CAPSTONE PROJECT

REALTY INVESTMENT IN PARIS AREA (FRANCE)

by Dimitri BONDARENKO,

9 May 2020

INTRODUCTION

Paris is a worldwide known city which attract people from all over the world because of its historical sightseeings, French cuisine and fashion industry. Alongside with that Paris itself has an early mediaeval planification with dense building and very few areas for new constructions. All these factors contributed to a significant increase in real-estate market. However, more and more young people moving to Paris for career reason, are interesting to become property owners.

The aim of this project is to guide young families in their first investment project to become a realty owner in Paris suburb area. It is not an easy decision that often conducts to an engagement with bank for multiple years. With this project I am going to reveal several criteria to facilitate decision making.

DATA COLLECTION

First, before the data description and collection let's define the area of interest and make some assumptions. As we are focusing on a young family with an average income, the expensive city of Paris remains outside the scope of the analysis. Instead, multiple *communes* (i.e. towns) in Ile-de-France region can hide interesting investment cases. So, in this study we will cover the following districts (**www.kelcodepostal.fr**):

- Hauts-de-Seine with 36 communes;
- Seine-Saint-Denis with 40 communes;
- Val-de-Marne with 47 communes.

Moreover, in this study we will focus on **official real-estate market rates** only (**www.meilleursagents.com**). We will distinguish two realty categories which are **apartments** and **houses**. In addition, we will consider different evaluation metrics on communes given by inhabitants especially on **Environment, Transport, Security, Healthcare, Sports & leisures, Culture, Education, Commerce** and **Quality of life** (**www.ville-ideale.fr**). Another kind of data that we will use for visualization reason is GPS coordinates (**Longitude, Latitude**) of each commune (**www.sql.sh**) and **GEO map of France** (**www.data.gouv.fr**). It should be outlined that in France each commune has its own unique **INSEE code** which we will use to simplify all manipulations with the data.

We note that all mentioned data is open sourced. On the contrary, French laws in realty sector (ex. *Carrez* law, *Pinel* law, etc.) as well as infrastructure development plans (ex. opening new metro station) remain outside the presented analysis.

Now let's take a look at the data itself and transform it to *Pandas* dataframes. Here we omit the details on webscrapping of the selected data from mentioned websites as it stands outside the project.

In [2]:
```
commune_gps = pd.read_csv("commune_gps.csv",delimiter=";")
commune_gps.head()
```

Out[2]:

	Code INSEE	Commune	Postal code	Longitude	Latitude	Surface
0	92051	NEUILLY-SUR-SEINE	92200	2.26667	48.8833	3.73
1	92020	CHATILLON	92320	2.28333	48.8000	2.92
2	92009	BOIS-COLOMBES	92270	2.26667	48.9167	1.92
3	92062	PUTEAUX	92800	2.23333	48.8667	3.19
4	92023	CLAMART	92140	2.26667	48.8000	8.77

Where column "Surface" stands for the squared area of each commune.

In [3]:
```
commune_metrics = pd.read_csv("commune_metrics.csv",delimiter=";")
commune_metrics.head()
```

Out[3]:

	Code INSEE	Environment	Transport	Security	Healthcare	Sport	Culture	Education	Commerce	Quality of life	Price_sq_m_apart	Price_sq_m_house
0	92002	7.63	7.30	6.97	8.10	8.04	7.67	7.89	7.38	7.97	4963	5400
1	92019	7.07	5.54	5.15	5.84	6.93	7.43	5.89	5.16	6.20	4390	5701
2	92060	8.23	5.72	8.46	7.74	7.81	7.43	7.12	7.33	8.12	5084	5878
3	92071	8.47	6.86	7.51	7.57	7.80	7.86	8.89	7.79	8.41	6131	7810
4	92014	6.88	8.55	7.88	7.67	7.12	6.81	7.97	7.43	8.33	5542	6653

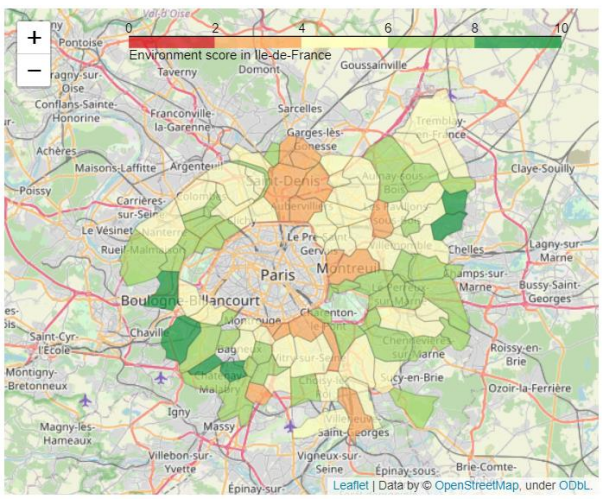
Where columns "Price_sq_m_apart" and "Price_sq_m_house" stand for average price per square meter in apartment and house respectively. Next, let's merge these two dataframes by column "Code INSEE":

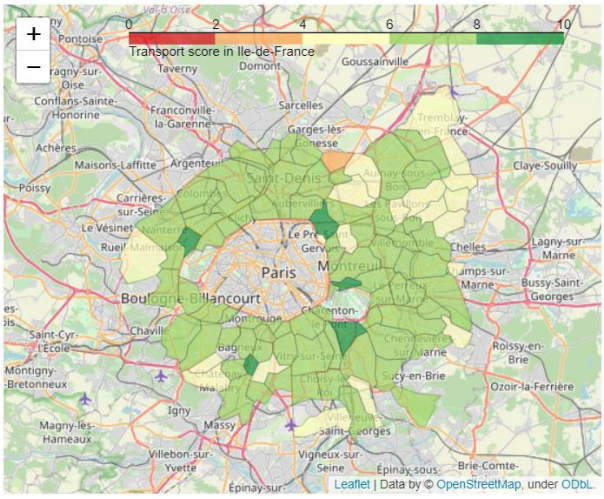
In [4]:
```
commune_data = pd.merge(commune_gps, commune_metrics, on=['Code INSEE'])
commune_data.head()
```

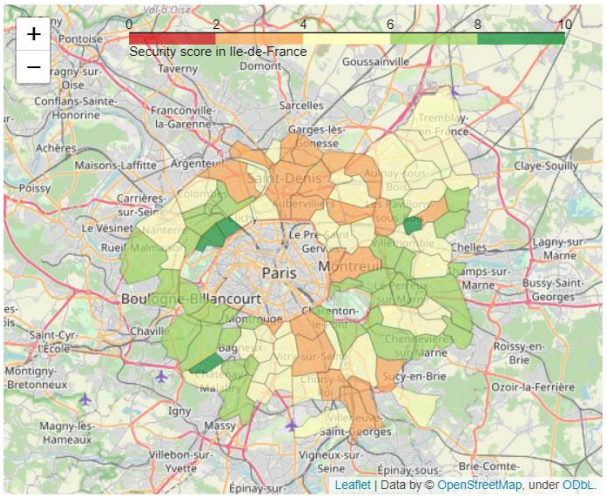
Out[4]:

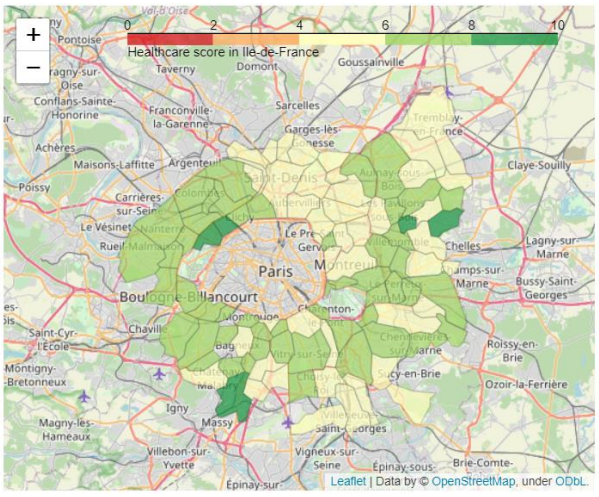
	Code INSEE	Commune	Postal code	Longitude	Latitude	Surface	Environment	Transport	Security	Healthcare	Sport	Culture	Education	Commerce	Quality of life
0	92051	NEUILLY-SUR-SEINE	92200	2.26667	48.8833	3.73	7.60	7.80	8.44	8.00	6.80	6.73	7.53	6.93	8.04
1	92020	CHATILLON	92320	2.28333	48.8000	2.92	6.63	7.52	6.63	7.03	7.07	6.55	6.52	7.03	6.97
2	92009	BOIS-COLOMBES	92270	2.26667	48.9167	1.92	6.89	6.81	7.28	6.80	7.19	7.09	7.38	7.15	7.74
3	92062	PUTEAUX	92800	2.23333	48.8667	3.19	6.67	8.28	7.99	7.29	8.59	7.85	6.12	7.56	7.65
4	92023	CLAMART	92140	2.26667	48.8000	8.77	7.12	6.38	6.52	7.56	7.19	6.80	7.03	6.70	7.41

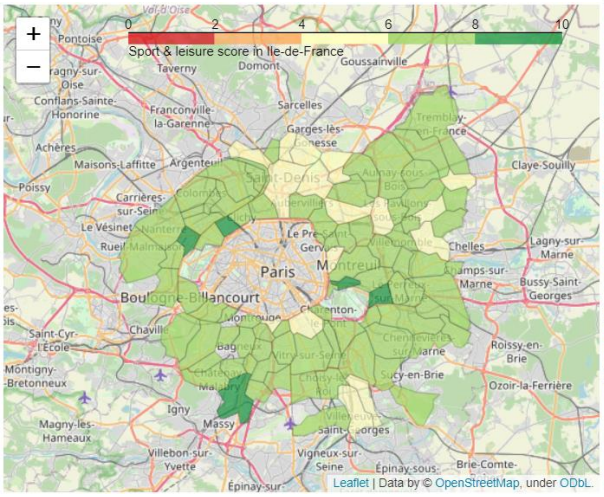
The created dataframe **commune_data** will be used as initial data in the following study on realty investment. It contains the information on metrics, prices and GPS coordinates for 101 selected communes around Paris city. To "feel" the data, it would be useful to visualize all extracted metrics on the GEO map centred on Paris city.

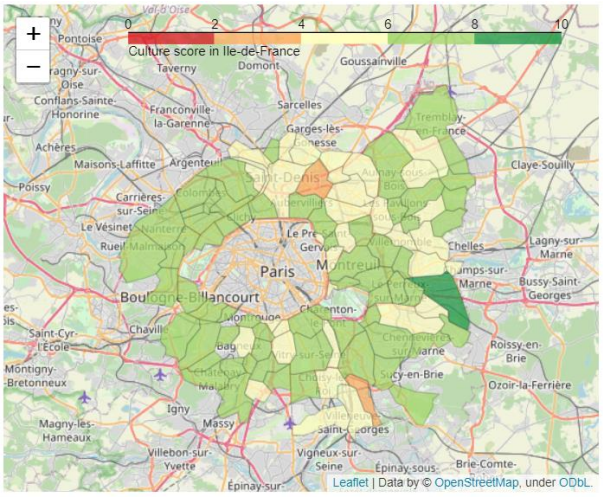


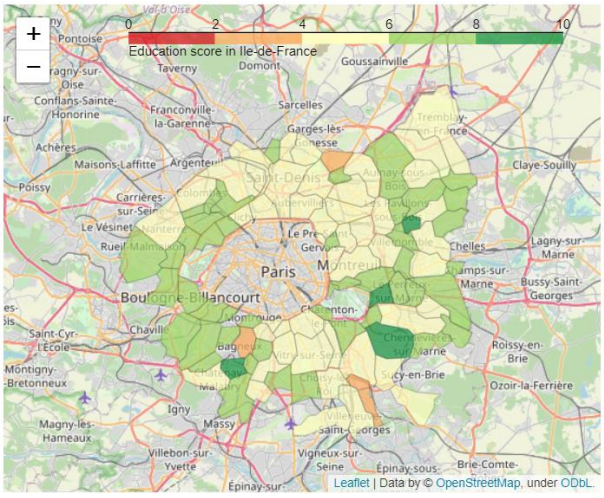


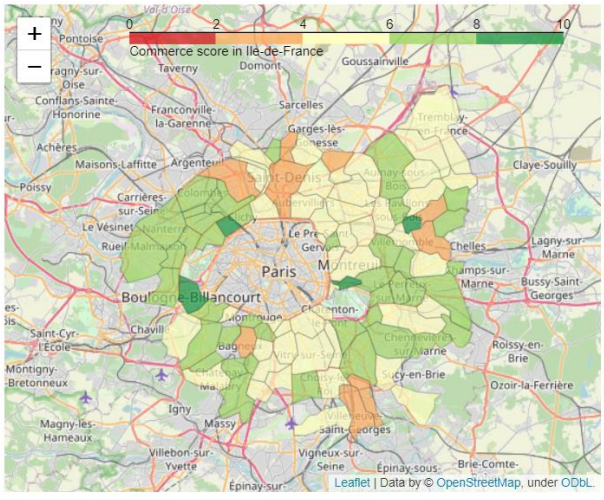

Security score in Ile-de-France
0 2 4 6 8 10
Leaflet | Data by © OpenStreetMap, under ODbL.

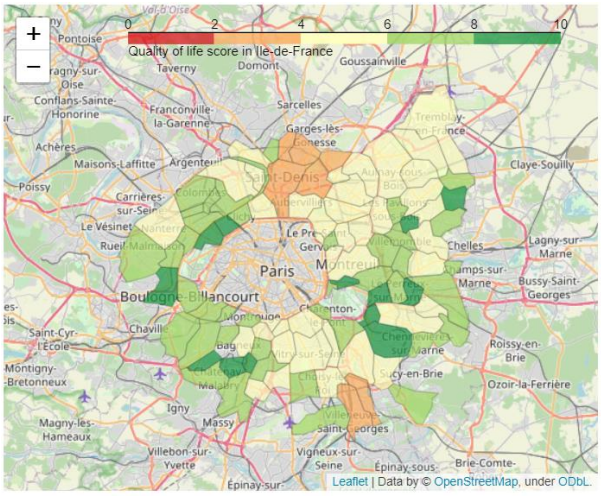

Healthcare score in Ile-de-France
0 2 4 6 8 10
Leaflet | Data by © OpenStreetMap, under ODbL.


Sport & leisure score in Ile-de-France
0 2 4 6 8 10
Leaflet | Data by © OpenStreetMap, under ODbL.

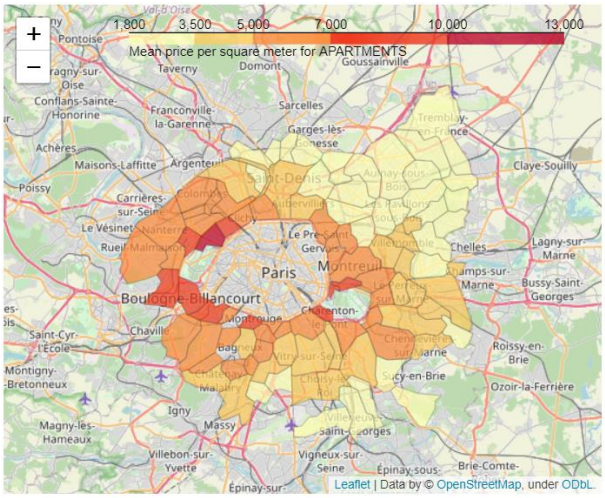

Culture score in Ile-de-France
0 2 4 6 8 10
Leaflet | Data by © OpenStreetMap, under ODbL.

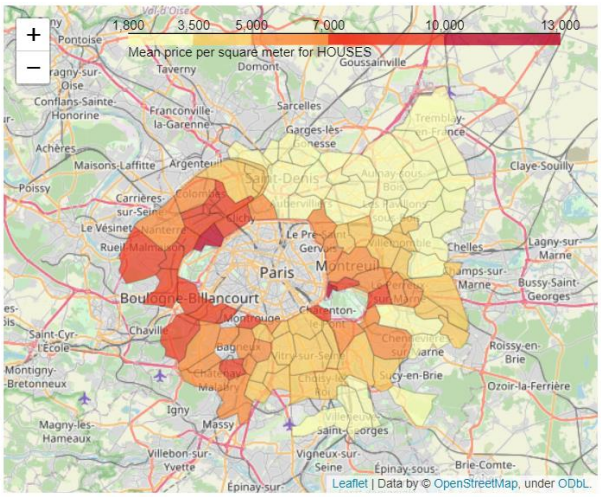

Education score in Ile-de-France
0 2 4 6 8 10
Leaflet | Data by © OpenStreetMap, under ODbL.


Commerce score in Ile-de-France
0 2 4 6 8 10
Leaflet | Data by © OpenStreetMap, under ODbL.



Now, let's take a closer look on a distribution of realty rates for apartments and houses. Attention, another pattern of colours is used for two next figures:





Looking at all these choropleth maps its not so obvious to make an optimum choice of commune for investment in both apartment and house. Nevertheless, some observations could be outlined for the given dataset:

- Realty in west side suburb areas of Paris is the most expensive, as well as in communes locating close to the city of Paris. The effect is more outlined in house segment. On the other hand, most afforable housing can be found in north and north-east communes;
- Distribition of realty rates per commune does not correlated well with different rating scores given by dwellers. For exemple, the score for Education, Culture and Healthcare attributed for several noth-east communes is similar to expensive west side communes. Moreover, the distribution of the Environment score introduces some uncertainty to decision making in whith commune to invest.

In the next sections we will apply machine learning techniques to classify communes and to enhance the decision making.

METHODOLOGY

In order to achieve the designated goal, following data driven algorithm is proposed:

- Computing total average rating for each commune based on scrapped user rating data;
- Clustering communes with k-means method;
- Selecting cluster with highest rating values;
- Aligning selected communes with price rates in apartment and house categories;
- Narrowing the list of selected communes by maximizing the living space area for a given family budget;
- Enhancing the list of communes attractive for investment with some details on most visited local venues. Such information could affect the final decision as it reveal local habits.

To illustrate the described algorithm, some use case will be considered.

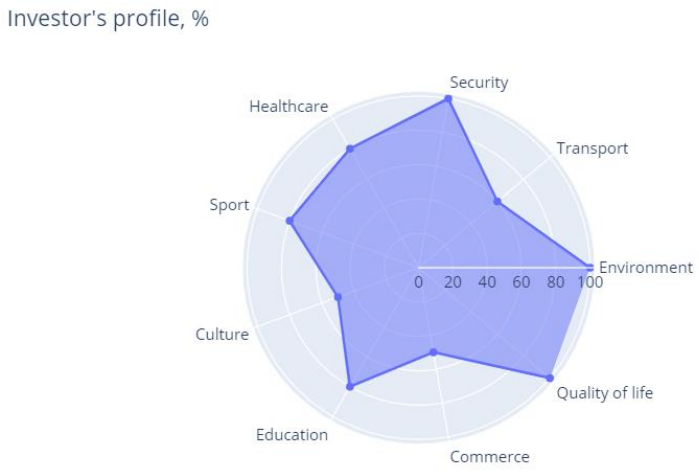
ANALYSIS

Let's first imagine and describe our investor. His demands on Environment, Healthcare, Education, Sport and so on form the investor profile. It can be visualized with a radar chart as follows.

Out[113]:

	Environment	Transport	Security	Healthcare	Sport	Culture	Education	Commerce	Quality of life
0	1.0	0.6	1.0	0.8	0.8	0.5	0.8	0.5	1.0

It can be illustrated with the following radar chart:



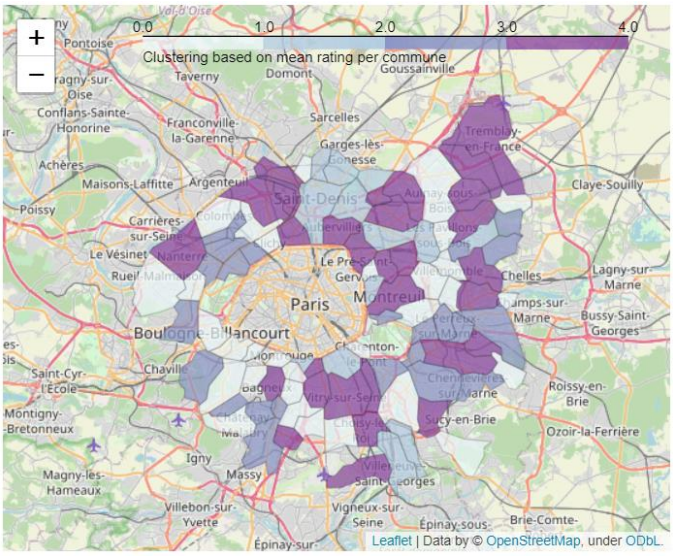
Next, we compute a weighted average rating for each commune based on scrapped statistical values and correlated with the given investor profile. Such weighted mean rating will be used hereafter as indicator of attractiveness for selected communes. The arithmetic average rating is also calculated for comparison. Then, k-means clustering is applied to group communes with similar weighted average rating.

In [237]:
```
commune_data['Arithmetic mean rating']=commune_data.iloc[:, 6:15].sum(axis=1)/9
w = np.array(list(weights.loc[0,labels].values))
commune_data['Weighted mean rating']=commune_data.iloc[:, 6:15].dot(w)/sum(w)
commune_data.head()
```

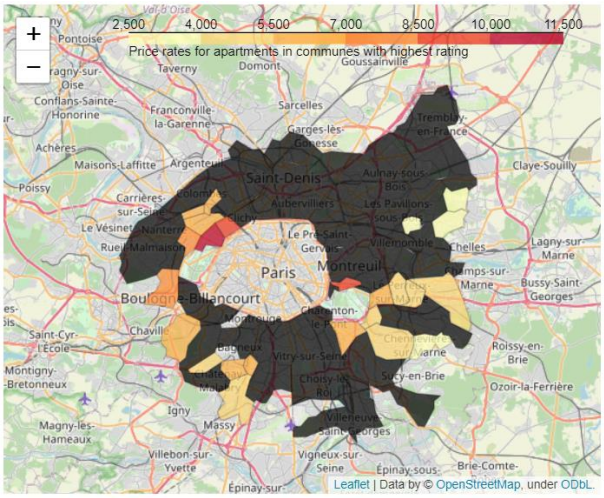
Out[237]:

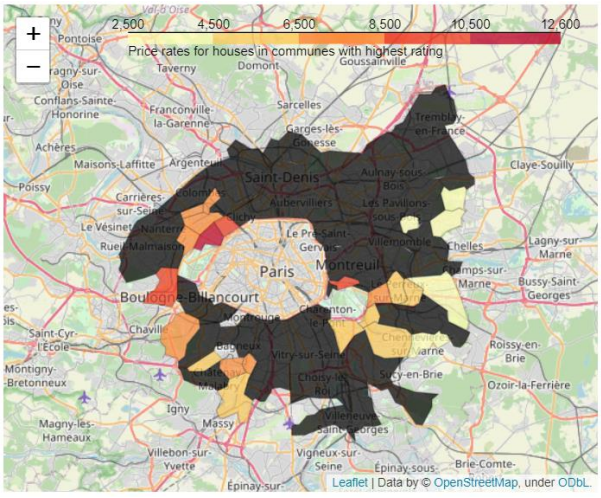
nment	Transport	Security	Healthcare	Sport	Culture	Education	Commerce	Quality of life	Price_sq_m_apart	Price_sq_m_house	Arithmetic mean rating	Weighted mean rating	Cluster labels
7.60	7.80	8.44	8.00	6.80	6.73	7.53	6.93	8.04	11230	12545	7.541111	7.636286	2
6.63	7.52	6.63	7.03	7.07	6.55	6.52	7.03	6.97	5774	6472	6.883333	6.861143	0
6.89	6.81	7.28	6.80	7.19	7.09	7.38	7.15	7.74	6233	7675	7.147778	7.173143	2
6.67	8.28	7.99	7.29	8.59	7.85	6.12	7.56	7.65	7496	8302	7.555556	7.511857	2
7.12	6.38	6.52	7.56	7.19	6.80	7.03	6.70	7.41	5638	6931	6.967778	7.007429	0

Result of k-means clustering with 4 clusters:



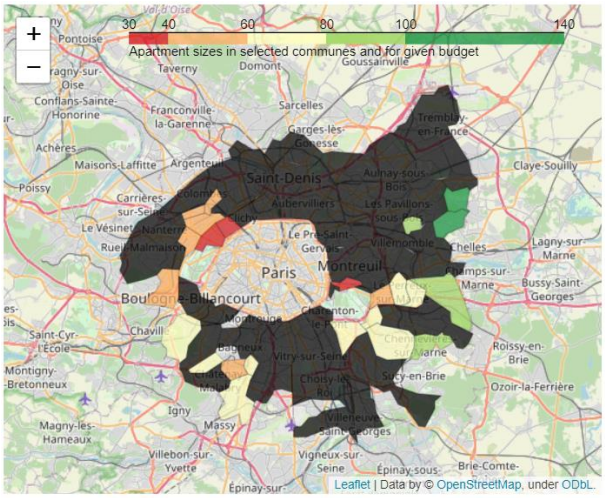
The 3rd cluster corresponds to highest values in weighted average rating that was computed for our investor. There is a list with 24 communes to choose from. Let us correlate the selected communes with the scraped realty average rates for apartment and house categories.





There is still large number of communes to make a proper decision on investment. We need some additional data from our investor which is his estimation on budget. Let's say he is ready to pay 350000 €. With this information we are going to fit our list of 24 selected communes with highest possible size of living area corresponded to the budget. Note, that for the sake of reality in the

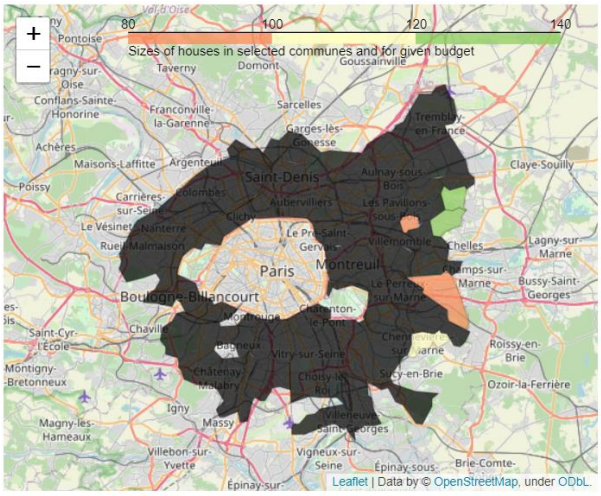
following computations the minimum available living area is limited with 30 square meters for apartments and 80 square meters for houses.



Let's summarize in a table the information on TOP 5 attractive communes for our investor in terms of their weighted mean rating and maximum theoretical living space area for APARTMENTS that corresponds to the budget of 350000€. The interested communes are coloured with green and light green on the figure above.

Out[307]:

	Code INSEE	Commune	Postal code	Arithmetic mean rating	Weighted mean rating	Mean_apart_square
53	93047	MONTFERMEIL	93370	7.261111	7.380429	133.130468
38	93015	COUBRON	93470	7.172222	7.338143	107.131925
94	94019	CHENNEVIERES-SUR-MARNE	94430	7.423333	7.477429	99.431818
40	93062	LE RAINCY	93340	7.776667	7.813857	88.006035
44	93051	NOISY-LE-GRAND	93160	7.474444	7.401000	86.185669



Let's summarize in a table the information on TOP 5 attractive communes for our investor in terms of their weighted mean rating and maximum theoretical living space area for HOUSES that corresponds to the budget of 350000€. Note, that we find same communes in the list as for apartments but in different order.

Out[306]:

	Code INSEE	Commune	Postal code	Arithmetic mean rating	Weighted mean rating	Mean_apart_square
53	93047	MONTFERMEIL	93370	7.261111	7.380429	133.130468
38	93015	COUBRON	93470	7.172222	7.338143	107.131925
94	94019	CHENNEVIERES-SUR-MARNE	94430	7.423333	7.477429	99.431818
40	93062	LE RAINCY	93340	7.776667	7.813857	88.006035
44	93051	NOISY-LE-GRAND	93160	7.474444	7.401000	86.185669

Now, let's go further and provide to our investor some details on popular venues per commune from above tables. Such information could influence the final decision on investment as it reflects local habits. For that we will use database from [Foursquare.com]

In [223]:
```
num_top_venues = 5

for hood in idf_grouped['Commune']:
        print("----"+hood+"----")
        temp = idf_grouped[idf_grouped['Commune'] == hood].T.reset_index()
        temp.columns = ['venue','freq']
        temp = temp.iloc[1:]
        temp['freq'] = temp['freq'].astype(float)
        temp = temp.round({'freq': 2})
        print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
        print('\n')
```

```
----CHENNEVIERES-SUR-MARNE----
              venue  freq
0               Gym  0.25
1  French Restaurant  0.25
2       Sports Club  0.25
3    Tennis Stadium  0.25
4  Food & Drink Shop  0.00


----COUBRON----
                venue  freq
0         Supermarket  0.33
1           Gastropub  0.33
2       Auto Workshop  0.33
3    Auto Dealership  0.00
4  Japanese Restaurant  0.00


----MONTFERMEIL----
                venue  freq
0                Park  0.50
1       Burger Joint  0.25
2        Supermarket  0.25
3    Auto Dealership  0.00
4  Health Food Store  0.00


----LE RAINCY----
                  venue  freq
0    Japanese Restaurant  0.17
1                Bistro  0.17
2       Sushi Restaurant  0.17
3  Furniture / Home Store  0.17
4     Light Rail Station  0.17


----NOISY-LE-GRAND----
                venue  freq
0    Auto Dealership  0.33
1        Supermarket  0.33
2  Food & Drink Shop  0.33
3  Health Food Store  0.00
4   Sushi Restaurant  0.00
```

Following observations could be outlined by inspecting results of analysis on popular venues:

- Commune CHENNEVIERES-SUR-MARNE could be attractive for investors interesting in different sport activities (note, that this correlates to user ratings scrapped from [www.ville-ideale.fr]);
- Car enthusiasts may turn toward communes COUBRON and NOISY-LE-GRAND;
- Commune MONTFERMEIL could be interesting for investors with children as there is a park that attracts locals;
- Finally, japaneese food lovers should focus on the commune LE RAINCY.

RESULTS AND DISCUSSION

Obtained results are individual and applicable for the given investor with defined budget and his own demands on different aspects of life. Starting from the list of 102 communes around Paris we selected five best communes to invest for apartment or house. Moreover, some additional information on popular venues is given for each selected commune.

Out[319]:

	Code INSEE	Commune	Postal code	Arithmetic mean rating	Weighted mean rating	Mean_apart_square	Mean_house_square	Good for
0	93047	MONTFERMEIL	93370	7.261111	7.380429	133.130468	138.999206	Families with children
1	93015	COUBRON	93470	7.172222	7.338143	107.131925	128.818550	Car enthusiasts
2	94019	CHENNEVIERES-SUR-MARNE	94430	7.423333	7.477429	99.431818	105.042017	Active and sport lovers
3	93062	LE RAINCY	93340	7.776667	7.813857	88.006035	85.616438	Japaneese food lovers
4	93051	NOISY-LE-GRAND	93160	7.474444	7.401000	86.185669	96.578366	Car enthusiasts

The presented algorithm could be enhanced by statistical analysis on social rating and realty price rates distributions in order to evaluate the accuracy and confidence. But some advanced skills in web scrapping is needed to conduct such kind of analysis.

CONCLUSION

Real estate investing is a challenge that a successful modern person will face sooner or later. With this project, we presented an algorithm for choosing a city for investment, based on social ratings, average market prices, as well as investor expectations based on criteria such as Environment, Transport, Security, Healthcare, Sports & leisure, Culture, Education, Commerce and Quality of life. The suburbs of Paris were considered in this study. But it can be easily applied to any other places.

ACKNOWLEDGEMENT

This work was carried out as part of the final project completing a series of nine courses entitled "IBM Data Science Professional Certificate" and mentored by IBM via **Coursera.org**.