

# Paris – City of Light

## Coursera - Capstone Project

Paul Preda

January 2020

### I. Introduction: Business Problem

Paris is a tremendous city of 2 million inhabitants, with a very rich offer of culture, history but also gastronomy and nature. Around, the "Ile de France" region is the biggest European region in terms of population and GDP. Real estate is quite expensive, and there are important discrepancies among the 20 Paris districts ("Arrondissements") and the cities which are very close to Paris.

#### I.1 Problem

In this project we will try to help the **Mayor of Paris** and his colleagues nearby Paris to adjust the **City policies** in order to encourage **social diversity** and to increase the **attractiveness of the neighborhoods**. We will try to bring them practical levers allowing to increase this attractivity and we assume – as hypothesis – that increasing the attractivity of each neighborhood will contribute to the global welfare of their electors.

#### I.2 Approach

Our project will focus on East Paris (10 arrondissements) and 14 surrounding cities close to East Paris. Data collection and interpretation will focus on two distinct parts:

- first, we will examine the **real estate** price per "Arrondissement" or city, based on the available public data of Real Estate transactions. We will calculate the average price per square meter per city.
- second, we will find the most **popular venues** and segment the neighborhoods according to the typology of their venues. Of course, Foursquerra will be used to collect the associated data.

Once we have the Real Estate prices associated to each Quartier and the Venues profile, we will try to find a **correlation** between the two types of data and describe what is the **venue segment associated with a "expensive", "medium" and "low price" neighborhood**.

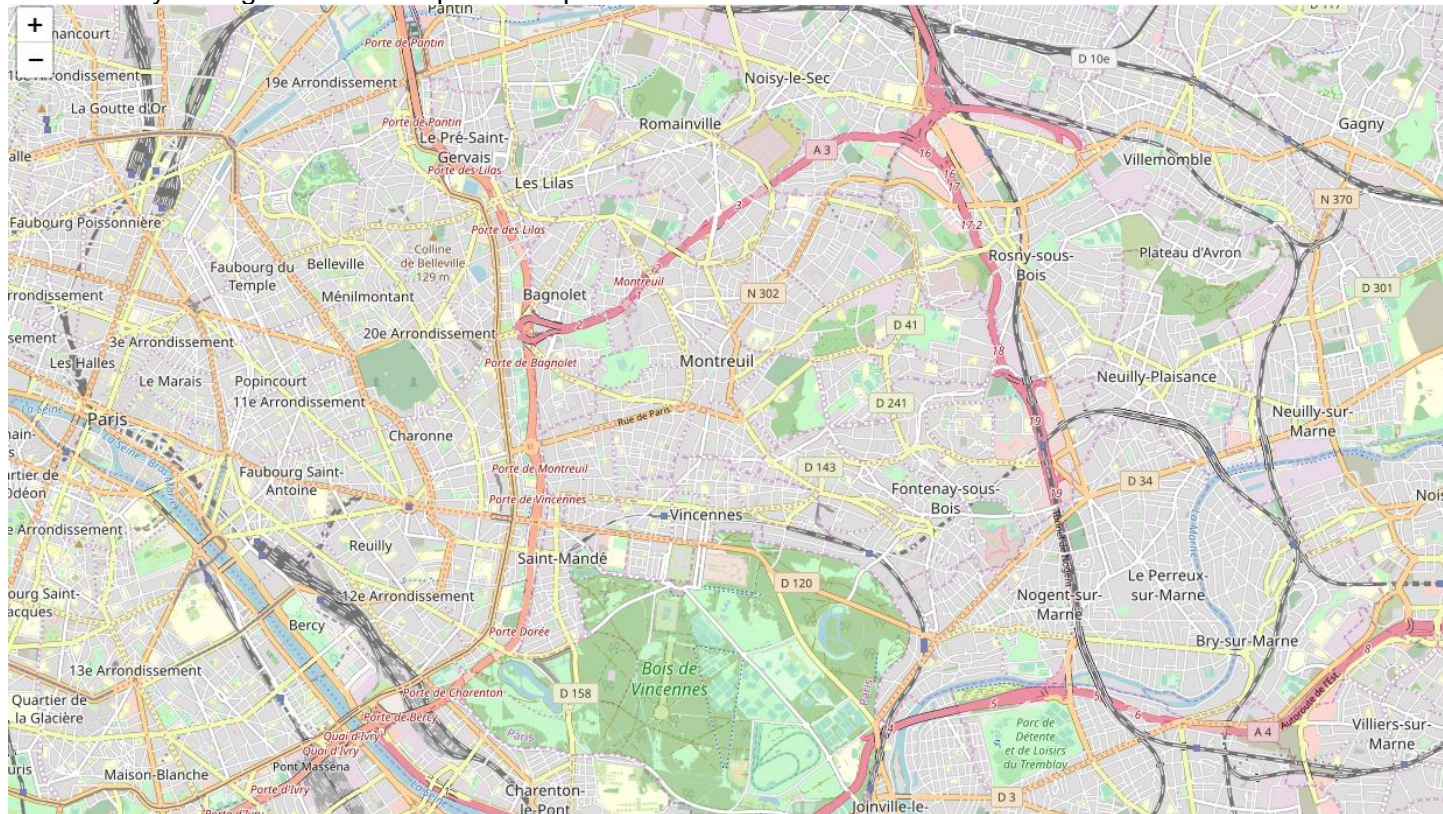
We will then use then our data science powers to make some recommendation for the Mayor in order to adjust the City Policies. Particularly, we will try to find **what kind of venues** should be encouraged in order to increase the **attractiveness of the low-price neighborhoods**.

## II. Data acquisition and cleaning

### II.1 The geographical region we will study

The geographical region we will focus on is East Paris and the cities close to Paris on it's East side.

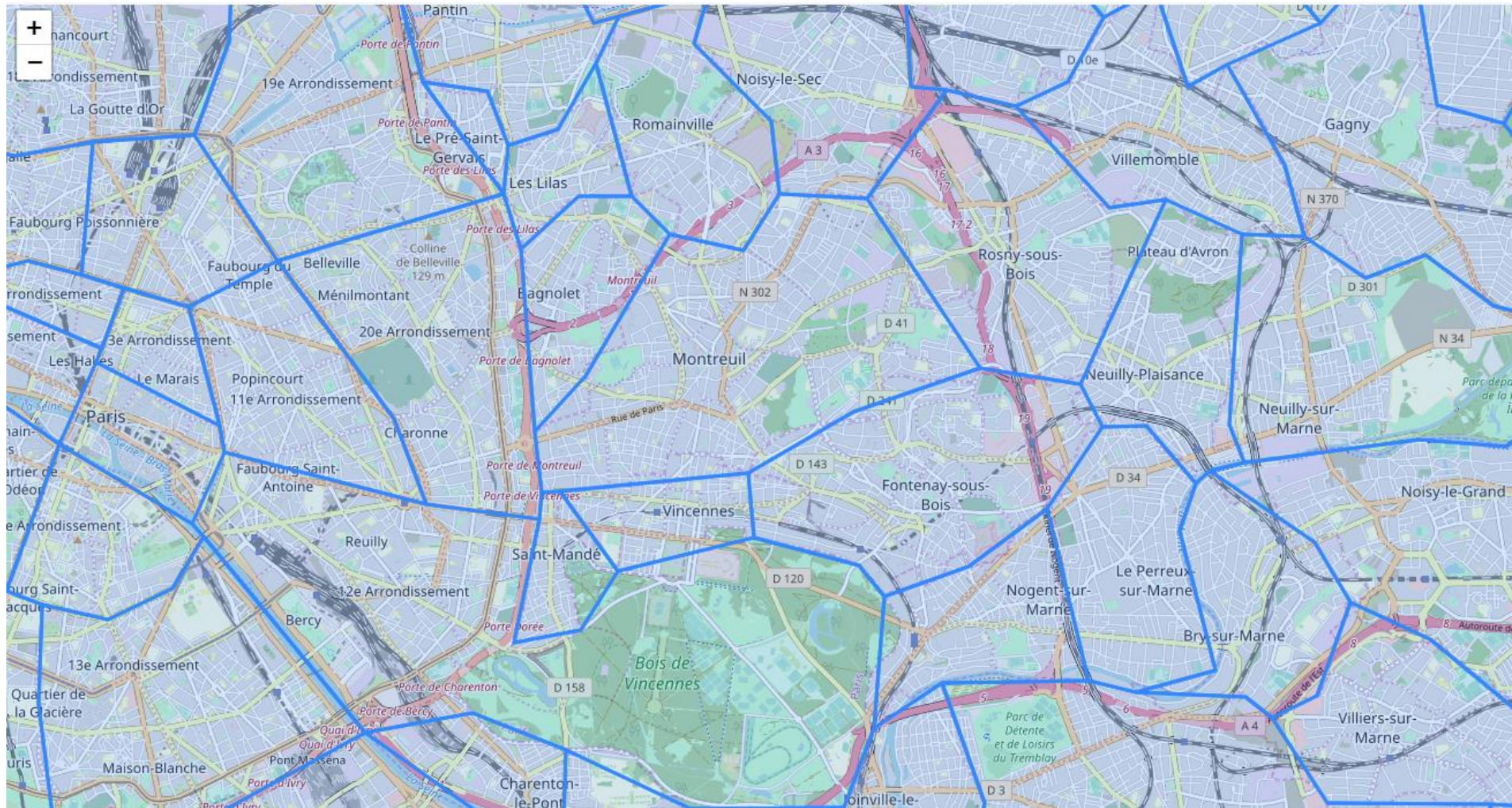
Let's start by having a look on this part of map.



Let's mark on the map the boundaries of the **Cities** and "**Arrondissements**" we are interested in.

We will use for this the GeoJson file available on <https://france-geojson.gregoireddavid.fr>





## II.2 Data related to real estate transactions

Data associated with real estate transaction is public, and a full description (in French) can be found on the site below:

<https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres-geolocalisees/>

The brute data can be downloaded using the site <https://www.data.gouv.fr/fr/datasets/les-communes-d-ile-de-france-idf/>. Data contains one line for each real estate element which was sold. This includes apartments, houses, terrain but also dependencies, shops and industrial buildings.

I have downloaded and studied data corresponding to the real estate sales in 2018 in 3 departments from Ile de France:

- 75 – Paris
- 93 - Seine Saint Denis
- 94 - Val de Marne

They correspond to the Paris East Region we want to study.

Brute data contains. 156,000 lines corresponding to unitary transactions on flats, houses, parkings, shops etc. sold in 2018 in these 3 departments.

We will keep only 14 useful columns which correspond to:

- transaction id and date
- address
- price of the transaction
- real estate type of element
- surface and number of pieces
- latitude and longitude.

Let's take a look on how many elementary transactions we have per Arrondissement or City for our **city\_list**:

Postal code	City Name	Number of elementary transactions
75003	Paris 3e Arrondissement	1347
75004	Paris 4e Arrondissement	1378
75005	Paris 5e Arrondissement	1383
75010	Paris 10e Arrondissement	3254
75011	Paris 11e Arrondissement	4832
75012	Paris 12e Arrondissement	3703
75013	Paris 13e Arrondissement	2883
75018	Paris 18e Arrondissement	6245
75019	Paris 19e Arrondissement	5134
75020	Paris 20e Arrondissement	3658

93100	Montreuil	2693
93110	Rosny-sous-Bois	1813
93170	Bagnolet	984
93230	Romainville	5685
93260	Les Lilas	1011
93310	Le Pré-Saint-Gervais	495
93500	Pantin	1650
94120	Fontenay-sous-Bois	1630
94130	Nogent-sur-Marne	1342
94160	Saint-Mandé	869
94220	Charenton-le-Pont	940
94300	Vincennes	1777
94340	Joinville-le-Pont	527
94410	Saint-Maurice	553

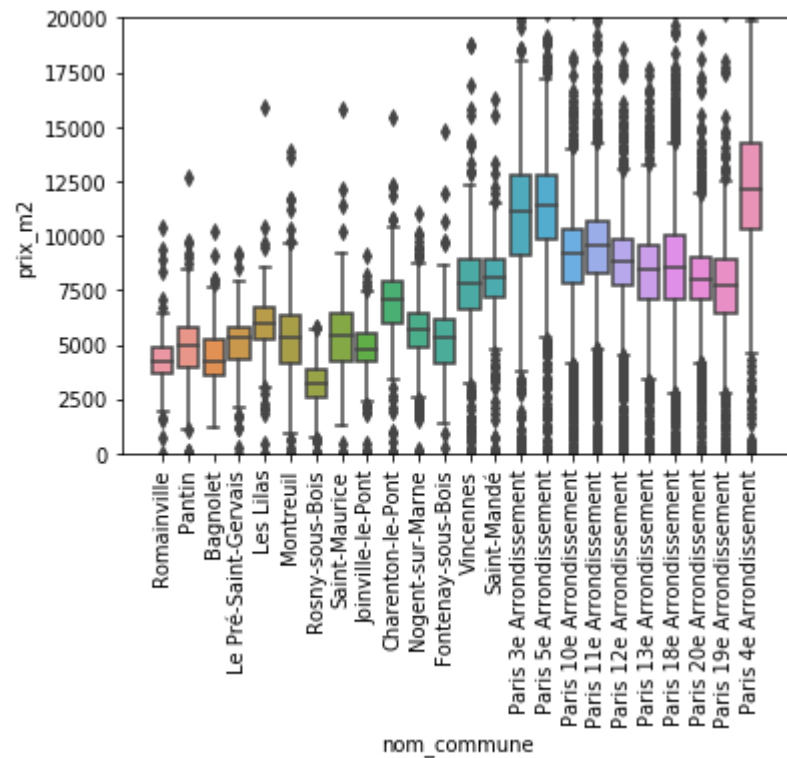
Through the data cleaning process, we will perform the following actions:

- group the elements per transaction; we will keep a **single line per transaction** (while in the original files there was a line per object sold)
- keep only sales transactions (drop auctions, donations etc.)
- keep only transactions concerning the apartments and drop everything else

We will now calculate and enrich the dataframe with the **price per square meter**.

Let's take a look at the distribution of price per square meter in our data, city by city:





So, depending on the cities, median price goes between **3000€/m²** and more than **12000€/m²**.

Let's take a look at the outliers, and simply drop them (prices per square meter above 20000€ or below 500€). The number of transactions we will keep per city is here below.

Postal code	City Name	Number of elementary transactions
75003	Paris 3e Arrondissement	760
75004	Paris 4e Arrondissement	524
75005	Paris 5e Arrondissement	824
75010	Paris 10e Arrondissement	1627
75011	Paris 11e Arrondissement	2561
75012	Paris 12e Arrondissement	1675
75013	Paris 13e Arrondissement	1576
75018	Paris 18e Arrondissement	3626

75019	Paris 19e Arrondissement	1829
75020	Paris 20e Arrondissement	2123
93100	Montreuil	1035
93110	Rosny-sous-Bois	472
93170	Bagnolet	312
93230	Romainville	205
93260	Les Lilas	275
93310	Le Pré-Saint-Gervais	198
93500	Pantin	643
94120	Fontenay-sous-Bois	430
94130	Nogent-sur-Marne	572
94160	Saint-Mandé	407
94220	Charenton-le-Pont	427
94300	Vincennes	849
94340	Joinville-le-Pont	206
94410	Saint-Maurice	211

So, at the end of the data cleaning process, we have a dataframe containing approx. **23000 real estate transactions** done in **2018** in the **24 cities of our choice**, with the associated **price / square meter**.

## II.3 Divide geographical space in square tiles

Analyzing data only per city is a too coarse approach, as cities can be quite large in surface and number of inhabitants. So we will proceed to a more refine approach, by covering the geographical zone of interest with a grid of 18x18 tiles.

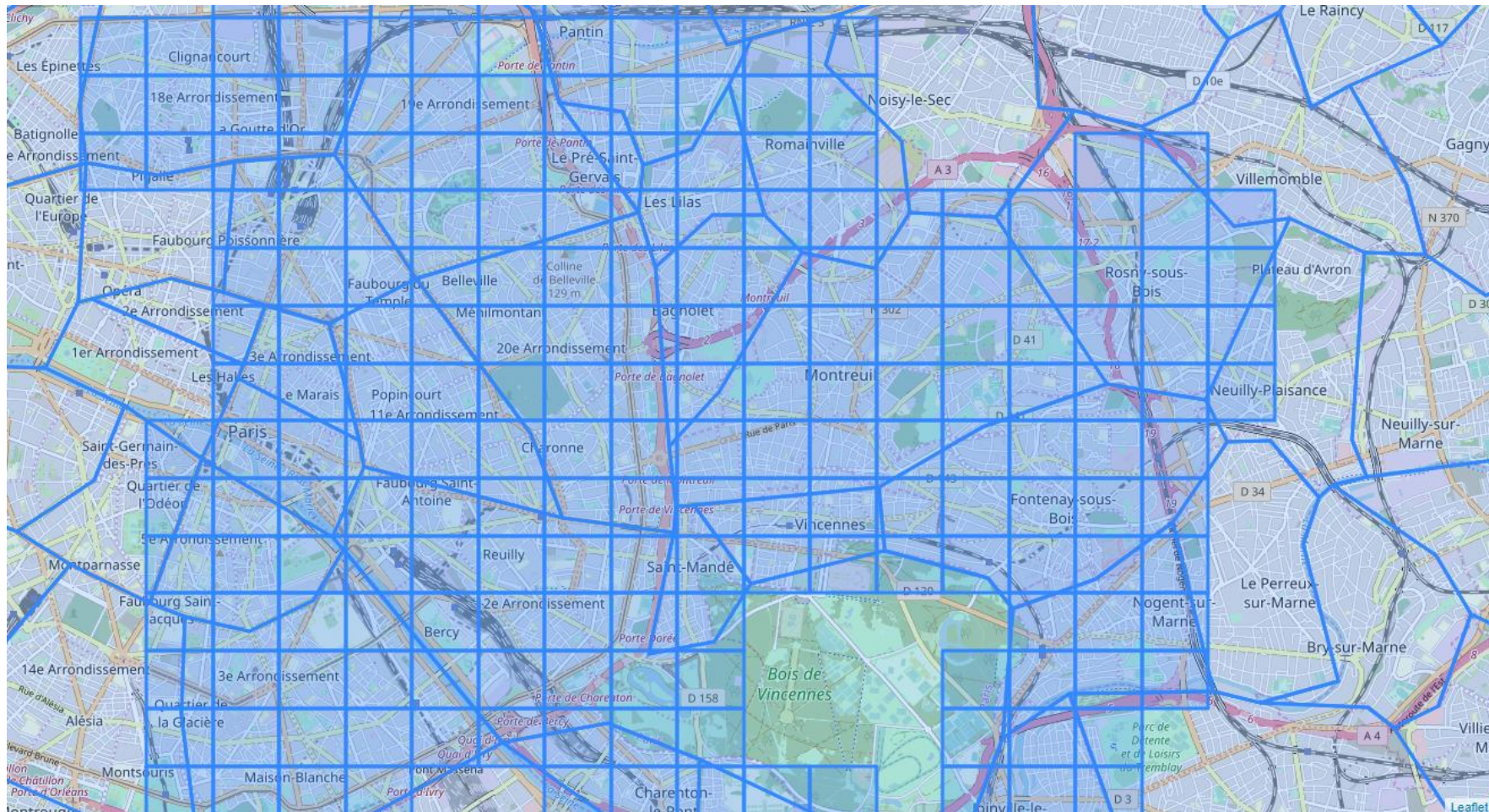
We will enrich this tile list with the geographical information (latitude, longitude, square borders), and also the postal address in the center of each tile.

We will also calculate the number of real estate transactions that occurred in each tile, and the mean price / square meter for the associated transactions. At the end, our geographical list of tile looks like this:

	bins_lat	bins_long	prix_m2	nombre_transactions	lat_s	lat_n	latitude	long_w	long_e	longitude	tile_id	address
0	0	12	5189.112510	6	48.811961	48.817690	48.814826	2.446647	2.456692	2.451669	00_12	Avenue Joffre, 94700, Maisons-Alfort
1	0	13	4731.721471	131	48.811961	48.817690	48.814826	2.456692	2.466738	2.461715	00_13	Rue du Maréchal Leclerc, 94410, Saint-Maurice
2	0	14	4788.432368	49	48.811961	48.817690	48.814826	2.466738	2.476783	2.471761	00_14	Villa de la Grotte, 94340, Joinville-le-Pont
3	0	15	4262.313225	14	48.811961	48.817690	48.814826	2.476783	2.486829	2.481806	00_15	Rue Diderot, 94500, Champigny-sur-Marne
4	1	1	8161.830850	42	48.817690	48.823419	48.820555	2.336146	2.346191	2.341168	01_01	Avenue de la Sibelle, 75014, Paris 14e Arrondi...

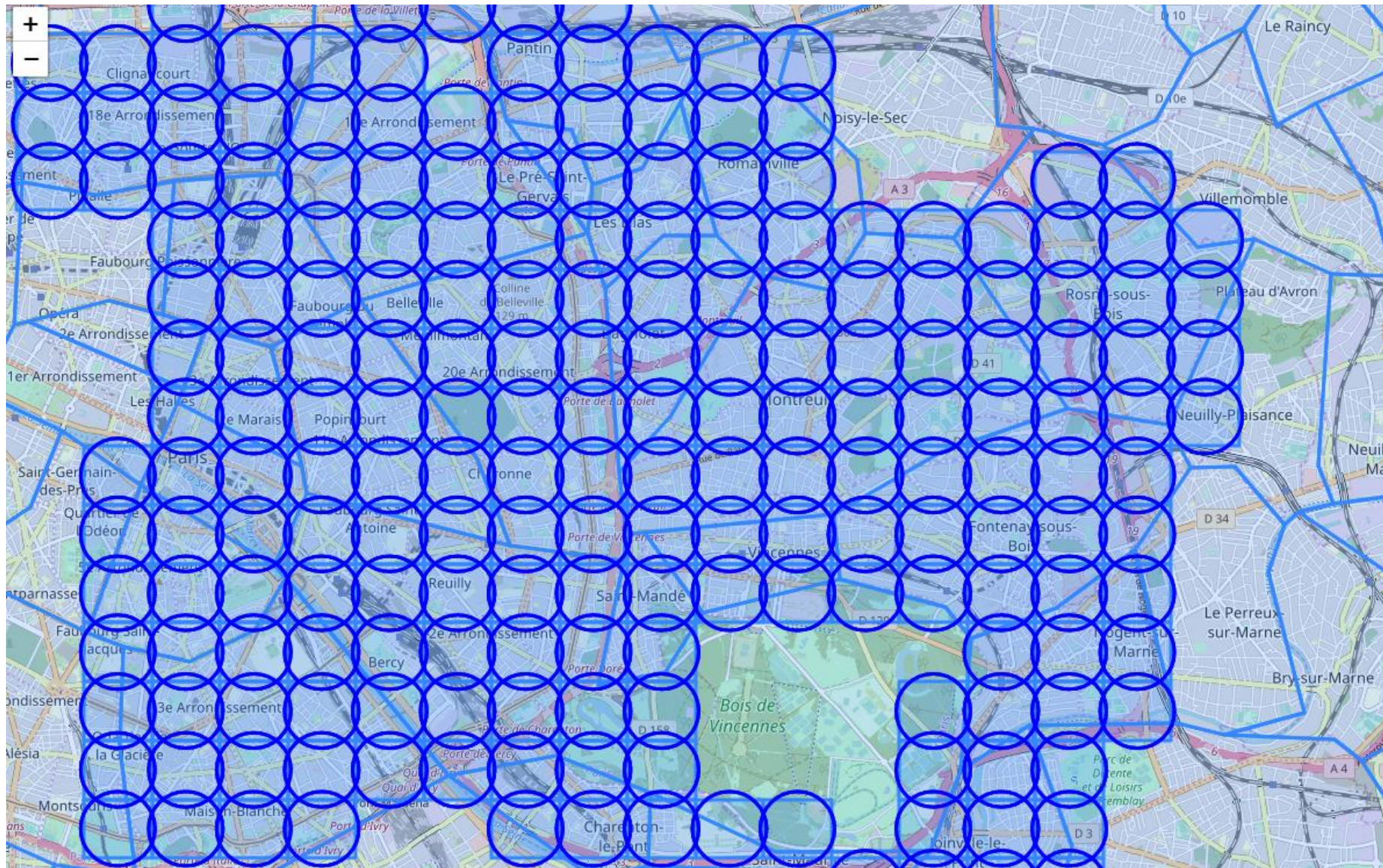


At the end, our tile grid, superposed on our geographical zone of interest looks like this:





On the top of that, we will center the circles in which we will look for Foursquera venues. The optimal radius seems to be 400 meters. The centers of the circles are in the center of each tile. The circles are superposing a little bit. However, we can accept this, as this gives the venues in the proximity of each tile center.



Looks good, but it is a little bit clumsy :-)

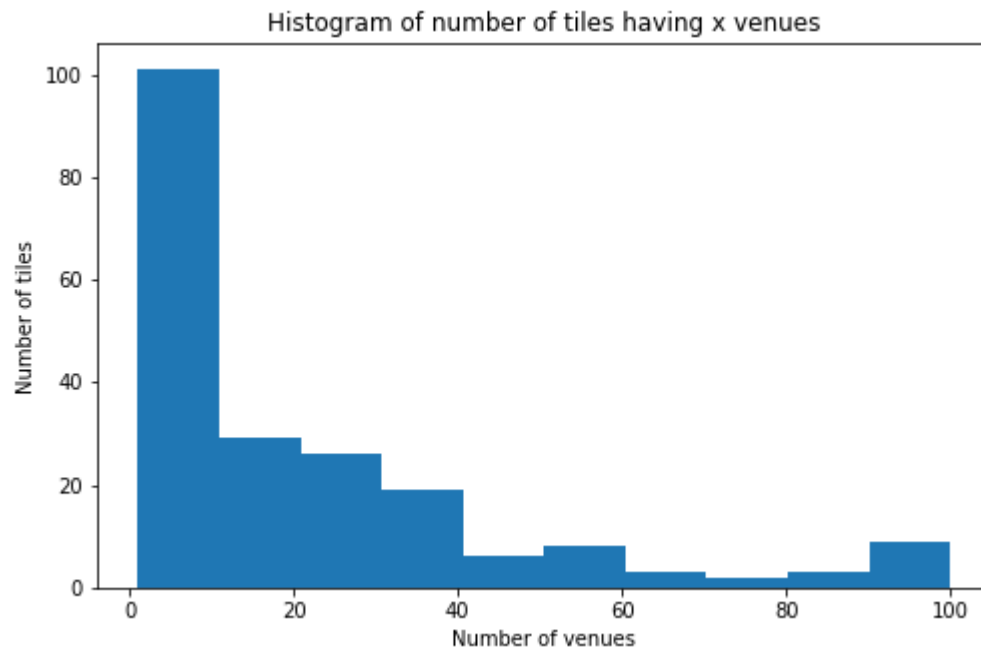


## II.4 Foursquera data collection and cleaning

Using the Foursquera API, we will collect the venues for the list of tiles we have built:

	tile_id	Tile Latitude	Tile Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
4405	15_06	48.900762	2.391396	Place Auguste Baron	48.900572	2.387162	Plaza
4406	15_06	48.900762	2.391396	Super Insolite - boutique en ligne de cadeaux ...	48.899533	2.395620	Gift Shop
4407	15_07	48.900762	2.401442	MurMur	48.902748	2.401252	Climbing Gym
4408	15_07	48.900762	2.401442	Halle Papin	48.904150	2.400839	Performing Arts Venue
4409	15_07	48.900762	2.401442	La Cité Fertile	48.898687	2.398871	General Entertainment
4410	15_07	48.900762	2.401442	RER Pantin [E]	48.897989	2.400438	Train Station
4411	15_07	48.900762	2.401442	Arrêt Pantin Gare	48.897897	2.400650	Bus Stop
4412	15_08	48.900762	2.411487	Galerie Thaddaeus Ropac - Pantin	48.899350	2.408075	Art Gallery
4413	15_08	48.900762	2.411487	Mimoza	48.899457	2.409012	Restaurant
4414	16_06	48.906491	2.391396	Sidi bousaïd (cuisine tunisienne)	48.904839	2.392877	Mediterranean Restaurant
4415	16_06	48.906491	2.391396	Superlav	48.905455	2.394582	Laundromat
4416	16_06	48.906491	2.391396	La Villa Mais d'Ici	48.906391	2.387434	Performing Arts Venue
4417	16_06	48.906491	2.391396	Klingooroo	48.907874	2.395802	Theme Park Ride / Attraction
4418	16_07	48.906491	2.401442	Halle Papin	48.904150	2.400839	Performing Arts Venue
4419	16_07	48.906491	2.401442	Compagnie d'arc de Pantin	48.909057	2.401340	Gym / Fitness Center

We get more than 4400 venues, covering all our tile space. A quick histogram will help to understand how many venues we have per tile.



So, there are approx. 100 tiles on our area, having a very small (less than 10) number of interesting venues. And 10 of them have the maximum number of venues (i.e.100).

We will now enrich our tile list dataframe with number of venues associated to each tile. From the final list of tiles of interest, we will drop the tiles which have less than 20 real estate transactions and less than 4 venues per tile.

Our **tiles** dataframe contains now a list of 151 geographical square zones (tiles), with the **average price per square meter**, the number of real estate transactions and the **number of venues** of interest for each tile. The tile dataframe looks now like this:

	bins_lat	bins_long	prix_m2	nombre_transactions	lat_s	lat_n	latitude	long_w	long_e	longitude	tile_id	address	Tile Latitude	Tile Longitude	nombre_venues
0	0	13	4731.721471	131	48.811961	48.817690	48.814826	2.456692	2.466738	2.461715	00_13	Rue du Maréchal Leclerc, 94410, Saint-Maurice	48.814826	2.461715	8
1	0	14	4788.432368	49	48.811961	48.817690	48.814826	2.466738	2.476783	2.471761	00_14	Villa de la Grotte, 94340, Joinville-le-Pont	48.814826	2.471761	4
2	1	1	8161.830850	42	48.817690	48.823419	48.820555	2.336146	2.346191	2.341168	01_01	Avenue de la Sibelle, 75014, Paris 14e Arrondi...	48.820555	2.341168	11
3	1	2	7737.428188	52	48.817690	48.823419	48.820555	2.346191	2.356237	2.351214	01_02	Rue Gouthière, 75013, Paris 13e Arrondissement	48.820555	2.351214	14
4	1	3	7066.006600	112	48.817690	48.823419	48.820555	2.356237	2.366282	2.361259	01_03	Rue Gandon, 75013, Paris 13e Arrondissement	48.820555	2.361259	32

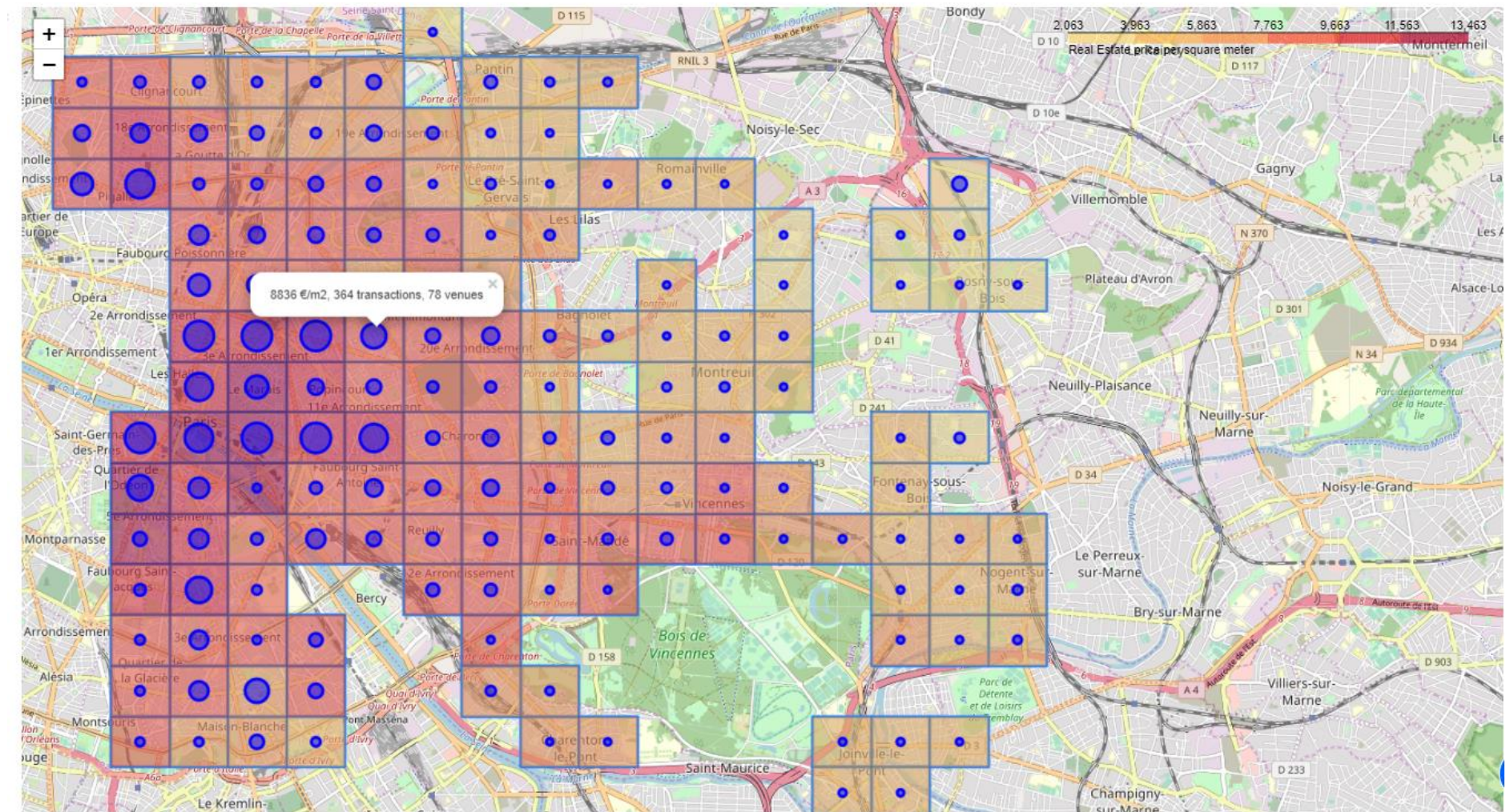
## II.5 Show real estate and number of venues on the same map

Our final part of the data analysis is to figure out - on the same map:

- the average price per square meter
- the number of venues



for all the tiles which cover our interest area. We get a final map like this one:



The choropleth is colored according to the average price per square meter, which goes between 2000€ and 13500€. The radius of each blue circle is proportional to the number of venues of interest in the tile, which goes between 4 and 100.

As a first conclusion of this Data Analysis section, it seems that we have a **direct correlation** between the **real estate price per square meter** in a zone and the **number of venues on interest** in the same region.

# III. Methodology

In this project we try to find if there is a **correlation** between the real estate prices and the type of interest venues in the neighborhood. We will try to describe what is the **venue segment associated with a "expensive", "medium" and "low price" neighborhood**.

We will then use then our data science powers to make some recommendation for the Mayor in order to adjust the City Policies. Particularly, we will try to find **what kind of venues** should be encouraged in order to increase the **attractiveness of the low price neighborhoods**.

As a first step we have collected the required data:

- we have defined a list of geographical zones, represented as tiles, for which we have calculated the average real estate price for apartments;
- for each tile, we have collected also the number and the categories of venues of interest (according to Foursquare categorization).

In the second step in our analysis we will perform three correlation approaches:

- first we will do a simple linear regression between the number of interest venues in a tile and the average real estate prices of the apartments;
- second, we will segment the tiles according to the type of venues present in each one; we will try to identify the segments associated with low, medium and high price neighborhoods;
- third, we will perform a multiple linear regression between the detailed type of venues and the real estate price in each tile.

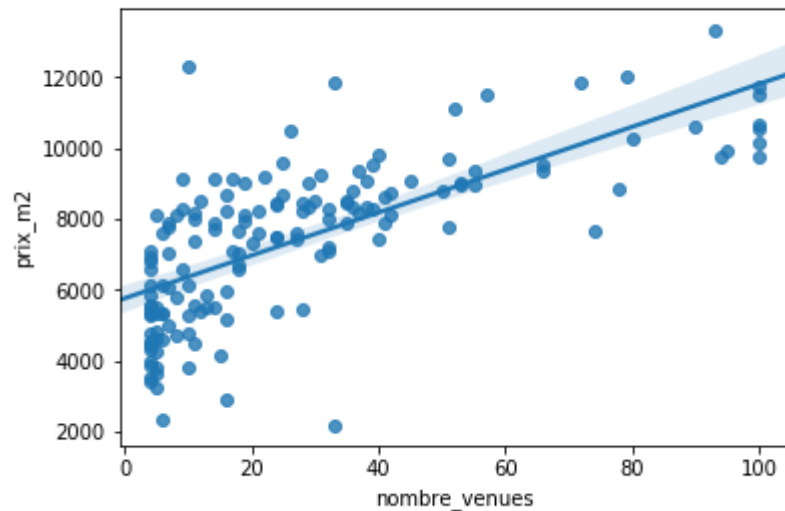
Finally, after scoring each approach, we will try to identify the precise segments or categories of venues with a strong positive correlation with the real estate prices. The associated City policies will be to encourage the implantation of those type of venues which increase the attractivity of each neighborhood.

## IV. Analysis

### IV.1 Simple linear regression

Let's perform a simple linear regression between the number of venues of interest and the real estate price per square meter associated to each tile.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7efff1c36668>
```



We use the basic LinearRegression model and measure its performance using the R2 (variance) score. We get a **variance score of 0.47** which shows a **weak positive correlation** between the number of venues of interest and the price per square meters of the apartments in each tile.

### IV.2 Segment the venues per tile

The second type of analysis will be to segment and cluster the tiles, based on the categories of venues present in each tile.

There are 324 different types of categories among the Foursquera venues found for our geographical area.



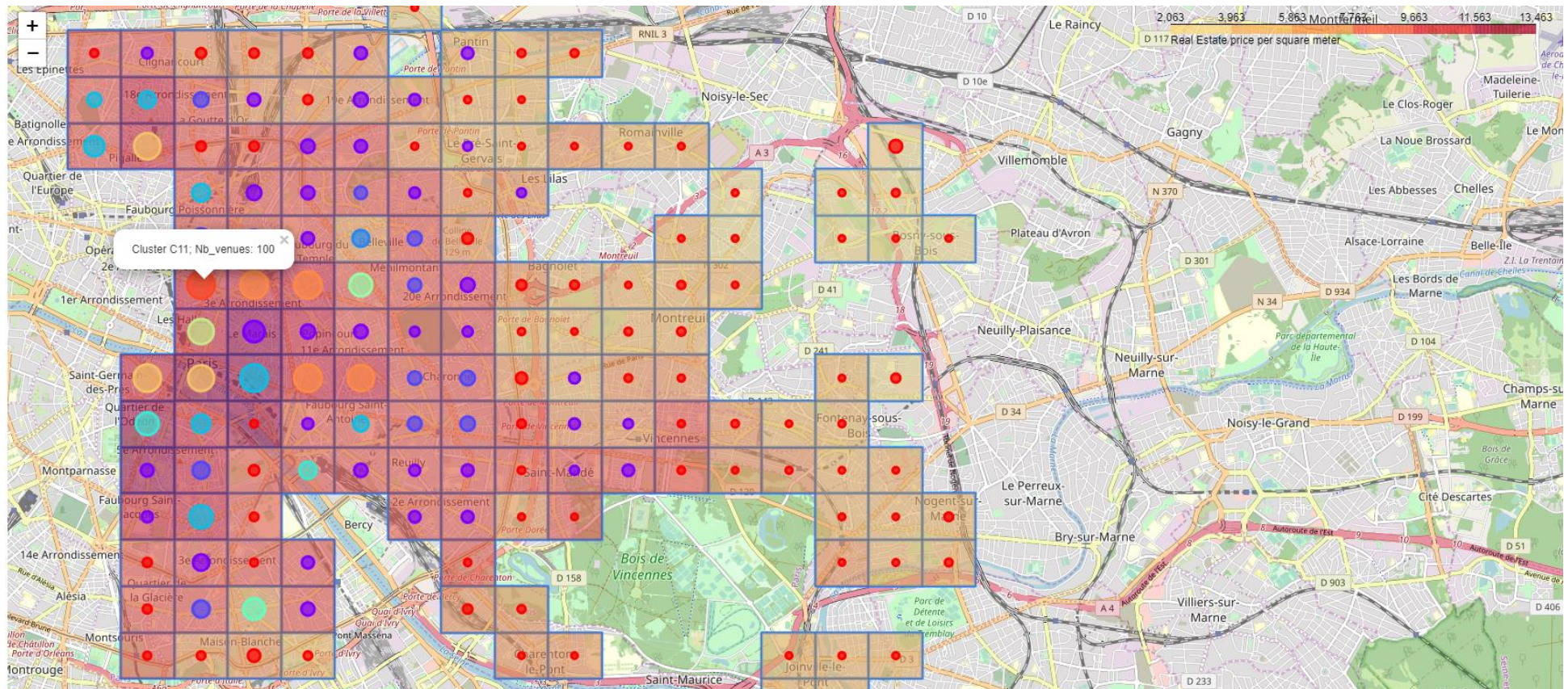
Segmentation will be done based on the number of venues per category present in each tile. We will use the KMeans algorithm in order to segment the tiles in 12 different clusters.

We will not normalize data, as each tile has a maximum number of 100 venues (number of venues is already normalized).

We will now associate a cluster label to each tile. Labels are from 0 to 11.













Now let's show our map with the 3 features associated to each tile:

- the average price per square meter (the choropleth color of each tile)
- the number of venues per tile, represented by the size of the circle marker
- the segment of the tile, represented by the color of each circle marker.



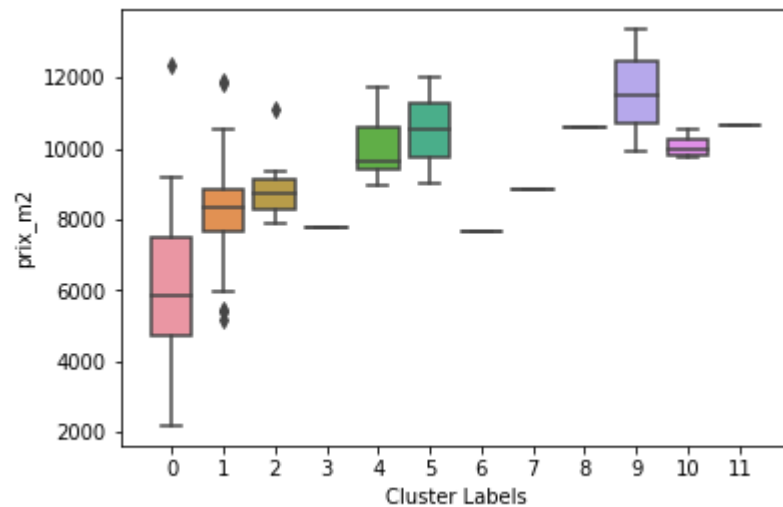


The 12 different clusters can be described as following

Cluster label	Mean number of venues	Price per square meter	Description	Most common venues					
0 - 	10	Low	Out of Paris, few venues, most utilities	0.63 Supermarket	0.52 Hotel	0.46 French Restaurant	0.31 Bakery	0.31 Plaza	0.25 Café
1 - 	30	Medium	Paris or close to Paris mix area	3.08 French Restaurant	1.83 Hotel	1.25 Bar	1.19 Café	1.14 Italian Restaurant	1.14 Japanese Restaurant
2 - 	44	Medium	Paris residential area	5.92 French Restaurant	4.50 Bar	1.75 Hotel	1.50 Bistro	1.50 Pizza Place	1.33 Coffee Shop
3 - 	51	Medium	Paris – Bars and Asian cuisine	9.00 Bar	6.00 Chinese Restaurant	6.00 Vietnamese Restaurant	4.00 French Restaurant	3.00 Supermarket	2.00 Dim Sum Restaurant
4 - 	62	High	Central touristic area	11.62 French Restaurant	3.88 Hotel	3.62 Italian Restaurant	2.62 Bakery	2.00 Café	1.62 Bar
5 - 	66	High	Central touristic area	9.00 Hotel	6.00 French Restaurant	3.50 Sandwich Place	2.00 Indie Movie Theater	2.00 Nightclub	2.00 Cocktail Bar
6 - 	74	Medium	Chinese neighborhood	16.00 Vietnamese Restaurant	11.00 Asian Restaurant	9.00 Thai Restaurant	5.00 Chinese Restaurant	4.00 French Restaurant	3.00 Supermarket
7 - 	78	Medium	Paris Republique – Bars & Bistros	21.00 Bar	3.00 Pizza Place	3.00 French Restaurant	3.00 Restaurant	2.00 Middle Eastern Restaurant	2.00 Bistro
8 - 	90	High	Central touristic area	8.00 French Restaurant	5.00 Bakery	4.00 Hotel	3.00 Furniture / Home Store	3.00 Cosmetics Shop	3.00 Plaza
9 - 	96	High	Central touristic area	19.00 French Restaurant	4.00 Plaza	3.67 Hotel	3.33 Ice Cream Shop	3.00 Bakery	3.00 Italian Restaurant
10 - 	98	High	Central touristic area	12.75 French Restaurant	6.50 Bar	4.25 Bistro	3.75 Hotel	3.50 Wine Bar	2.75 Restaurant
11 - 	100	High	Central “chic” area	9.00 Cocktail Bar	6.00 French Restaurant	6.00 Wine Bar	5.00 Hotel	5.00 Bakery	4.00 Thai Restaurant

We can look for a correlation between the prices and the clusters, as defined above. Data are quite dispersed, and we can just associate the various clusters with Low, medium and high price neighborhoods:

- Low price: cluster 0
- Medium price: clusters 1, 2, 3, 6 and 7
- High level price: clusters 4, 5, 8, 9, 10, 11



Now let's use a cleverer linear regression. We will use a one hot encoding in order to classify the cluster of each tile. Then, we will do a multiple linear regression in order to model the price per square meter of each tile through the associated cluster of its venues.

We will split the feature X dataset (represented by the Cluster labels per tile) and the y dataset (the average price per square meter per tile) into a train and test set.

Then we will apply a Ridge multiple regression algorithm - with regularization. The RidgeCV method will automatically chose the optimal alpha parameter, which will be chosen between 0.00001 and 10000.

```
Intercept: [9205.33191732]
Alpha: 0.3831186849557293
Variance score train: 0.55
Variance score test: 0.37
```

In this case, the model fits quite poor. However, we have a **final variance score of only 0.37**.

## IV.3 Multiple linear regression based on venue categories

Now let's try the last approach and perform a multiple linear regression based on the category of venues present in each tile. Our feature set X will be the number of venues - by category - in each tile.

The y dataset is the same: the average price per square meter per tile.

As in the previous case, we will use a Ridge multiple linear regression algorithm with Cross Validation embedded. The algorithm will calculate the optimal alpha.

```
Intercept: [5848.54311144]
```

```
Alpha: 46.41588833612782
```

```
Variance score train: 0.71
```

```
Variance score test: 0.58
```

We have now a slightly **better variance score (0.58)**, the best among the three methods we have used by now.

Now let's examine which are the venue types with the best positive impact on the real estate price:

	Venue type	Coefficients
34	Bike Rental / Bike Share	90.629154
36	Boat or Ferry	91.252307
287	Sushi Restaurant	95.987113
229	Plaza	96.167862
193	Metro Station	104.286308
49	Burger Joint	104.464146
72	Church	106.892597
217	Pastry Shop	108.642213
310	Vegetarian / Vegan Restaurant	113.688283
62	Canal Lock	113.896667
77	Coffee Shop	116.578945
112	Falafel Restaurant	119.387966
141	Gourmet Shop	124.331205
218	Pedestrian Plaza	124.636646
135	Gastropub	127.436126
151	Historic Site	133.162531
70	Chinese Restaurant	134.736091
127	French Restaurant	146.000185
168	Japanese Restaurant	163.238495
35	Bistro	163.795533
256	Science Museum	186.993403
144	Gym	193.036590
132	Garden	207.822985
157	Hotel	210.575776
204	Museum	223.934375
216	Park	225.227832
56	Café	289.155960



Now let's do a last visualization. We will build a heat map which contains only the venues having a strong positive impact on the price. We will compare it with the choropleth map of the real estate prices. Here is what we get:

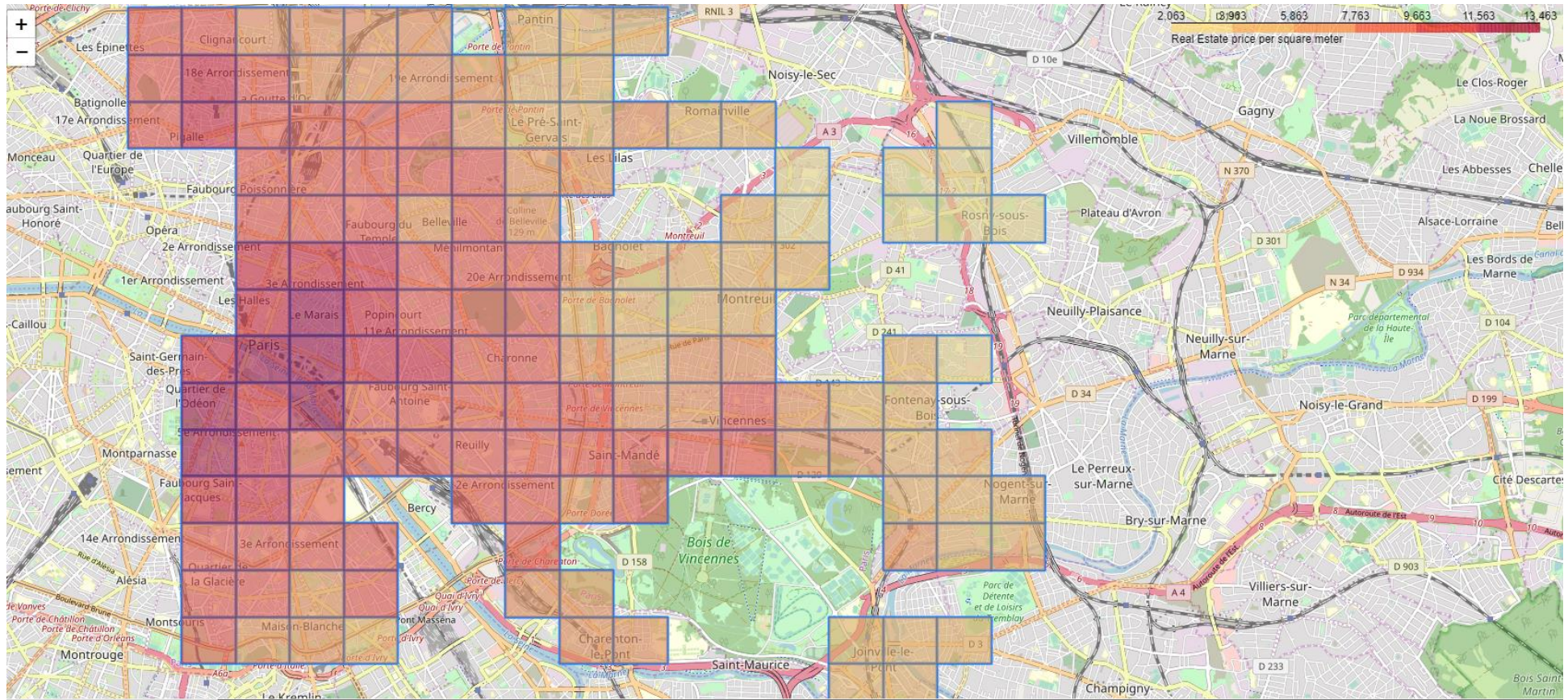


Figure 1 - Average price per square meter



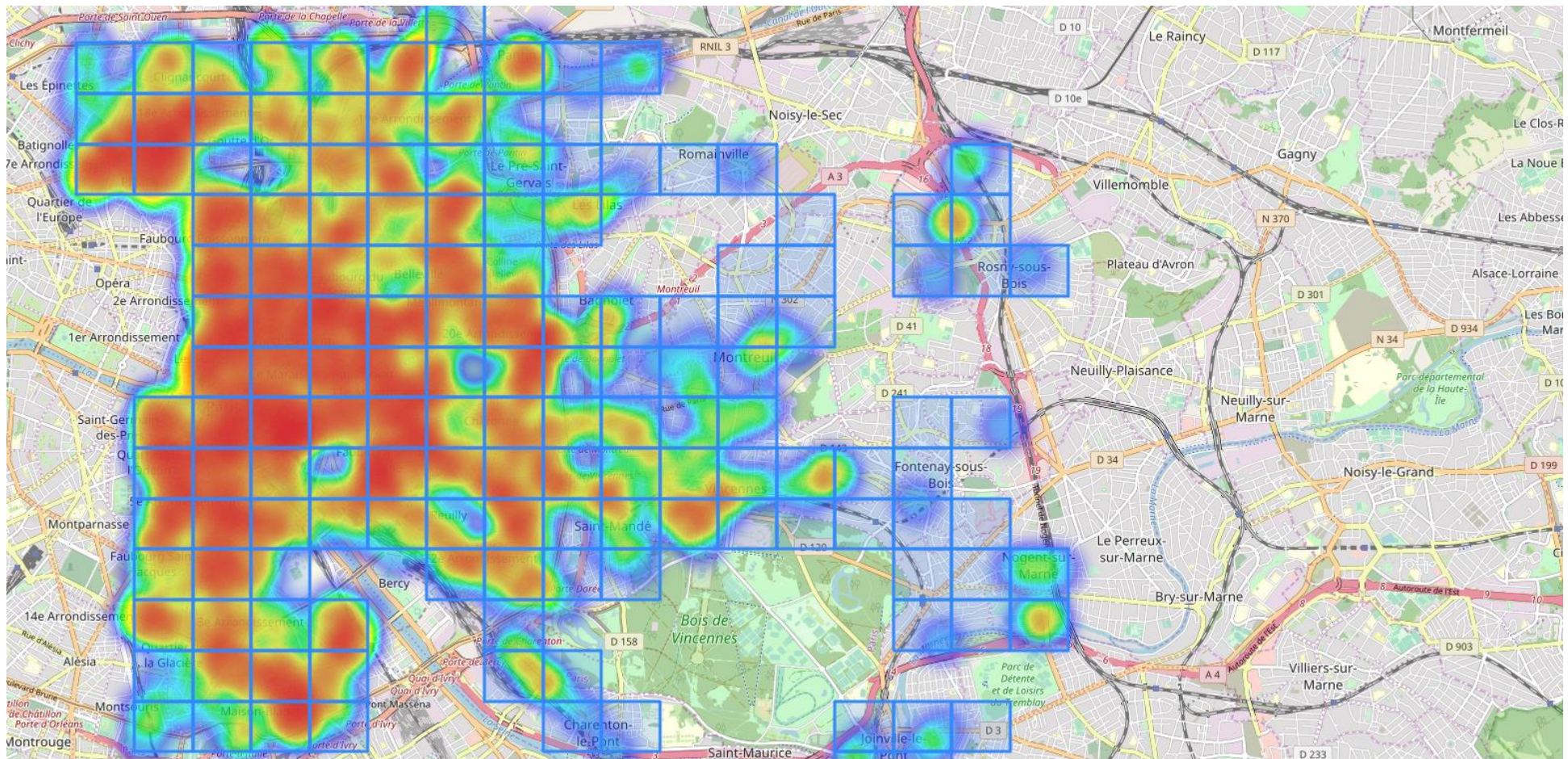


Figure 2 - Heat map of positive impact venues

## V. Results and Discussion

Our analysis shows that although there is a correlation between the number of venues of interest in each neighborhood and the real estate price per square meter, this simple linear correlation is quite weak (variance score of 0.47).

The strongest correlation is obtained when we analyze the type of venues associated with each neighborhood. And we can detect - in this case - type of venues which seem to have a positive influence on the real estate prices. Among the most prominent ones, some of them are actionable via public policies, such as:

Venue category	Correlation coefficient
Bike Rental / Bike Share	90.629154
Plaza	96.167862
Metro Station	104.286308
Pedestrian Plaza	124.636646
Historic Site	133.162531
Science Museum	186.993403
Garden	207.822985
Museum	223.934375
Park	225.227832

Some are related simply to geography, as "Boat or Ferry" or "Canal Lock". Others, even if not directly actionable, are signs of attractiveness and specific to the French \*art de vivre\*, as related to socializing and gastronomy:

Venue category	Correlation coefficient
Sushi Restaurant	95.987113
Burger Joint	104.464146
Pastry Shop	108.642213
Vegetarian / Vegan Restaurant	113.688283
Coffee Shop	116.578945
Falafel Restaurant	119.387966
Gourmet Shop	124.331205
Gastropub	127.436126
Chinese Restaurant	134.736091
French Restaurant	146.000185
Japanese Restaurant	163.238495
Bistro	163.795533
Gym	193.036590
Hotel	210.575776
Café	289.155960

Of course, these results can be considered only as a starting analysis point. The real estate price is good indicator of the attractiveness of a neighborhood. It can be influenced by the presence of attractive public infrastructures as parks, pedestrian plaza, gardens and museums and also by conviviality places such as cafés, gym, Japanese or French restaurants, bars and gourmet shops. However, our analysis don't take into consideration the presence of other features, such as the quality of the public schools, the criminality rate or the local tax level. It is likely that these factors have even a stronger influence on the attractiveness of each neighborhood.

But this analysis will be part of a further study ;-)



## VI. Conclusion

The purpose of this project was to try to help the **Mayor of Paris** and his colleagues nearby Paris to adjust the **City policies** in order to encourage **social diversity** and to increase the **attractiveness of the neighborhoods**.

We have focused on East Paris districts (10 arrondissements) and 14 surrounding cities close to East Paris. We analyzed data coming from two distinct sources:

- the Real Estate prices, available as public information for the 24 cities;
- the most **popular venues** available from Foursquera for the same geographical zone.

Through our analysis, we divided our geographical study zone in **151 square areas (tiles)**, for which we were able to calculate an average price per square meter for the apartments. Calculation was based on approx. 23000 transactions realized in these areas in 2018. We considered only neighborhoods where we were able to collect more than 20 transactions through the observed period.

For the same geographical area (151 square tiles), we collected more than 4000 venues of interest via the Foursquera API. We also kept into the study only the neighborhoods with more than 4 venues of interest.

We noticed - through the study - that there are big discrepancies of the mean price per square meter between the neighborhoods studied. The mean price varies between 2000€/m<sup>2</sup> and 13500€/m<sup>2</sup>. We took as an assumption the fact that the real estate mean price is a good indicator of the attractiveness of the neighborhood.

If we use the number and category of interest venues as a predictor for the real estate prices, we get the following results:

- there is a linear weak correlation between the number of interest venues in a neighborhood and the real estate prices;
- there is a stronger correlation between the presence of some specific category of venues and these prices.

This second point provides to the mayor an action lever allowing to increase the attractiveness of their neighborhoods. Our study shows that some public or semi-public infrastructures have a clear positive impact. Among them parks, pedestrian plaza, theaters, gardens and museums. Mayors can also encourage the installation of cafés, French restaurants, bistros and gourmet shops which also seem to have a positive influence on the neighborhood.

Of course, this study - even helpful - is only a start point. Public local policies must also consider the quality of the schools, the level of criminality, the level of noise, the social dynamics. These have also a strong and major impact on the attractiveness of each neighborhood.