad',
        'lat': 48.82526662415111,
        'lng': 2.2575543895488503,
        'labeledLatLngs': [{'label': 'display',
          'lat': 48.82526662415111,
          'lng': 2.2575543895488503}],
        'distance': 317,
        'postalCode': '92130',
        'cc': 'FR',
        'city': 'Issy-les-Moulineaux',
        'state': 'Île-de-France',
              'country': 'France',}
```

The first we need to do is the extract the useful information related to our home relocating later.

Here is a list of extracted values used later for recommendations of new home location.

```
['venue.categories','venue.location.distance','venue.location.formattedAddress','venue.location.lat','venue.location.lng','venue.name']
```

After extraction, the data is not real for analyse yet. Here is a short example



venue.categories	venue.location.distance	venue.location.formattedAddress	venue.location.lat	venue.location.lng	venue.name	
0	[{'id': '52e81612bcbc57f1066b79f9', 'name': 'M...	317	[172 quai de Stalingrad, 92130 Issy-les-Moulin...	48.825267	2.257554	La Passerelle
1	[{'id': '4bf58dd8d48988d163941735', 'name': 'P...	454	[170 quai de Stalingrad, 92130 Issy-les-Moulin...	48.826249	2.255630	Île Saint-Germain
2	[{'id': '4bf58dd8d48988d10c941735', 'name': 'F...	434	[113 bis avenue de Verdun, 92130 Issy-les-Moul...	48.818860	2.255304	Issy Guinguette

3	[{'id': '4bf58dd8d48988d175941735', 'name': 'G...	374	[1-6 boulevard Garibaldi, 92130 Issy-les-Mouli...	48.823489	2.262604	MurMur

The contents are too long and can not be interpreted directly by humain. I then filtered the contents to get only the quantitive numbers and short, readable informations. Here gives an example after data cleaning.

category	distance	address	latitude	longitude	name	
0	Modern European Restaurant	317	172 quai de Stalingrad, 92130 Issy-les-Mouline...	48.825267	2.257554	La Passerelle
1	Park	454	170 quai de Stalingrad, 92130 Issy-les-Mouline...	48.826249	2.255630	Île Saint-Germain
2	French Restaurant	434	113 bis avenue de Verdun, 92130 Issy-les-Mouli...	48.818860	2.255304	Issy Guinguette
3	Gym / Fitness Center	374	1-6 boulevard Garibaldi, 92130 Issy-les-Moulin...	48.823489	2.262604	MurMur

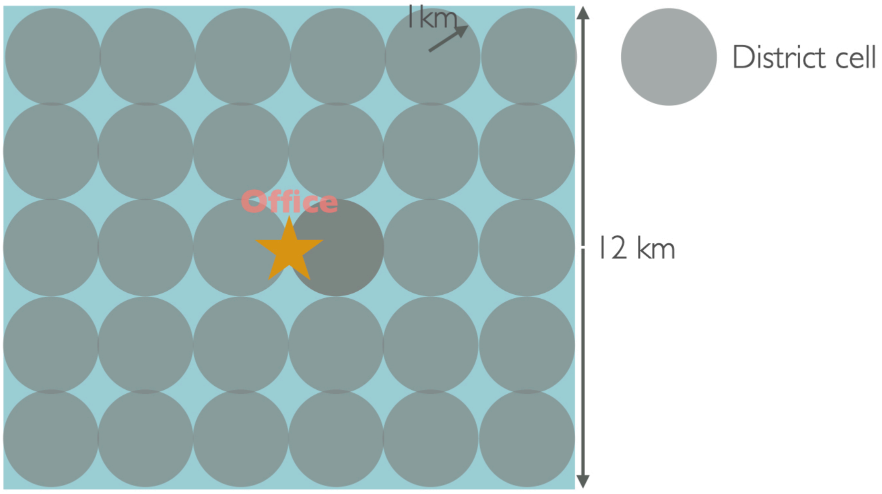
Feature preparation

I need to convert all these data into features that can be used for quantitative analysis. One important thing is to keep these features reflecting the four criteria that we discussed in the introduction. These four criteria are the rules for us to carry out further recommendations. I classify all the items into four categories commerce, transport, culture and environment (park and sport centres). I then count the numbers items and distances to office in each category. In each category, I use two values to describe it. One is the mean distance to home. The other is the density of numbers. In order to avoid influences of very small and very large numbers. I use a category value to describe the density of number. For example, for transport, the *num_transport* will have values of None, Few and Many. At the end, a concerned address will has the following shape of the feature vector.

dis_commerce	num_commerce	dis_culture	num_culture	dis_transport	num_transport	dis_environement	num_environement	
0.67481	high	0.659	median	0.7702	many	0.5485	low	

NEW HOME RECOMMENDATION METHOD

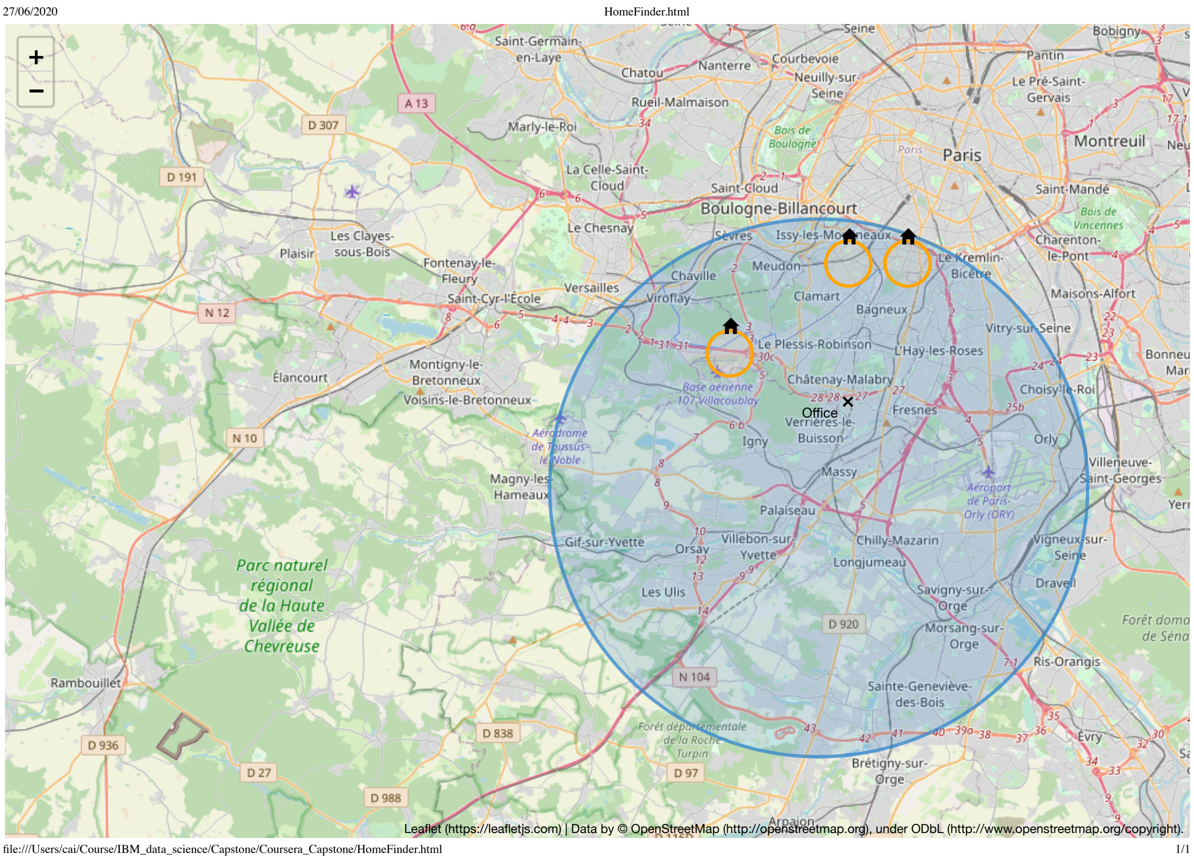
Grid mesh of searching region



I configured that only in a radius of 12 km around the office is considered as possible new home locations. In order to refine the searching region and perform Foursquare search. The searching region is a square 12x12 km with the office located at the center as shown above. The district cell is aligned inside of the searching region in a circle of radius 1 km. There are 36 district cells in total. The final propositions for new home location will be top three out of the 36 district cells.

In data preparation, we carried out data preparation separately for 36 district cells and the district cell of the reference home address. After the data preparation, we got a list of district cell features.

	dis_commerce	num_commerce	dis_culture	num_culture	dis_transport	num_transport	dis_environment	num_environment
0	0.67481	high	0.659	median	0.7702	many	0.5485	low
1	NaN	low	NaN	low	NaN	none	NaN	low



27/06/2020 HomeFinder.html

Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

file:///Users/cai/Course/IBM_data_science/Capstone/Coursera_Capstone/HomeFinder.html 1/1

2	0.66400	low	NaN	low	NaN	none	NaN	low
3	0.88100	low	NaN	low	NaN	none	NaN	low
4	NaN	low	0.612	low	NaN	none	0.7685	low

I then performed a one-hot encoding for all category data and replace the NaN data into 1 which means none of this kind shop is located in the district cell of the concern.

Recommendation criterion

I choose to use Euclidean distance to measure the difference of two distincts. Since we have a reference home address. The Euclidean distance between district of concern and the reference home district is a reasonable merit for new home recommendation.

Result

The interactive result is given at the link : https://github.com/CammieCo/Coursera_Capstone/blob/master/HomeFinder.html

The figure below gives a capture view. Noticed that I list the top three recommendations. They are given in orange circles. The searching region circle is just for visual view. The actual searching region is a square centred at office as we discussed above.

DISCUSSION & PERSPECTIVES

We see that my homeFinder makes it very simple to get a new home location recommendation in an unfamiliar region. The only required information is the new office address around which you would like to settle down your family and a reference home address. However, it is still limited by its data source.

- No school, hospital and real estate are used our current homeFinder. These informations are usually very important when we choose a place to relocate a family.
- Can not deal with recommendations with identical or very close scores. This is also the reason why we give the top three instead of top one. In future, this can be improved by introducing more customisable criteria.
- Must require a reference home address. This is usually not difficult to get but for those who look for a change environment. It might not always easy to get.