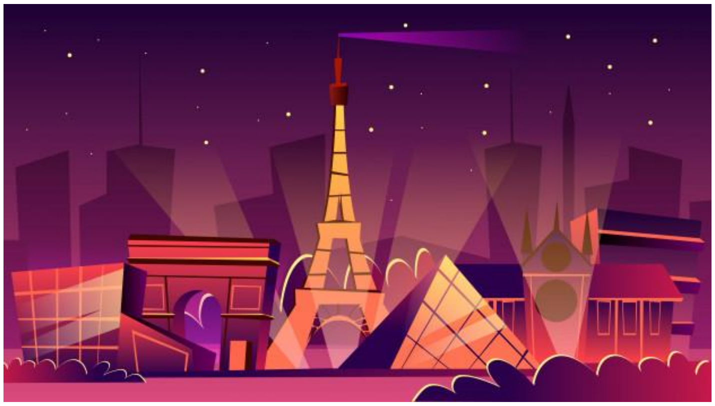
Capstone Project
The Battle of the Neighborhoods

Applied Data Science Capstone by IBM / Coursera

by Ludovic D'ALESSIO

Opening a traditional French bakery in Paris, France



Baguette e Vaquiere 1,60 pièce
Baguette du Chef 1,00€ (ou bien)

Contents

Introduction: Business Problem 3

Data 4

Data sources overview 4

Population density 4

Districts shapes 5

List and geo-localization of the bakeries 5

Methodology 8

Analysis 9

Complete list of bakeries 9

Division of Paris into hexagons 9

Clustering 12

Results 14

Discussion 17

Conclusion 18

INTRODUCTION: BUSINESS PROBLEM

In this project we will try to find suitable locations to open a **bakery** in **Paris**.

Let us imagine ourselves in the shoes of a young, gifted, traditional French baker willing to settle down in Paris. France is well known for its culinary wealth, and bakery makes no exception. An infinite variety of breads, croissants, pastries... hold a significant part of the ***French Way of Life***. French people are very proud of their bakeries, and you can find them everywhere. Every district, every block has its own bakery, which really plays an important part in the neighborhood's life.

So, the question is: **how to find a suitable place to open a new bakery in a city already crowded with bakeries of all kinds?**

Paris is a truly beautiful place, a mix of well-known landmarks, historical architecture, residential buildings and small local shops. Paris itself is quite small and homogeneous. Unlike the large U.S. cities for example, the business center is outside the city, and there are residential areas just everywhere.

Many factors could be taken into consideration to determine if an area is suitable to open a new traditional and high standing bakery, but we will only concentrate in this project on the three criteria below:

- *Density of population* in the area: you typically do not want to take you car to buy your bread for the day or the chocolate croissant for your breakfast, so the area's attractiveness is directly linked to the number of Parisians living around.
- *Number of bakeries already present in the area*: competition is good for the customers, but as a shop owner less competitors means more market share.
- *Distance of to the closest "quality bakery"*: all the bakeries are different, and you might want to walk a bit more to find an exceptional quality, hand-made product, within a reasonable range; so, an area where the closest top bakery is more than one kilometer away offers a true opportunity for a baker able and willing to provide this level of service.

DATA

Data sources overview

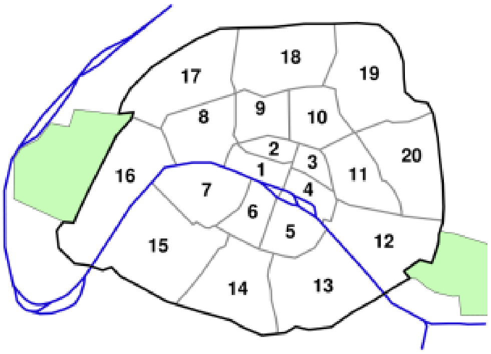
The following data sources will be used to generate the required information:

- **Paris population density** can be found on Wikipedia per borough and per administrative district.
- **Paris boroughs and districts shapes**, in `geojson` format, can be downloaded for free on opendata.paris.fr website.
- **The list and geo-localization** of all the bakeries in Paris will be retrieved through Foursquare API standard requests.
- **Bakeries ratings**, that will be used to identify top bakeries, will be retrieved through Foursquare API premium requests.
- The `folium` and `geopy.geocoders` packages will be used respectively to visualize data on a map and to retrieve map coordinates from given addresses.

All those pieces of data are completely and freely available on the internet. The following sections describe the data sources in detail and the data once retrieved.

Population density

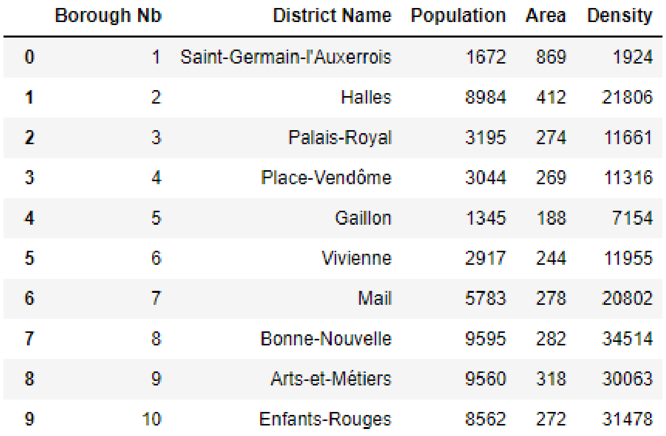
Paris is conveniently divided into 20 boroughs, called *arrondissements*, arranged in spiral and numbered from 1 to 20 starting from the center:



However, those boroughs are sometimes quite big, and the population density is not homogeneous. Fortunately, each of them is also divided into 4 administrative districts, which makes a total of **80 districts** covering the whole city. The list of the districts and their characteristics can be found here on Wikipedia (it's in French as the similar page in English does not directly show the density figures).

We will also need the `geojson` coordinates of the districts, that can be found on the opendata.paris.fr website.

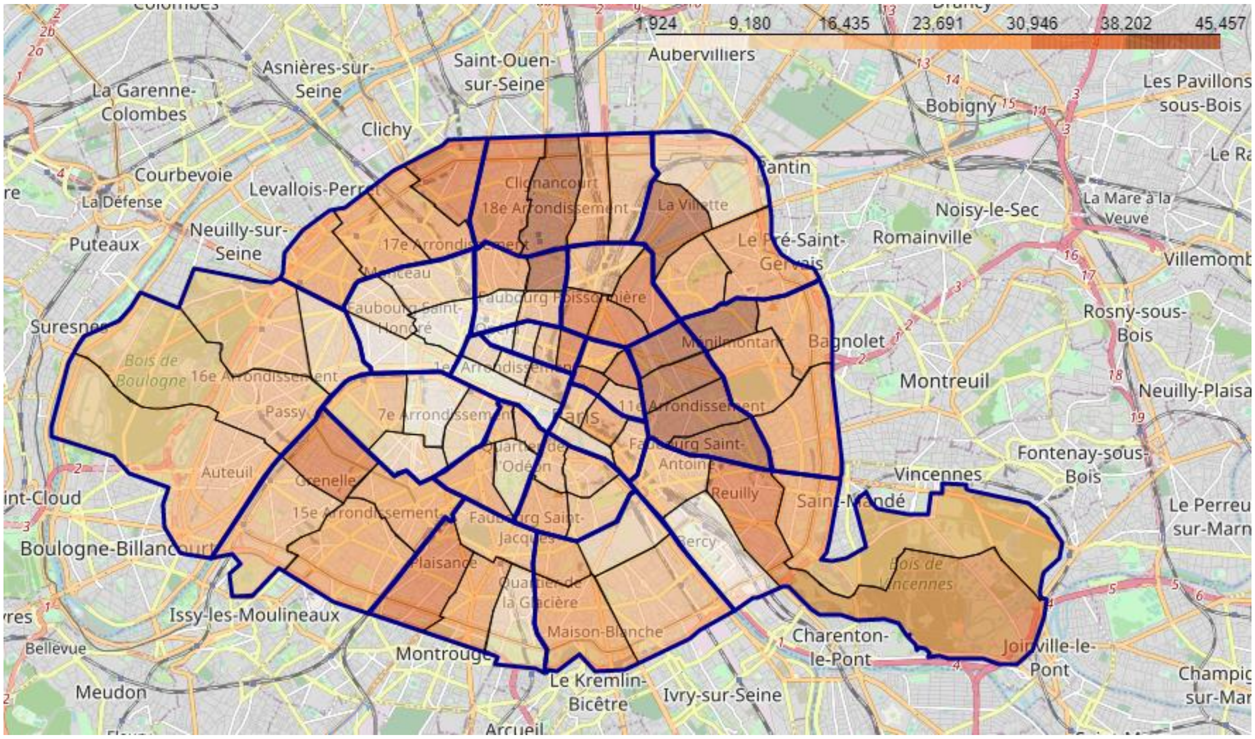
We can download the Wikipedia table directly into a Pandas `DataFrame`. Below the first 10 lines:



	Borough Nb	District Name	Population	Area	Density
0	1	Saint-Germain-l'Auxerrois	1672	869	1924
1	2	Halles	8984	412	21806
2	3	Palais-Royal	3195	274	11661
3	4	Place-Vendôme	3044	269	11316
4	5	Gaillon	1345	188	7154
5	6	Vivienne	2917	244	11955
6	7	Mail	5783	278	20802
7	8	Bonne-Nouvelle	9595	282	34514
8	9	Arts-et-Métiers	9560	318	30063
9	10	Enfants-Rouges	8562	272	31478

Districts shapes

Now we download the `geojson` files from opendata.paris.fr for the districts and the boroughs and display the boroughs as a choropleth map according to their population density:

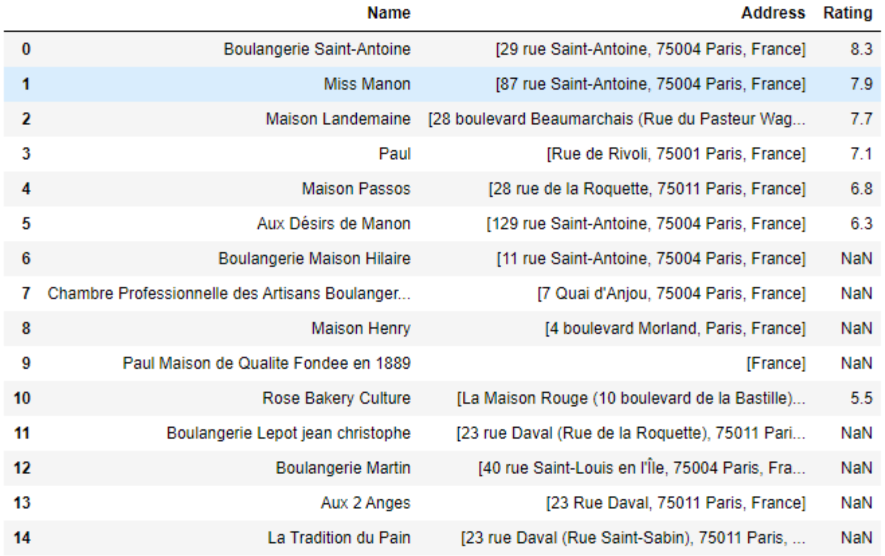


List and geo-localization of the bakeries

To retrieve the list of bakeries around a given point we use the Foursquare API with a verified account, that allows for 99,500 standard and 500 premium requests per day.

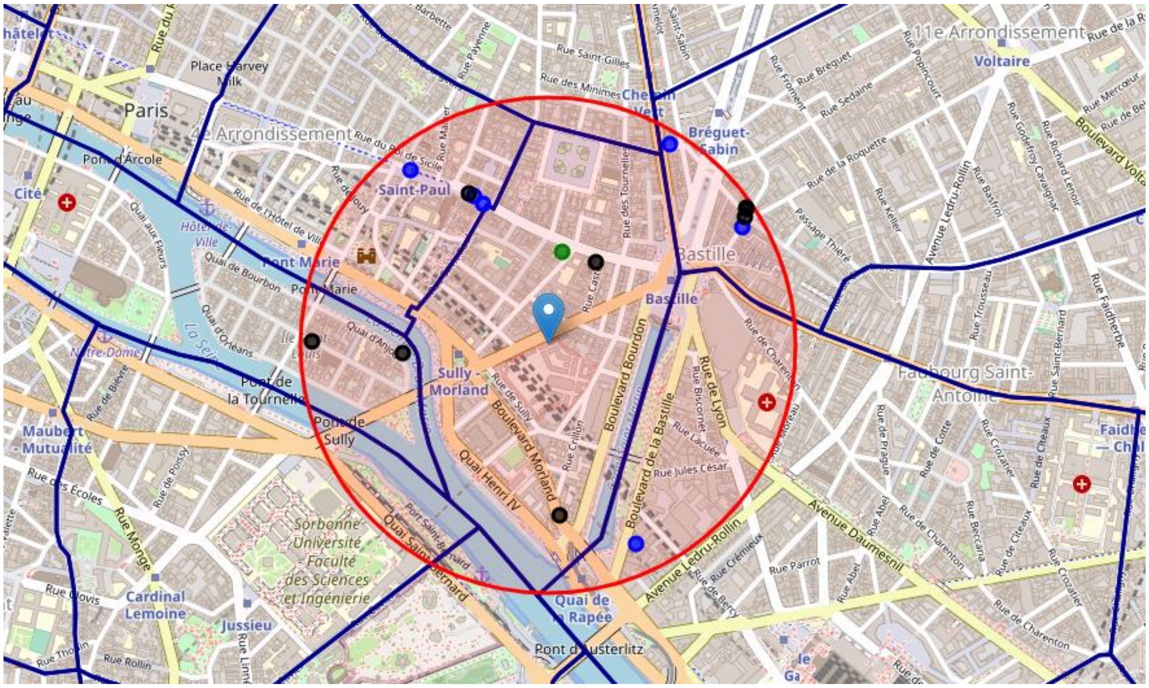
To demonstrate how Foursquare API works, we retrieve the bakeries within 600 meters of the center of one random district; for the sake of simplicity, we'll use the first district listed in the `geojson` file retrieved from opendata.paris.fr.

We must send one single standard Foursquare request to retrieve the list of bakeries, and to retrieve the rating of each bakery we need to send one premium request per bakery. The results are shown in the table below. It is important to note that not all venues in Foursquare are rated, so if the request returns no results, we assign `NaN` as the bakery's rating.



	Name	Address	Rating
0	Boulangerie Saint-Antoine	[29 rue Saint-Antoine, 75004 Paris, France]	8.3
1	Miss Manon	[87 rue Saint-Antoine, 75004 Paris, France]	7.9
2	Maison Landemaine	[28 boulevard Beaumarchais (Rue du Pasteur Wag...	7.7
3	Paul	[Rue de Rivoli, 75001 Paris, France]	7.1
4	Maison Passos	[28 rue de la Roquette, 75011 Paris, France]	6.8
5	Aux Désirs de Manon	[129 rue Saint-Antoine, 75004 Paris, France]	6.3
6	Boulangerie Maison Hilaire	[11 rue Saint-Antoine, 75004 Paris, France]	NaN
7	Chambre Professionnelle des Artisans Boulanger...	[7 Quai d'Anjou, 75004 Paris, France]	NaN
8	Maison Henry	[4 boulevard Morland, Paris, France]	NaN
9	Paul Maison de Qualite Fondee en 1889	[France]	NaN
10	Rose Bakery Culture	[La Maison Rouge (10 boulevard de la Bastille)...	5.5
11	Boulangerie Lepot jean christophe	[23 rue Daval (Rue de la Roquette), 75011 Pari...	NaN
12	Boulangerie Martin	[40 rue Saint-Louis en l'Île, 75004 Paris, Fra...	NaN
13	Aux 2 Anges	[23 Rue Daval, 75011 Paris, France]	NaN
14	La Tradition du Pain	[23 rue Daval (Rue Saint-Sabin), 75011 Paris, ...	NaN

Finally, we plot the bakeries we have discovered on a map centered around the district. We use the following colors: in green the bakeries with a high rating (≥ 8.0), in orange the other rated bakeries, and in black those that do not have any rating.



On this small example, we observe that a small proportion of bakeries are highly rated (only 1 in 15 in our example is rated above 8.0), while most of the bakeries are not rated. These proportions need to be confirmed on a larger scale, but this is not surprising: people do not usually put comments on each and every place they go, but only on places they found truly exceptional, or at least unusual, or for which they had a high expectation, which represent only a minority of them.

Going forward we will split the bakeries in two categories:

- The "**top**" bakeries, with a rating equal to or above 8.0
- The "**ordinary**" bakeries, either with no rating or with a rating strictly below 8.0

METHODOLOGY

Now that we have familiarized ourselves with the data, we can design a methodology to determine good candidate areas to open our high-standing bakery.

1. The first step will be to ***retrieve the full list of bakeries*** with their geo-localization and their rating.
2. The second step will be to ***divide Paris*** into small, ***hexagonal-shaped***, areas of equal size. For each of those areas we will estimate the 3 metrics that will be considered in our decision: the population density, the average distance of any point in the area to the closest "top" bakery, and the number of bakeries per inhabitant.
3. The final step will consist in grouping the areas into ***different clusters according to their similarities*** on the 3 characteristics determined in the second step. We will then determine the cluster of areas that present the ***best mix of characteristics***.

To confirm the results, we will ***assign a score*** to each area according to how it performs on the three different measures, and eventually conclude on which areas seem to us the most suitable to welcome a new top bakery.

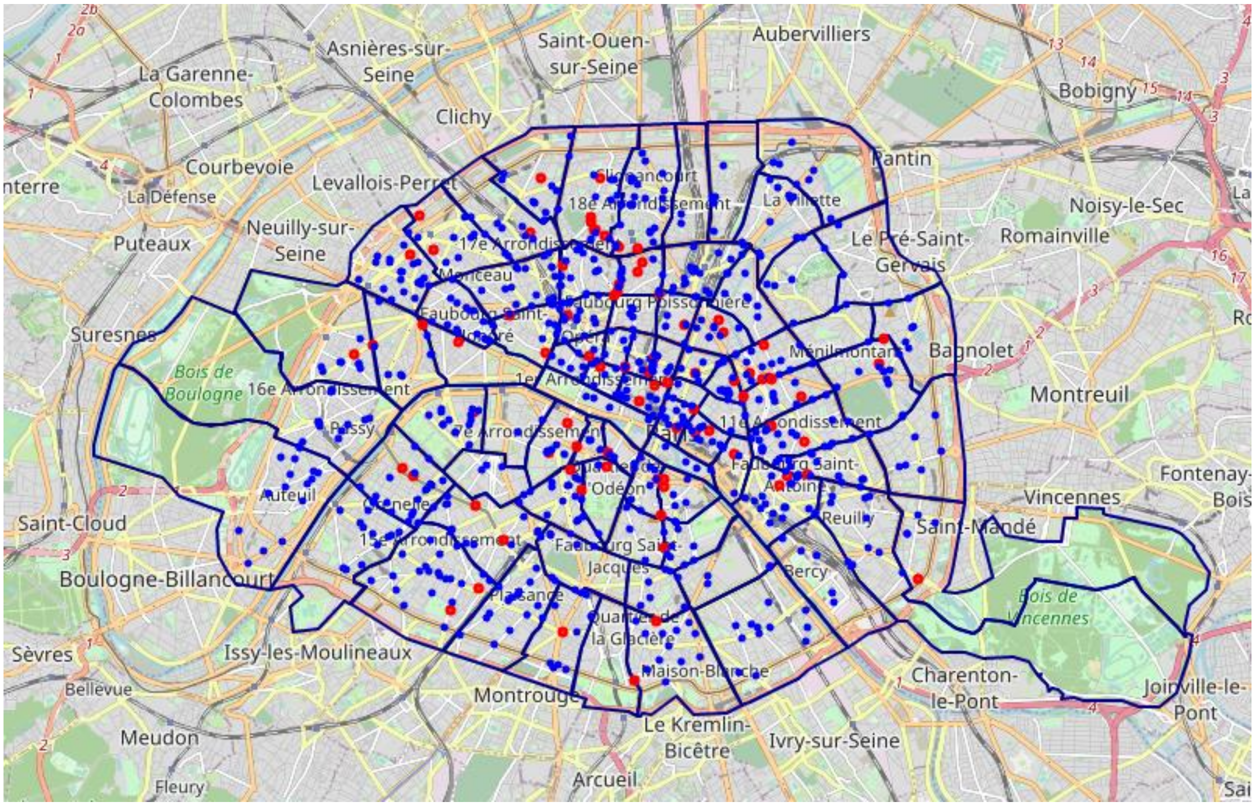
ANALYSIS

Complete list of bakeries

To limit the number of bakeries retrieved each time we request Foursquare, we will fetch them district by district. Each time, we will define the search area as the smallest circle that completely covers the shape of the district (by defining the radius as the largest distance between the center of the district and any of the vertexes of the polygon defining the shape of the district), and make sure that each bakery retrieved is really inside the district (in order not to double count bakeries that are inside a district and close enough to another district).

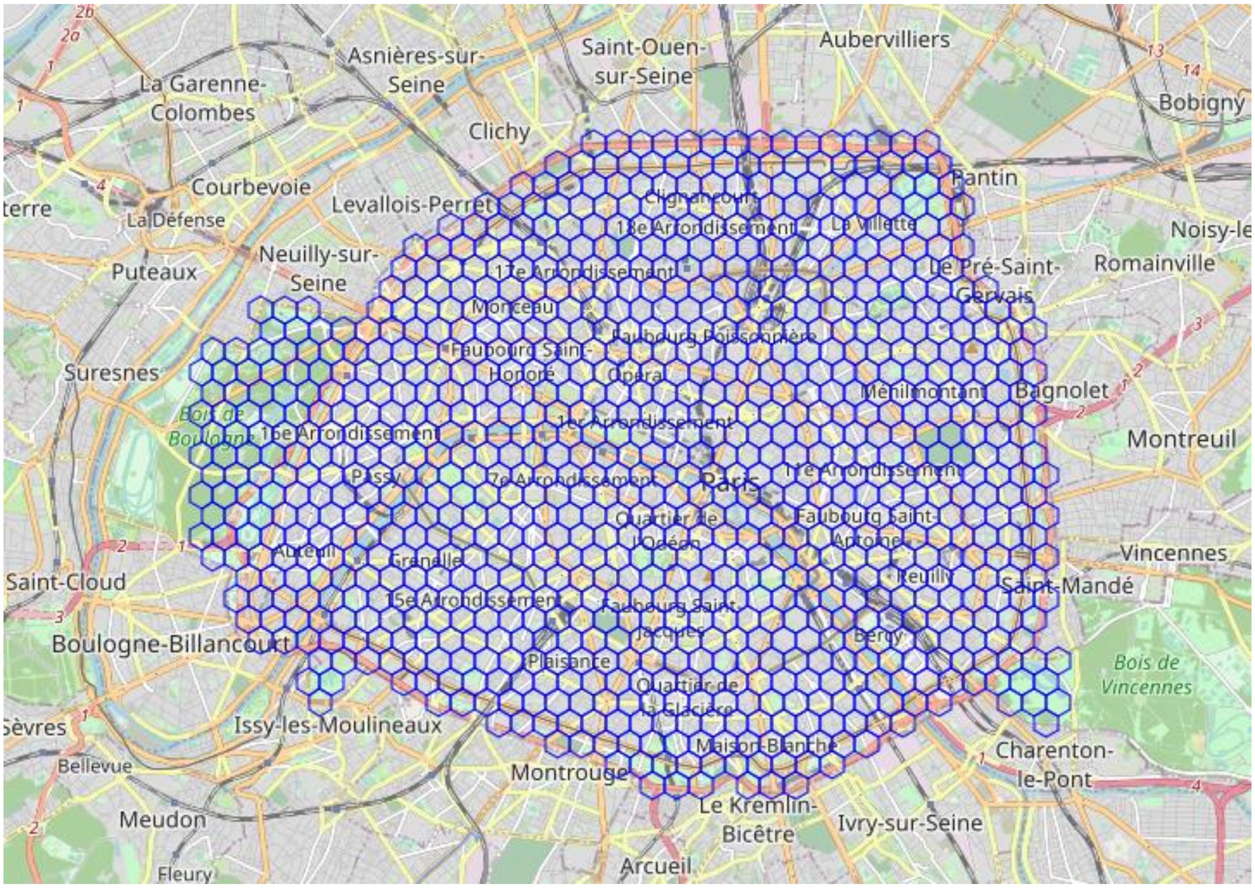
In total we have retrieved 804 bakeries in Paris, according to Foursquare. Only 82 of them, or about 10%, received a rating equal to or greater than 8.0. We identify those bakeries as "top" bakeries and copy them into a dedicated `DataFrame`.

Below is a visualization of all the bakeries on a map, with the "top" bakeries in green, and the others in blue (either not rated or rated below 8.0):



Division of Paris into hexagons

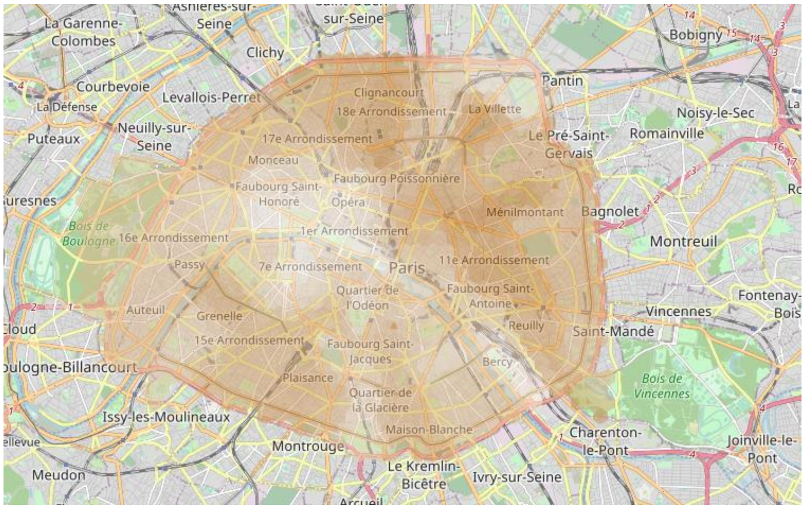
In this step we want to create a patchwork of hexagonal shapes that completely covers Paris. We choose 200m as the hexagon side, each shape therefore covering a surface of about 0.1 km^2. The shapes look like this:



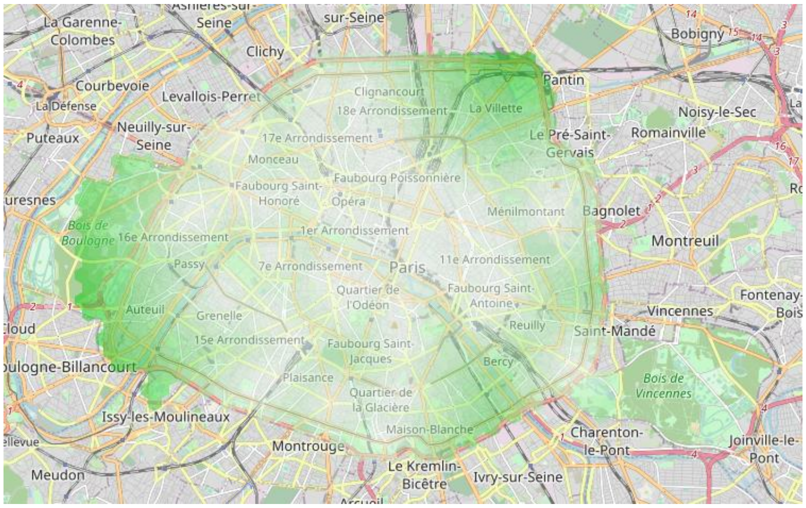
We have used 898 hexagonal shapes to cover Paris, and we need to attribute to each area the 5 following features:

- The **average population density**: considering the 7 points consisting in the 6 vertexes and the center, we determine the districts where the points are (on peripheral hexagons, some vertexes might be outside any district) and we compute the average of the corresponding densities.
- The **number of bakeries inside** the area: we simply count the number of bakeries located within the hexagon.
- The **number of bakeries around** the area: we count the number of bakeries inside all the 6 (or less for a peripheral hexagon) areas surrounding the considered hexagon.
- The **average distance to a top bakery**: considering the 7 points consisting in the 6 vertexes and the center, we compute the average of the distance between each point and the closest "top" bakery.
- The **number of bakeries per inhabitants**: it is calculated as `nb_bakeries_around / population`, where the population is the density multiplied by the surface of a hexagon in km^2.

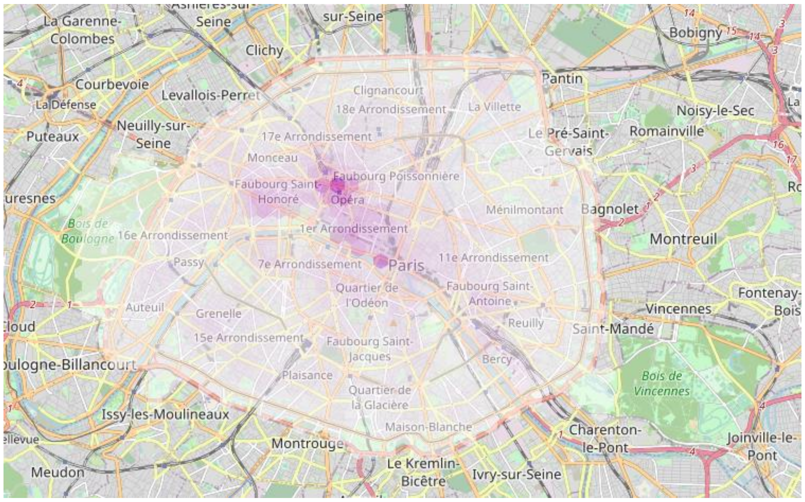
Once we have calculated everything, we can plot the results and get the following maps:



Hexagonal shapes colored according to the population density



Hexagoonal shapes colored according to the average distance to a “top” bakery

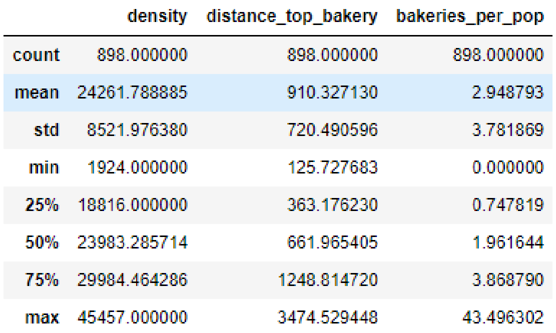


Hexagonal shapes colored according to the number of bakeries per inhabitants

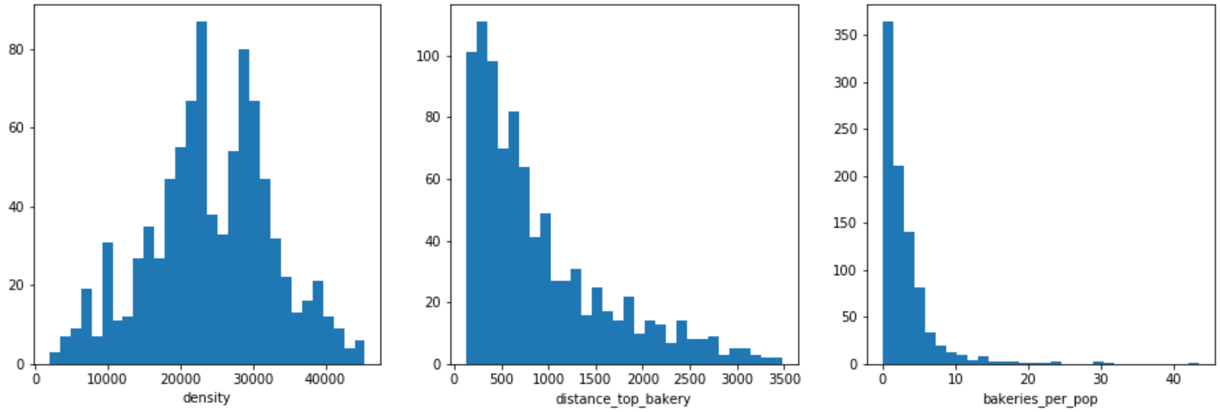
Clustering

Now that we have designed a patchwork of hexagonal shapes covering Paris, and we have computed the 3 metrics that we want to consider in the decision, we want to group the hexagons together according to their similarities on those 3 features. For this, we will use a clustering method, which is a machine learning unsupervised algorithm.

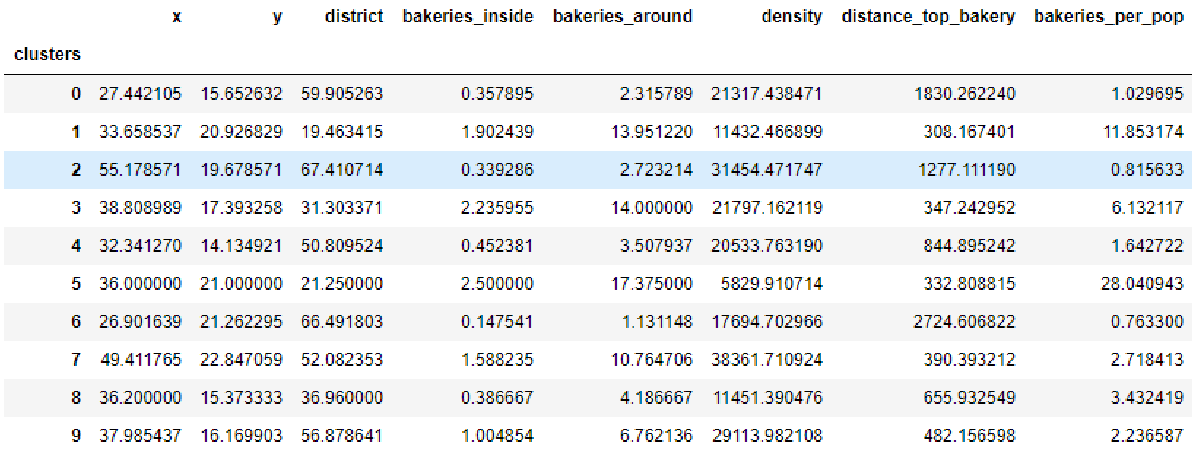
Let us first analyze the distribution of the 3 metrics over the whole set of hexagons:



	density	distance_top_bakery	bakeries_per_pop
count	898.000000	898.000000	898.000000
mean	24261.788885	910.327130	2.948793
std	8521.976380	720.490596	3.781869
min	1924.000000	125.727683	0.000000
25%	18816.000000	363.176230	0.747819
50%	23983.285714	661.965405	1.961644
75%	29984.464286	1248.814720	3.868790
max	45457.000000	3474.529448	43.496302



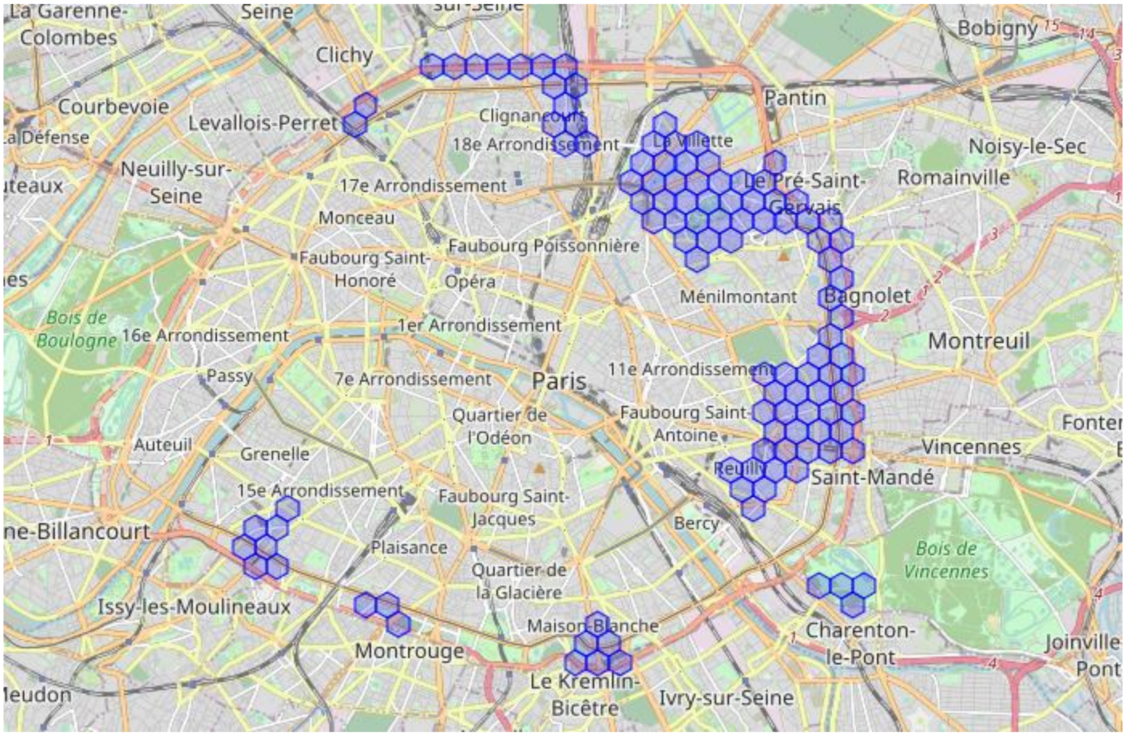
The three features span over quite different ranges, so we need to standardize them before running the clustering algorithm. We will use the **K-means algorithm** to group the areas into **10 clusters**. The results are shown in the table below:



	x	y	district	bakeries_inside	bakeries_around	density	distance_top_bakery	bakeries_per_pop
clusters								
0	27.442105	15.652632	59.905263	0.357895	2.315789	21317.438471	1830.262240	1.029695
1	33.658537	20.926829	19.463415	1.902439	13.951220	11432.466899	308.167401	11.853174
2	55.178571	19.678571	67.410714	0.339286	2.723214	31454.471747	1277.111190	0.815633
3	38.808989	17.393258	31.303371	2.235955	14.000000	21797.162119	347.242952	6.132117
4	32.341270	14.134921	50.809524	0.452381	3.507937	20533.763190	844.895242	1.642722
5	36.000000	21.000000	21.250000	2.500000	17.375000	5829.910714	332.808815	28.040943
6	26.901639	21.262295	66.491803	0.147541	1.131148	17694.702966	2724.606822	0.763300
7	49.411765	22.847059	52.082353	1.588235	10.764706	38361.710924	390.393212	2.718413
8	36.200000	15.373333	36.960000	0.386667	4.186667	11451.390476	655.932549	3.432419
9	37.985437	16.169903	56.878641	1.004854	6.762136	29113.982108	482.156598	2.236587

One cluster immediately stands out: the **cluster #2**. It has the 2nd highest population density, ranks 2nd for the number of bakeries per inhabitant with 0.816 for 1,000 (or 1 bakery every 1,226 people), and 3rd for the distance to a top bakery (1,277m in average).

Here is where the hexagons belonging to cluster #2 are located on the map:



RESULTS

At this point, here is what we have achieved:

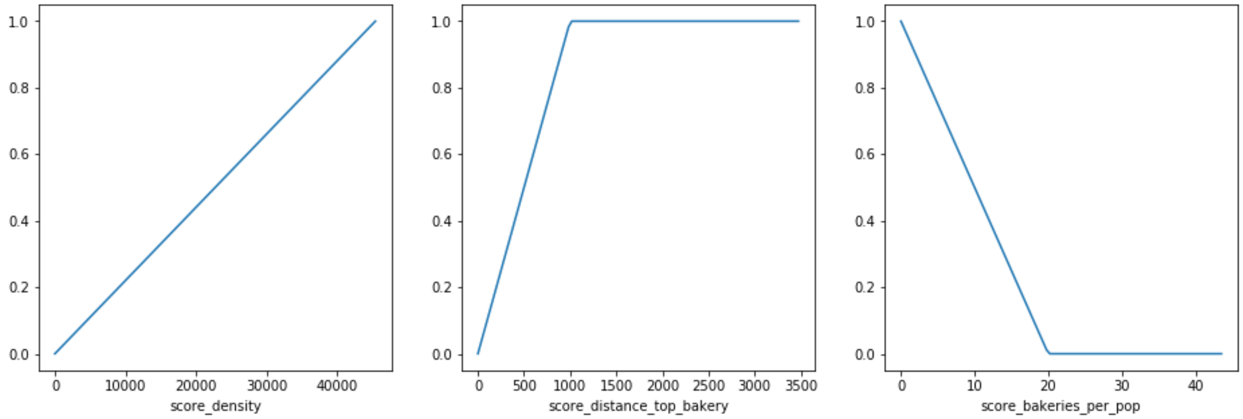
- we have covered Paris with a patchwork of small hexagonal shapes
- we have computed for each hexagonal shape a few metrics that we want to consider in our decision: the average population density, the number of bakeries around the area and the number of bakeries per inhabitant.
- we have clustered the hexagonal shapes according to how similar they look along those 3 metrics and have identified one particular cluster that shows interesting scores in average on all 3 metrics.

In this section, we want to confirm quantitatively what we found in our analysis. Also, as cluster #2 is still quite populous, we want to go further in identifying a handful of hexagons in this cluster that are the most suitable to open our top-tiered bakery.

To this end, we will assign a score between 0 and 1 to each hexagon according to each feature , using the following scoring functions:

- The score for the **density** should be fully linear: the more the population, the more attractive the area, so we assign each area a score **directly proportional to the value of the density**, with the densest hexagon receiving 1.
- For the **distance to a top bakery**, it should make no difference if the closest top bakery is far or very far away: if it's **far enough** that we can't go there by foot (let's take 1000m as threshold), **the score should be 1** (if we have to take the car, driving 1 or 3 km makes no difference). Below this threshold, the score is linear, with 0 corresponding to a distance of 0.
- For the **number of bakeries per inhabitants,** the score should be **inversely proportional** to the number of bakeries; however, beyond a certain threshold, there are just too many bakeries and it makes no sense opening a new one in the area, so **the score should be 0**. We will use 20 as threshold (which corresponds to 1 bakery every 50 persons).

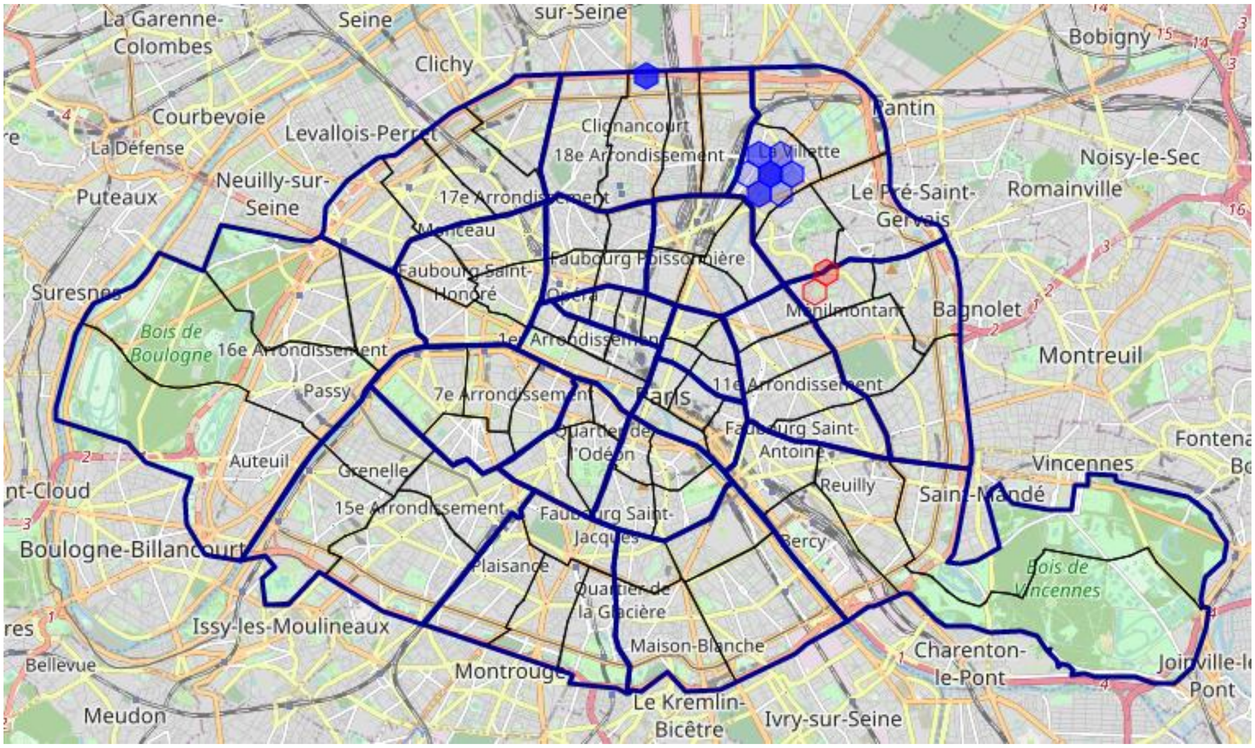
Graphically, the scores look like this:



Having implemented the scoring functions above, we can finally compute the total score and rank the hexagons. The 10 best areas are listed in the table below:

x	y	density	distance_top_bakery	bakeries_per_pop	clusters	score_density	score_distance_top_bakery	score_bakeries_per_pop	total_score
54	28	41718.000000	1824.576153	0.922624	2	0.917746	1.000000	0.953869	2.871615
43	33	39243.000000	1202.971214	0.245203	2	0.863299	1.000000	0.987740	2.851039
53	27	41718.000000	1479.405036	1.383936	2	0.917746	1.000000	0.930803	2.848550
53	29	41718.000000	2017.519234	1.383936	2	0.917746	1.000000	0.930803	2.848550
56	28	41718.000000	2009.207142	1.383936	2	0.917746	1.000000	0.930803	2.848550
55	29	41718.000000	2170.098599	1.614592	2	0.917746	1.000000	0.919270	2.837017
59	23	40264.857143	950.043531	0.477960	7	0.885779	0.950044	0.976102	2.811925
55	27	38400.571429	1668.152062	0.751747	2	0.844767	1.000000	0.962413	2.807180
52	28	38305.714286	1679.872141	0.753609	2	0.842680	1.000000	0.962320	2.805000
58	22	44328.000000	849.940465	0.651225	7	0.975163	0.849940	0.967439	2.792543

Unsurprisingly, **8 out of the 10 highest scores are from cluster #2**, including the top 6. This result allows us to confirm quantitatively what we inferred from the clustering operation. Besides, we now have, among cluster #2 and overall, the best areas of Paris to open our bakery. Let's plot those areas on a map, in blue those from cluster #2, in red the others, and with the highest ranked being the most opaque, together with Paris boroughs and districts, to get an idea of where they are.



The result is striking: a cluster of 7 areas grouped together is clearly identifiable, with the area with the highest score in the center. This is definitively this area that we should choose to open our bakery. The 7 hexagons in the cluster all lie in one single administrative district; by using `geopy.geocoders` package, we can easily find out which district it is, and the exact address of the center of the shape.

```
In [29]:
```

This concludes our study:

Considering the 3 features *population density*, *number of bakeries around the area* and *average number of bakeries per inhabitants* the ideal location where to open a high-standing, traditional French bakery is in the 19th borough, district of *La Vilette*, more precisely around the address: *11 Passage de Flandre*.

DISCUSSION

The goal of this project was to find an optimal place in Paris to open a new high-standing, traditional French bakery in Paris. We have defined 3 criteria that could be quantitatively assessed for any place in Paris using only resources freely available on the internet. We have come to a conclusion that identifies a single area particularly interesting within a popular neighborhood, with an average number of bakeries both in absolute and relatively to the population pretty low compared to the rest of Paris.

Of course, this result should be taken with a grain of salt. The metrics that have been used are just 3 among numerous possible criteria both quantitative and qualitative. In particular, the entrepreneur should be particularly mindful of the following:

- **The specificities of the neighborhood**: is the address easily accessible by foot, or even better is it in or close to a pedestrian area? it is close to a marketplace, or at least to other local shops, especially other food businesses, that could attract customers? is it close to a subway station? is there a historical landmark nearby?
- **The type of population**: traditional Parisians in average like their bakeries but what about a district with a high level of immigration? Are North-African and Chinese people, just to name a few, really keen to purchase a baguette every day from the bakery?
- **Economics**: how much does it cost to settle down in a specific area? how much is the local population willing to pay for a croissant?

There are certainly many more questions to ask, but they go largely beyond the scope and the purpose of this project.

CONCLUSION

In this project, we have used data science methodology and the Python language to answer a real-life question: what are the most suitable places in Paris to open a new, traditional French bakery.

We have demonstrated that, using only free resources and tools easily available on the internet, we could perform a thorough analysis based on a limited number of criteria and come up to a conclusion as to what area was the most attractive. While it is clear that a serious entrepreneur should run a much more detailed analysis, considering more criteria, and using larger sets of data, the methodology in this study could still be appropriate.

Moreover, the method that we have used here should not be restricted to traditional bakeries in Paris, instead it could benefit any competitive business willing to open a new store or shop in a big city: restaurants, barber shops, cinemas, nail salon etc.