Capstone Project
The Battle of Neighborhoods

KARIM BOUCHEKOURA

Coursera | IBM Data Science Professional Certificate | May 2019

Business problem

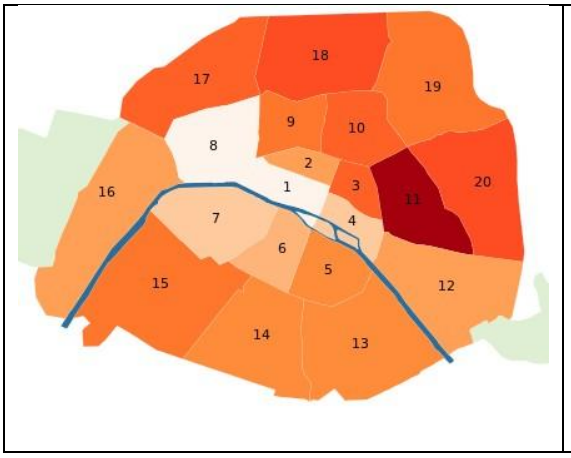
The objective of this report is to give an idea of "how data analysis can be a help to take a decision". The original questions, as we assume in the context of the final class, may have been asked by a restaurant chain: "***In a city of your choice, if someone is looking to open a restaurant, where would you recommend that they open it?***".

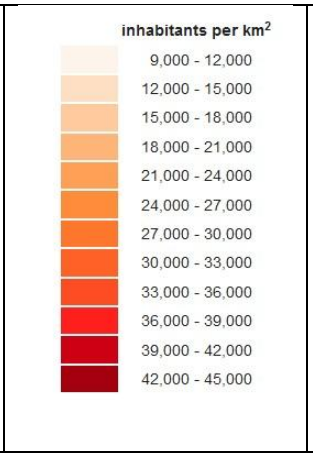
Between instinctive decision, and personal presumptions, we propose to see how data exploration, simple statistics and mathematics, as well as unsupervised machine learning algorithm help to take a decision.

Datas

A request for **neighborhoods of Paris** to FourSquare API give us the coordinates and names of those points:

We will also use **population density per borough** in Paris from Wikipedia





Arrondissement	Population per km²
1	9,457
2	22,740
3	31,364
4	17,532

Finally, thanks to reverse geocode we can request addresses in Paris from coordinates:
(48.830752,2.338778) -> https://opencagedata.com/ -> 14 Villa de Lourcine, 75014 Paris, France

Methodology

We start by creating a homogeneous grid of neighborhoods of Paris.
For every neighborhood, we will request coordinates, name and type of all the restaurants surrounding this neighborhood in a radius of 300 meters.
To every neighborhood, we will calculate and study the distribution of restaurants types in this neighborhood. This will lead us to separate the case of "French Restaurants", as we think that we should not advice to open a certain type of restaurant if this type is too numerous in the neighborhood.

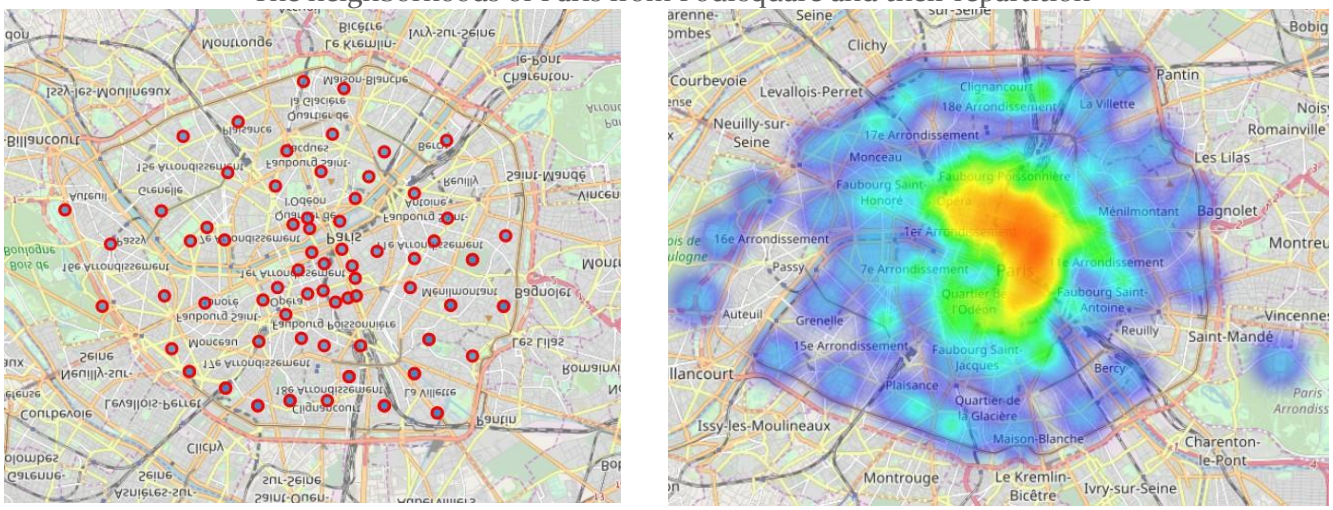
Every neighborhood will be associated with the shortest mean distance of all its restaurants to 20 monuments (as we would like to take tourism into account), population density of the borough it belongs to (as we believe a higher density of people would be better), and a grade reflecting the diversity of the types of restaurants it contains. A final grade for each neighborhood will reflect the sum of those 3 characteristics.

First, we will advise neighborhoods having a maximum score, regardless of the type of restaurant.
Secondly, we'll use Machine Learning algorithm K-Means to group neighborhoods by similarity of the distribution of the types of restaurants it contains. This mean that neighborhoods in the same cluster will show close proportions of types of restaurants.

Exploratory Data Analysis

We search for minimum and maximum latitudes and longitudes of the coordinates of all neighborhoods given by Foursquare, in order to create a homogeneous grid of neighborhoods:

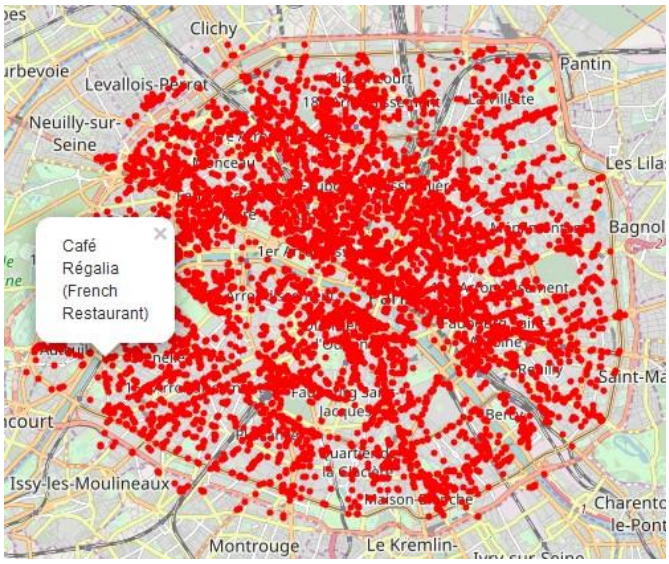
The neighborhoods of Paris from Foursquare and their repartition



The grid of neighborhoods we built to work with, homogeneously distributed

Note that the shortest distance between 2 neighborhoods of our grid is 558 meters.

The request of all restaurants in each neighborhood gives us more than 6 000 restaurants in Paris.



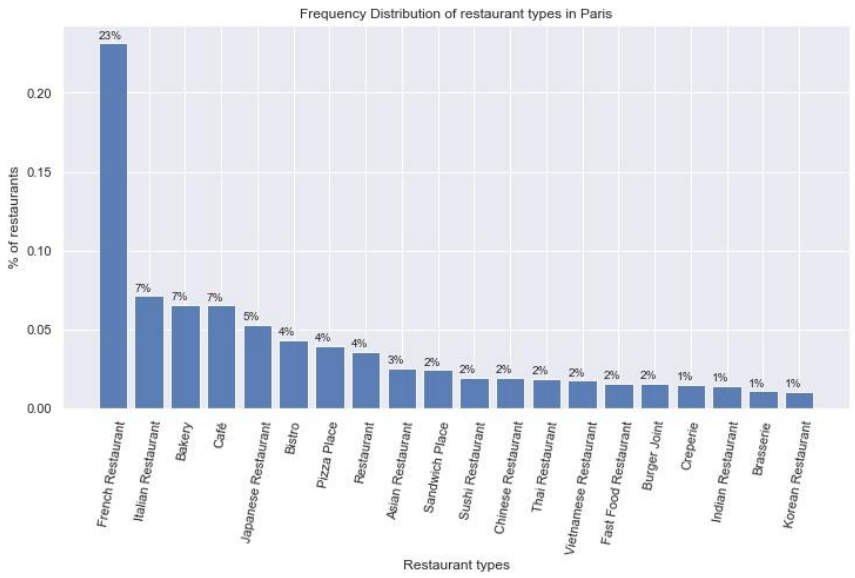
6485 restaurants of 129 different types

- 323 neighborhoods (Red points) of 129 different types
- Colors show restaurant density
- 20 monuments (**Green disks**)

We join addresses, postal code to our 323 neighborhoods dataset, then population density:

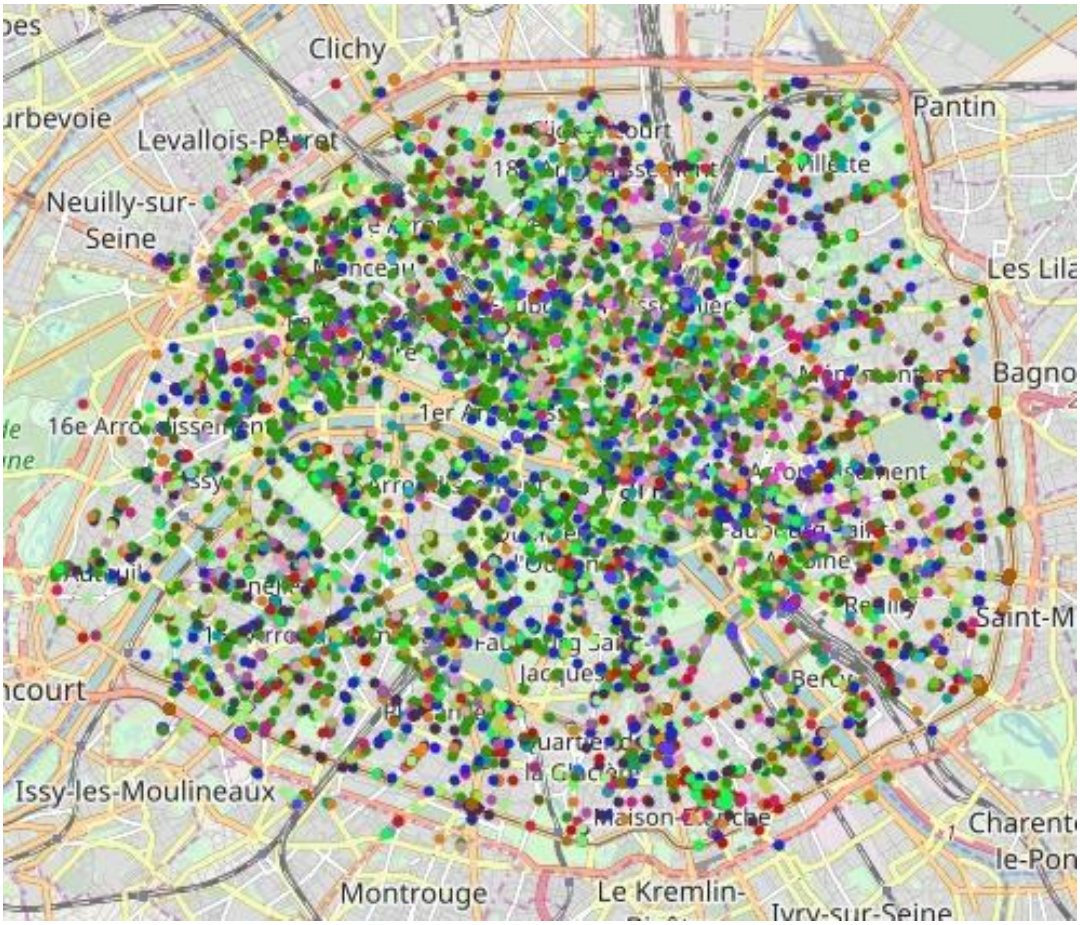
	Address	Latitude	Longitude	Name	Postal code	Density
108	53 Avenue de Ségur, 75007 Paris, France	48.849812	2.308265	N-109	75007	14228
142	3b Rue de l'Alboni, 75016 Paris, France métrop...	48.857436	2.285381	N-143	75016	21698
130	68 Rue Saint-André des Arts, 75006 Paris, France	48.853624	2.338778	N-131	75006	20499
131	2 Quai du Marché Neuf, 75004 Paris, France	48.853624	2.346406	N-132	75004	17532
107	7 Rue de la Cavalerie, 75015 Paris, France	48.849812	2.300637	N-108	75015	28314

Simple statistics

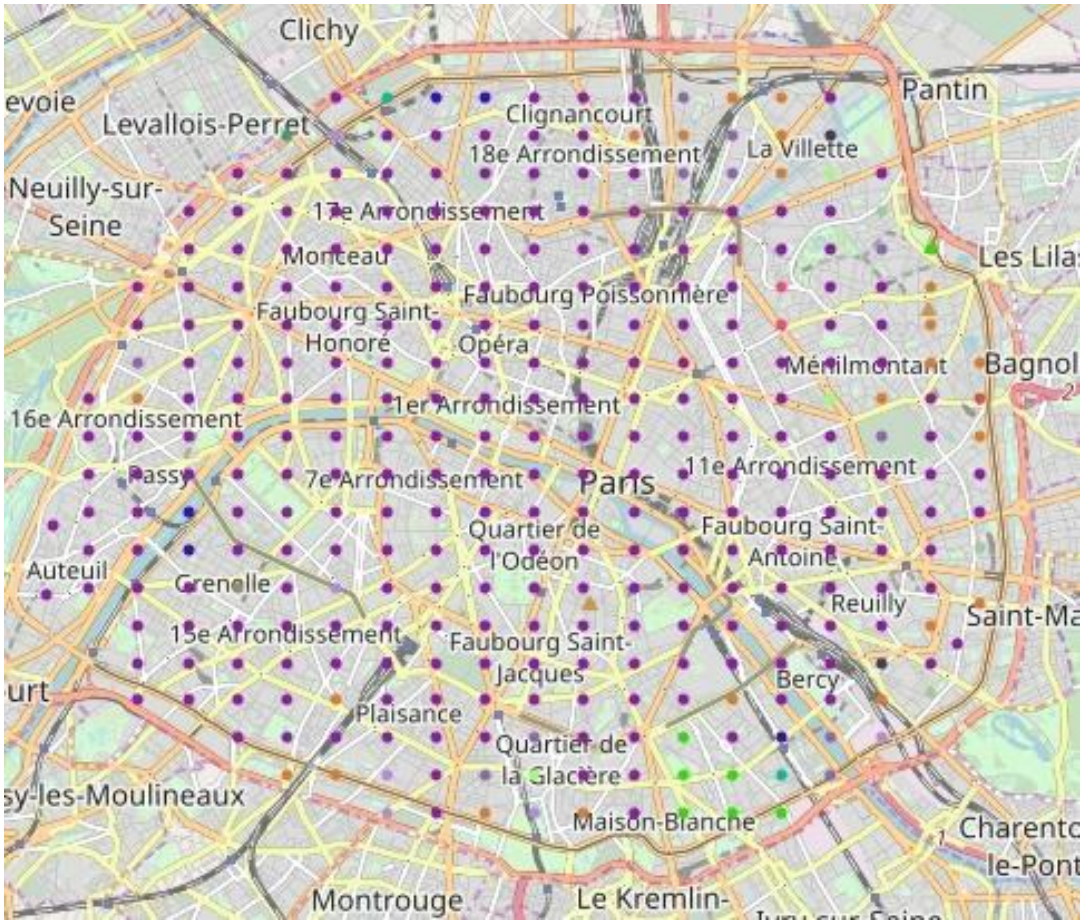


French Restaurant	1500
Italian Restaurant	462
Bakery	427
Café	423
Japanese Restaurant	344
Bistro	282
Pizza Place	257
Restaurant	234
Asian Restaurant	166
Sandwich Place	158
Sushi Restaurant	126
Chinese Restaurant	126
Thai Restaurant	123
Vietnamese Restaurant	114
Fast Food Restaurant	103
Burger Joint	102
Creperie	97
Indian Restaurant	93
Brasserie	71
Korean Restaurant	68

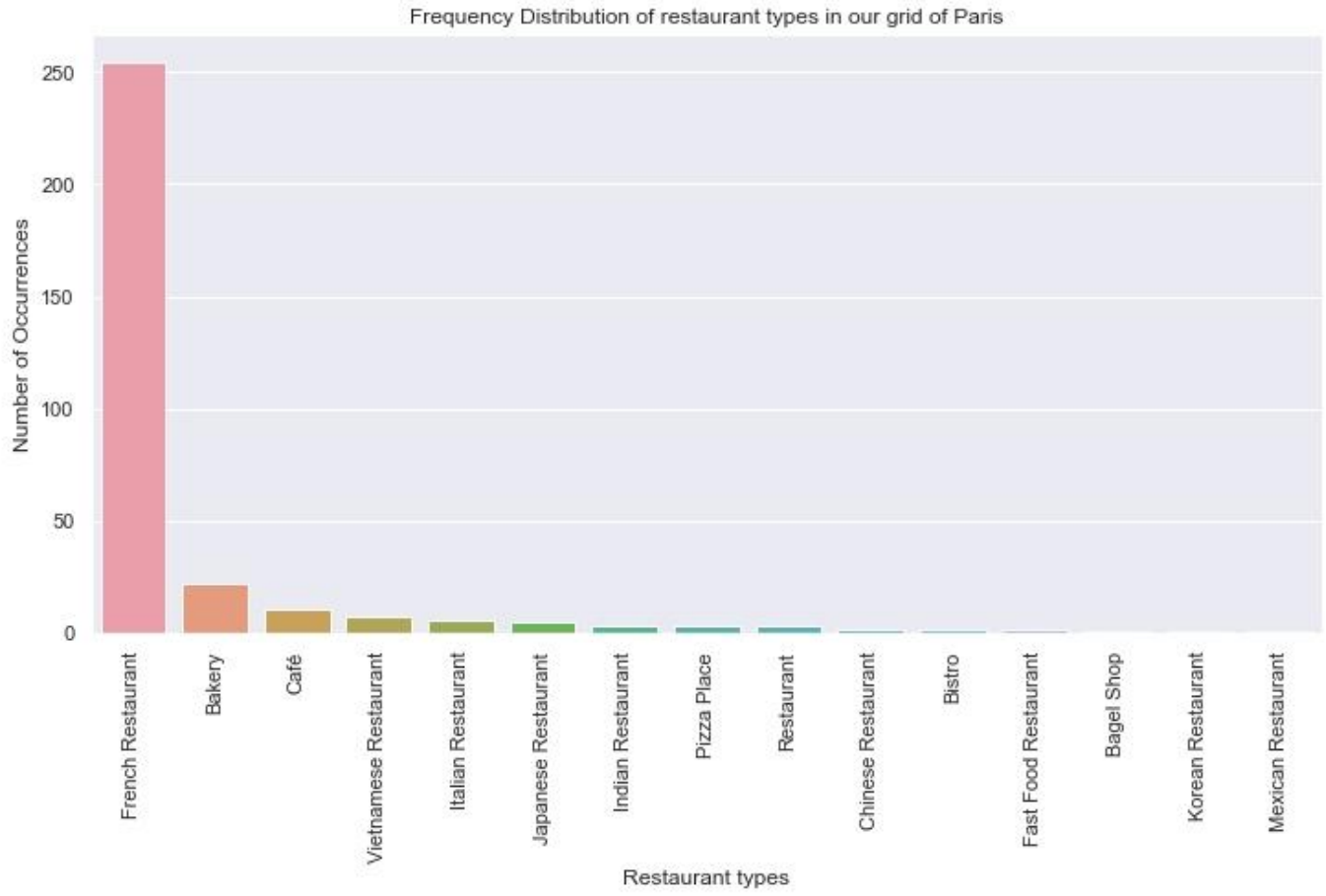
A map of restaurants colored by type is hardly readable because of the large number of french restaurants



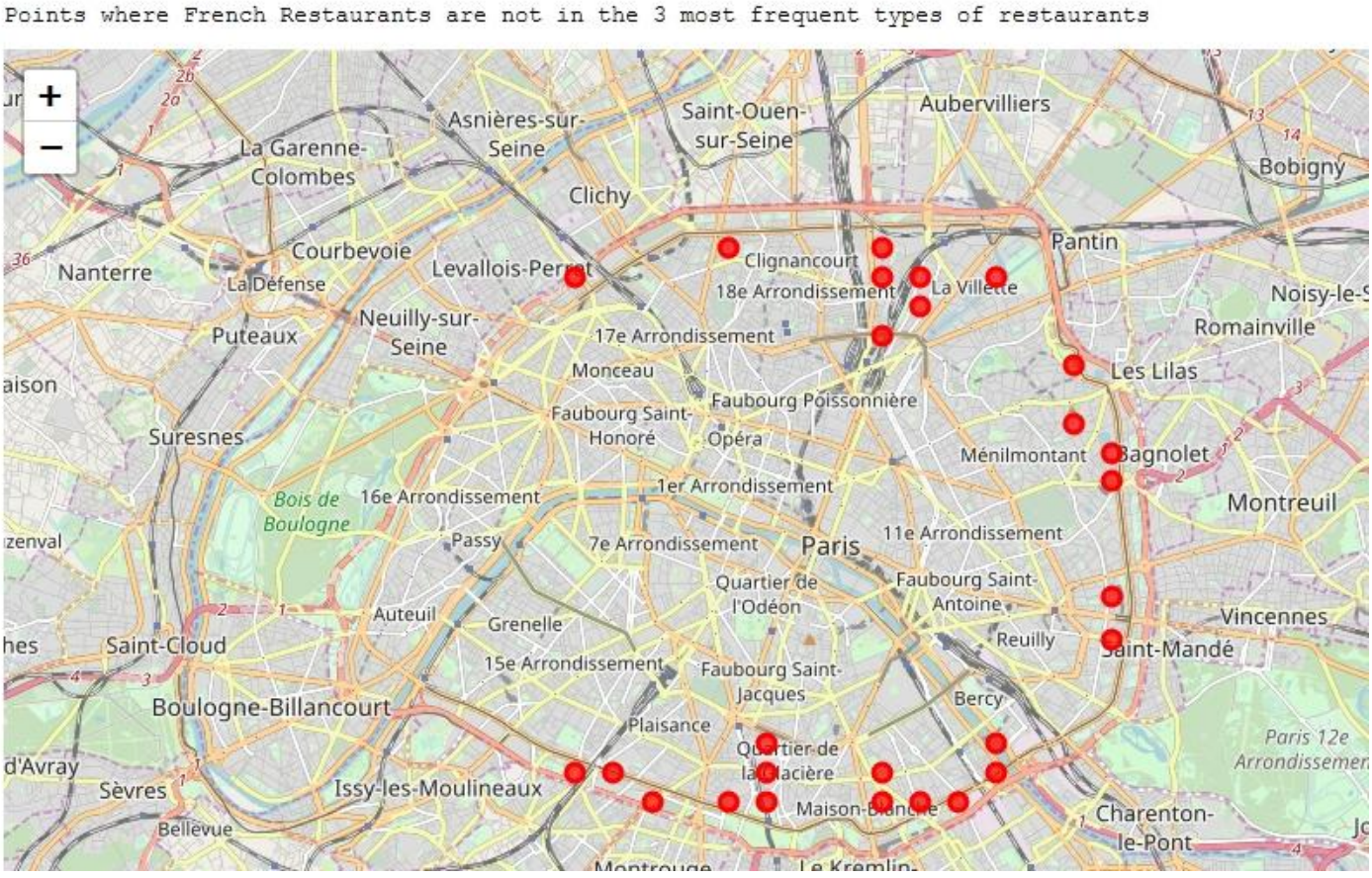
French restaurant is the first type of restaurant in most of our neighborhoods.



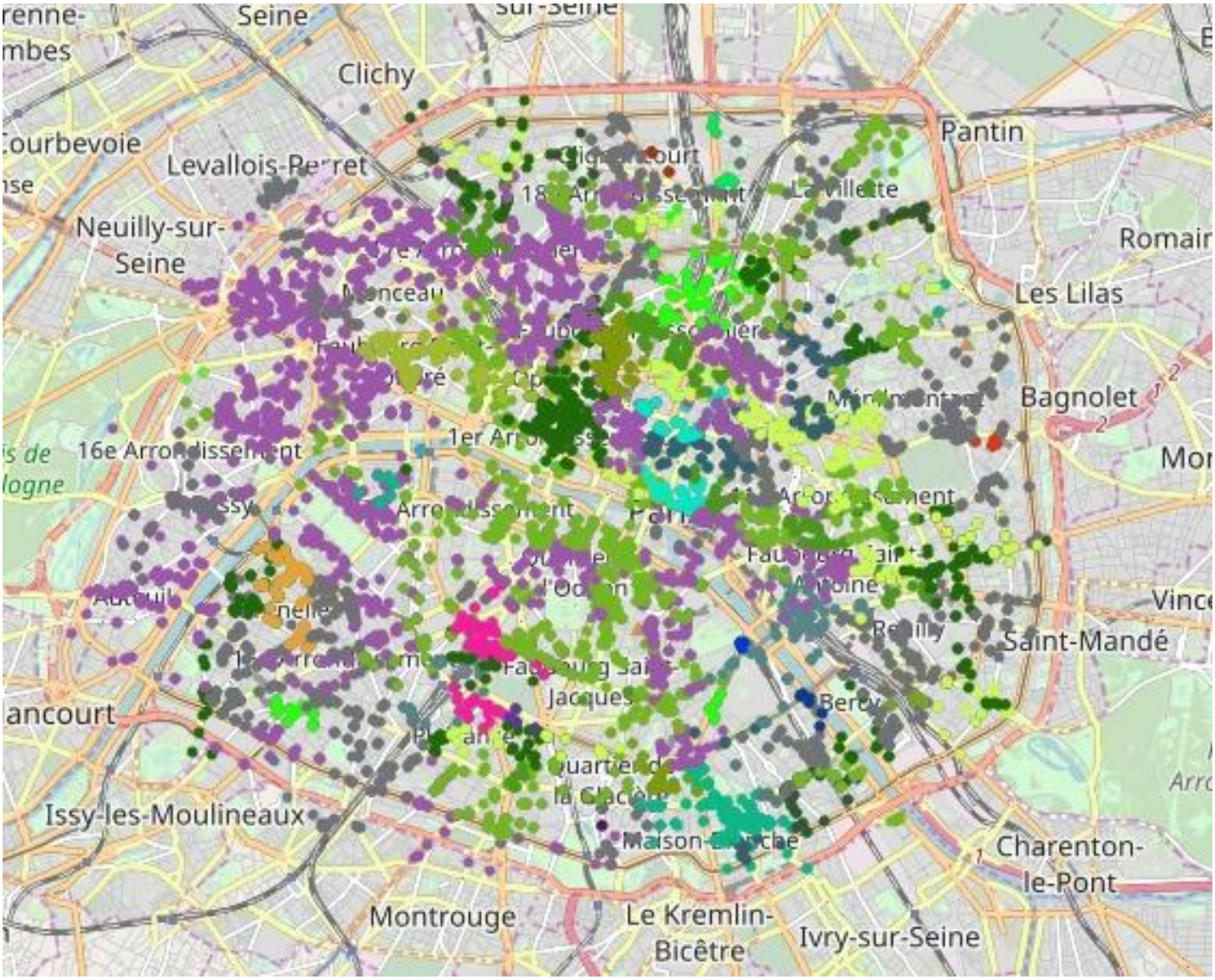
We start to think that opening a French restaurant may not be a good idea



In fact, French restaurants are not the majority in only 27 neighborhoods over 323



If French restaurants are removed from our set, restaurant types distribution become a little clearer



As we are exploring those informations, several questions arise:

- What means “good” when we say: a “good place” for a restaurant ?
- What kind of measure should we build to consider density of population, diversity of restaurants or proximity to a monument ?

Method 1: Advices neighborhoods that maximize a grade function

Each neighborhood is graded from 0 to 5. The grade 5 is given to neighborhoods with great density of population, great diversity of restaurants and which are located close to a monument in Paris.

Our grade function shall be the sum of

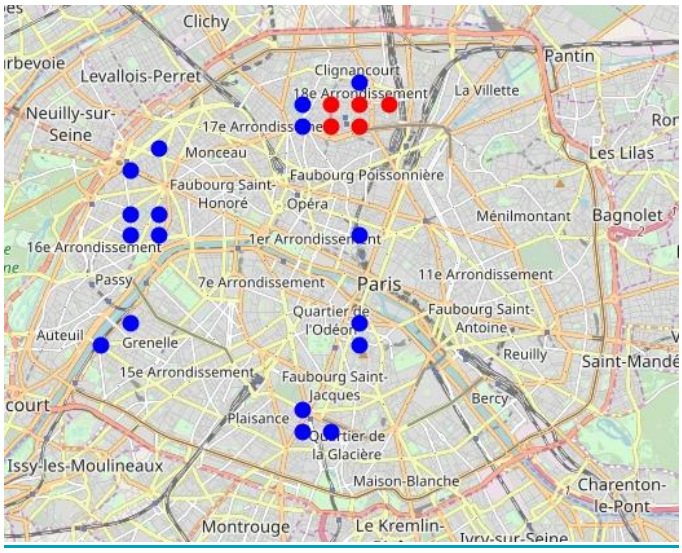
- **Diversity score**: (from 0 to 1): we set a score to each neighborhood, that should measure the diversity of the restaurant types surrounding it. We’re assuming that there’s less risk to open a restaurant in a neighborhood where lots of types of restaurant can be met. To do that we calculate the mean of the number of existing types of restaurant among the 10 first most numerous types of restaurant in every neighborhood, the we calculate the mean over every neighborhood, that is to say the mean in Paris. Then if a neighborhood has a diversity mean greater than the Paris’s mean, set its score to 1 (as many types are presents) or 0 if its mean is less than Paris’s mean.

- **Proximity to a monument score**: (from 0 to 2): We assume that a restaurant has more chance to perform if it's close to a monument, as it can benefit from tourism. We set a score that way: for every restaurant, we calculate its closest distance to any of 20 monuments in Paris. Then we calculate the mean of every neighborhood as the mean over every restaurants surrounding it. Paris' mean is the mean over every neighborhood. If a neighborhood's mean is in the first third-quantiles, set its score to 2 (close to a monument), if it's in the second, set it to 1, if it's in the third third, set its score to 0.
- **Density of population score**: (from 0 to 2): We assume that a restaurant has more chance to perform if it's located in a crowded borough. Here we just have to "bin" the density column in 3 parts, then assign 2 points to very dense neighborhoods, 1 point to dense neighborhoods, and 0 for others.

Neighborhoods with the highest grades of 5/5 and 4/5 are recommended.

As a result, we can advise **22 neighborhoods in Paris out of 323** based on diversity, proximity to monuments and population density

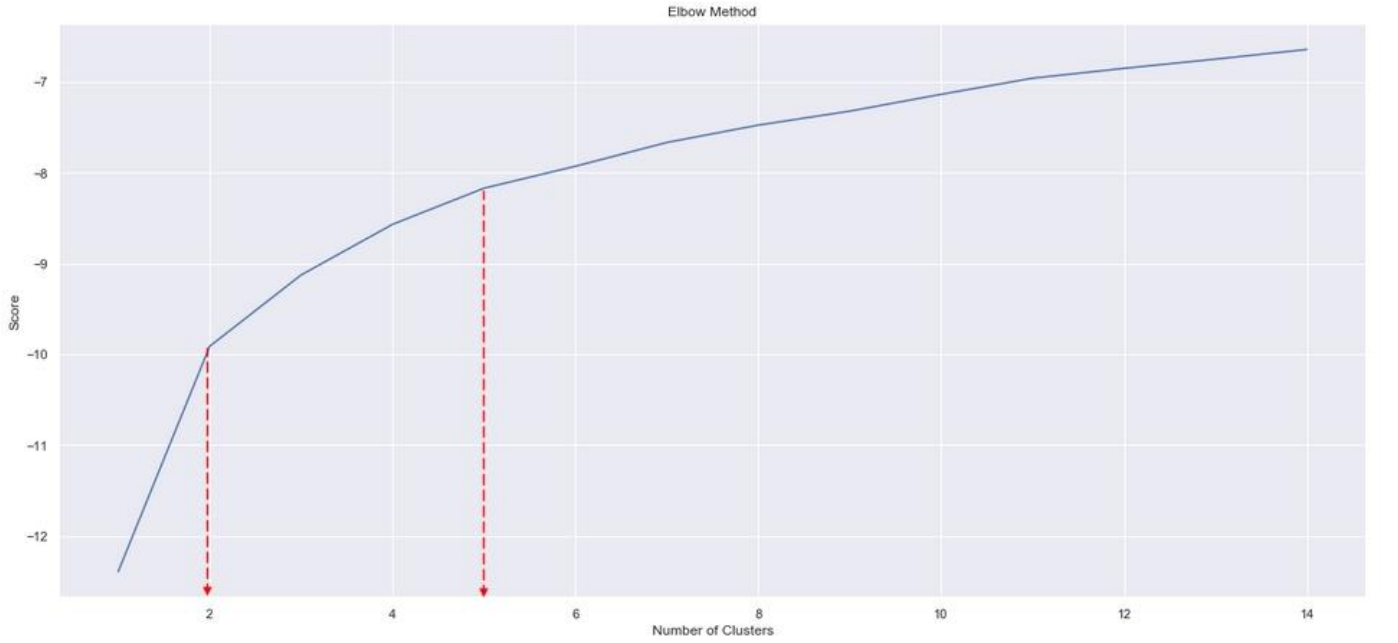
	Name	Postal code	Density	diversity_score	proximity_score	density_score	total_score
286	N-287	75018	33769	1	2	2	5
288	N-289	75018	33769	1	2	2	5
272	N-273	75018	33769	1	2	2	5
273	N-274	75018	33769	1	2	2	5
287	N-288	75018	33769	1	2	2	5
250	N-251	75017	30331	1	2	1	4
198	N-199	75016	21698	1	2	1	4
105	N-106	75015	28314	1	2	1	4
199	N-200	75016	21698	1	2	1	4
95	N-96	75005	24038	1	2	1	4
86	N-87	75015	28314	1	2	1	4
180	N-181	75016	21698	1	2	1	4
181	N-182	75016	21698	1	2	1	4
113	N-114	75005	24038	1	2	1	4
285	N-286	75018	33769	1	1	2	4
42	N-43	75014	25358	1	2	1	4
271	N-272	75009	27670	1	2	1	4
27	N-28	75014	25358	1	2	1	4
26	N-27	75014	25358	1	2	1	4
300	N-301	75018	33769	1	1	2	4
188	N-189	75002	22740	1	2	1	4
233	N-234	75016	21698	1	2	1	4



Method 2: Run k-Means clustering

k-Means is an unsupervised Machine Learning algorithm.
For a given integer k > 1, k-Means gather records in a set into k clusters.
This method will group neighborhoods of Paris into clusters in the sense that 2 neighborhoods in the same cluster have similar restaurant types repartitions.

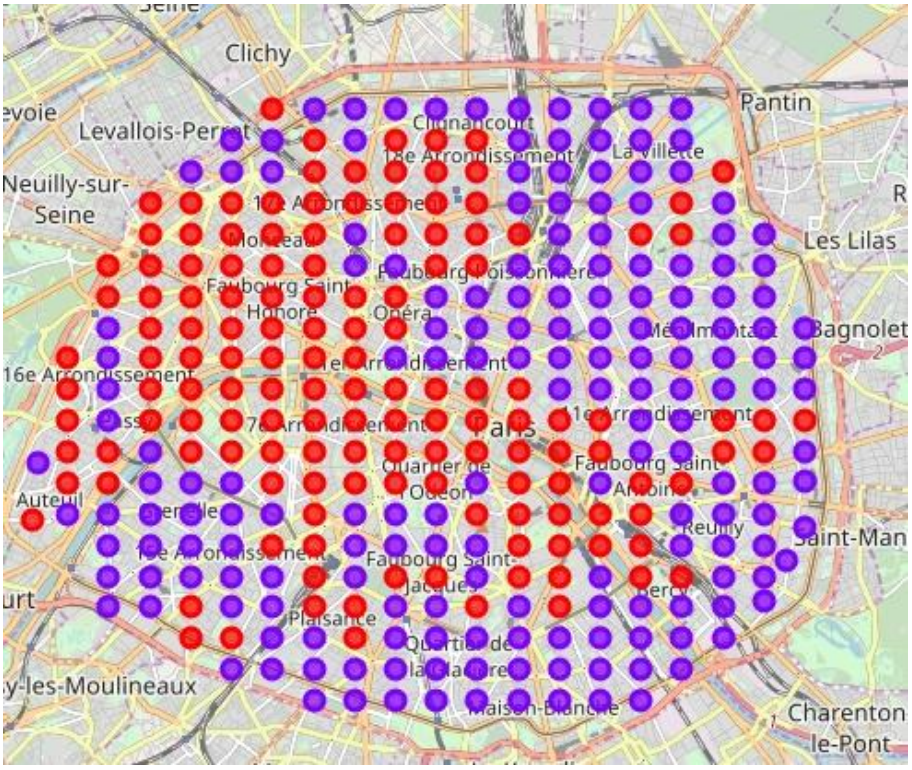
The elbow method leads us to optimum number of clusters: k = 2 and k = 5



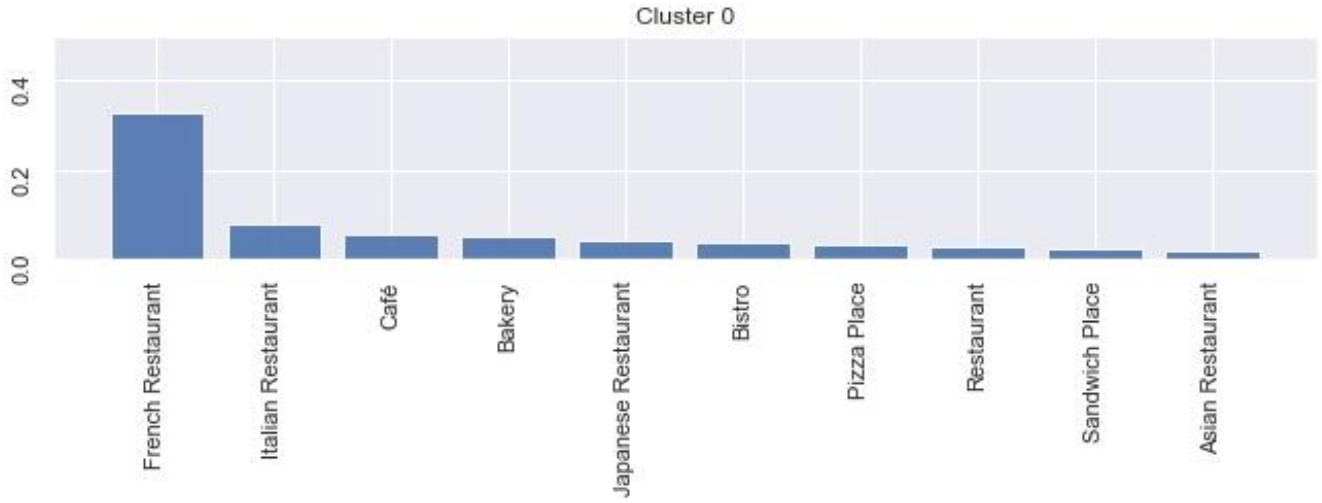
We will look at the clusters given by k-Means for k from 2 to 5. People knowing Paris will certainly recognize the designed clusters.

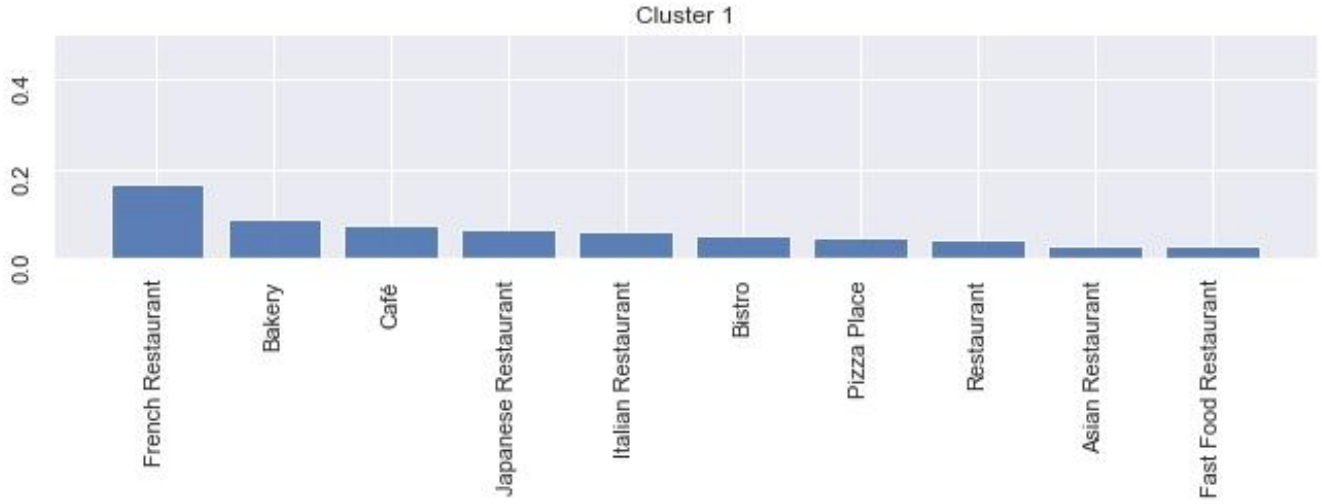
Note that k-Means does not provide any information about the clusters, we must then calculate the restaurant types distribution in each neighborhood.

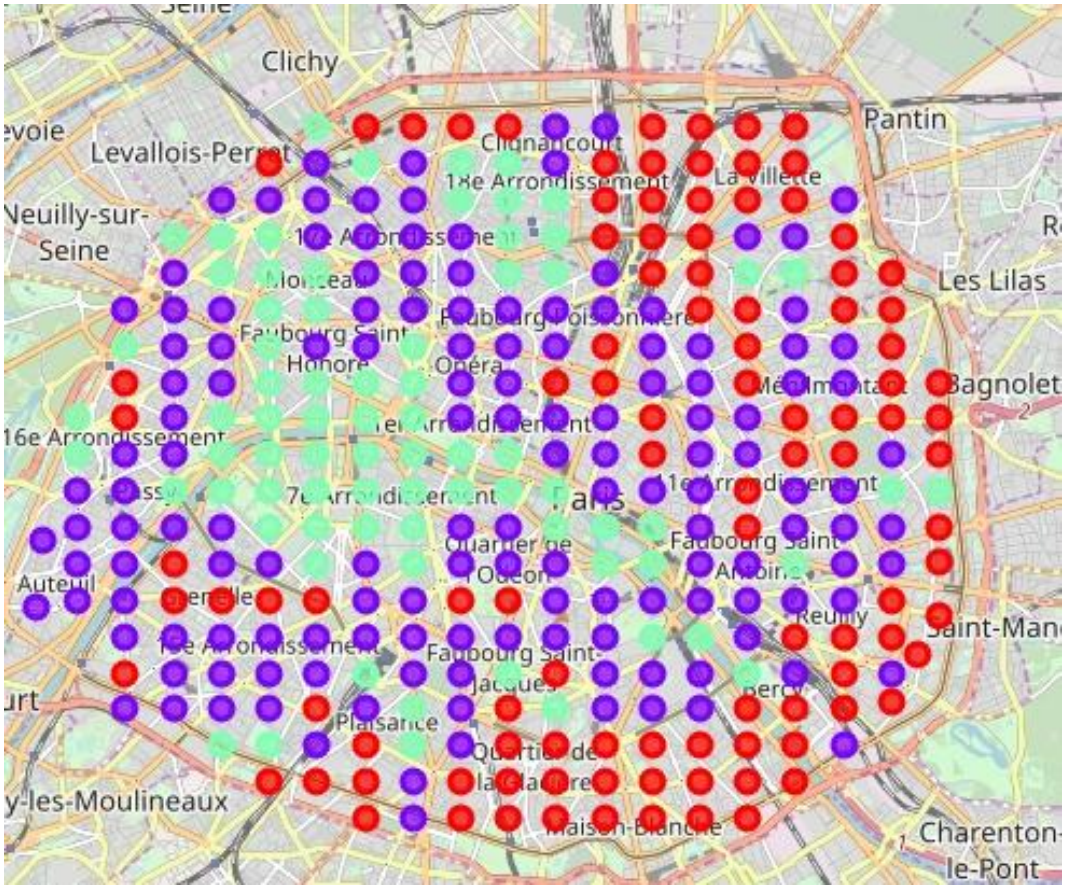
As the number of clusters increase, more precise regions in Paris will slowly show up



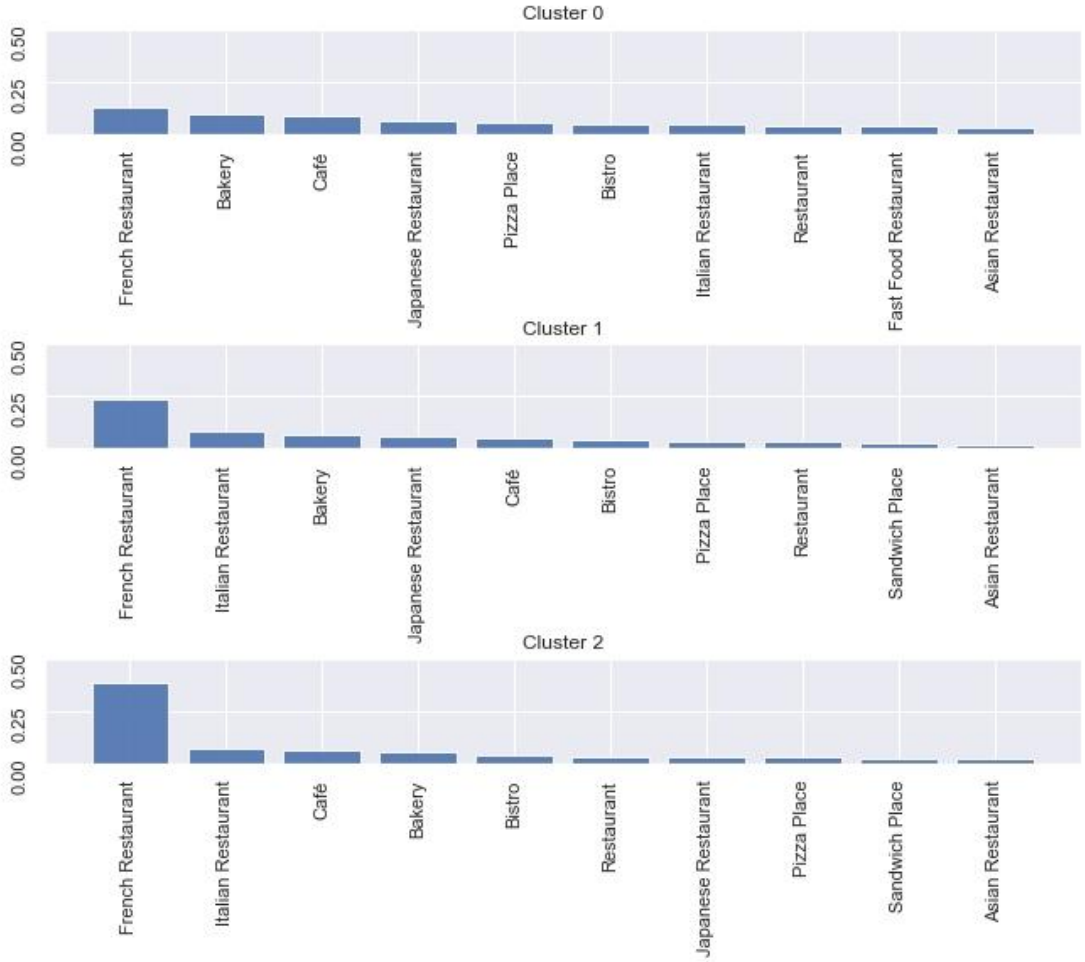
k-Means clusters for **k = 2**

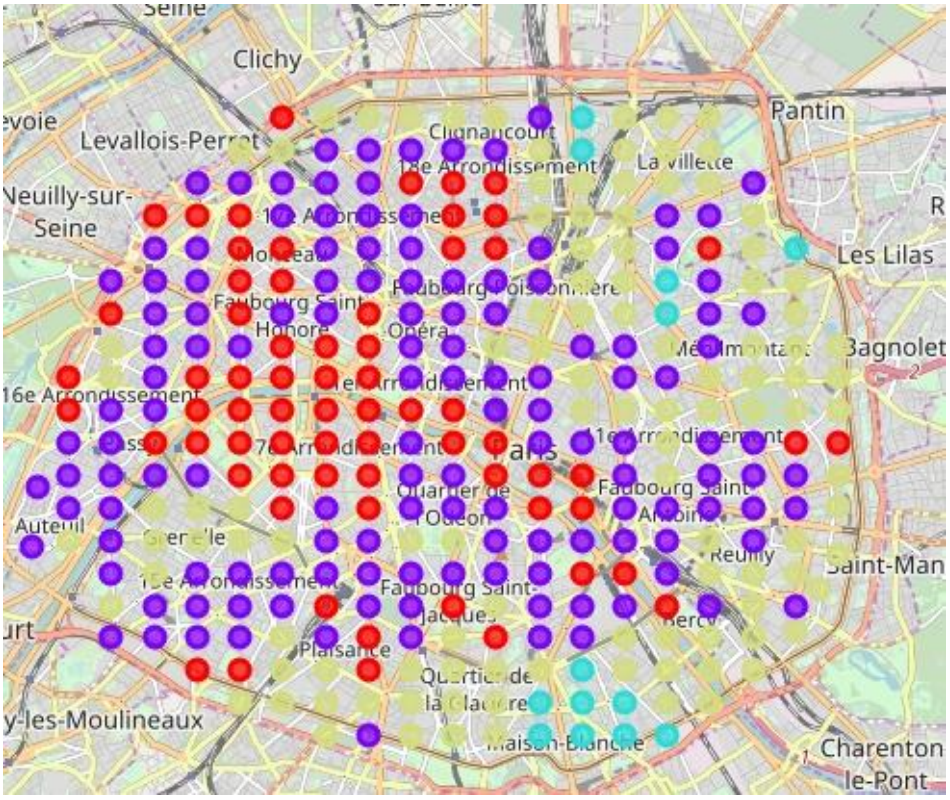




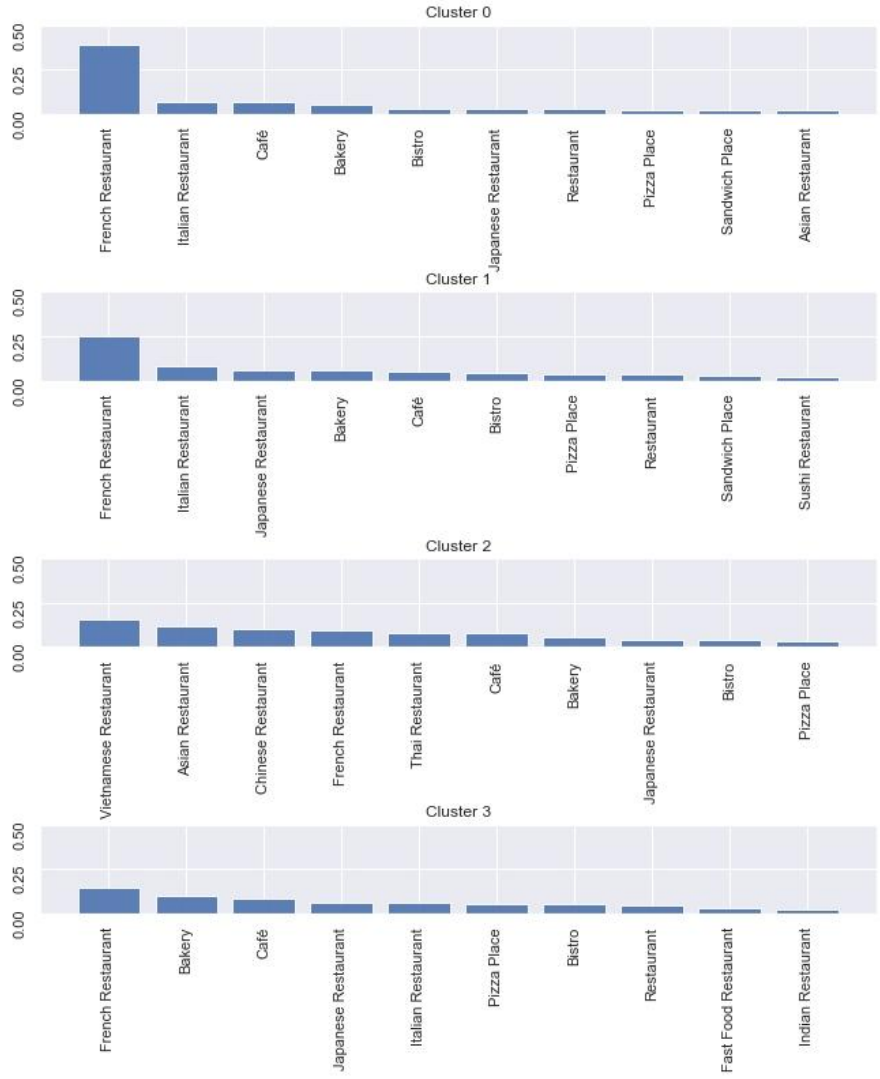


k-Means clusters for **k = 3**

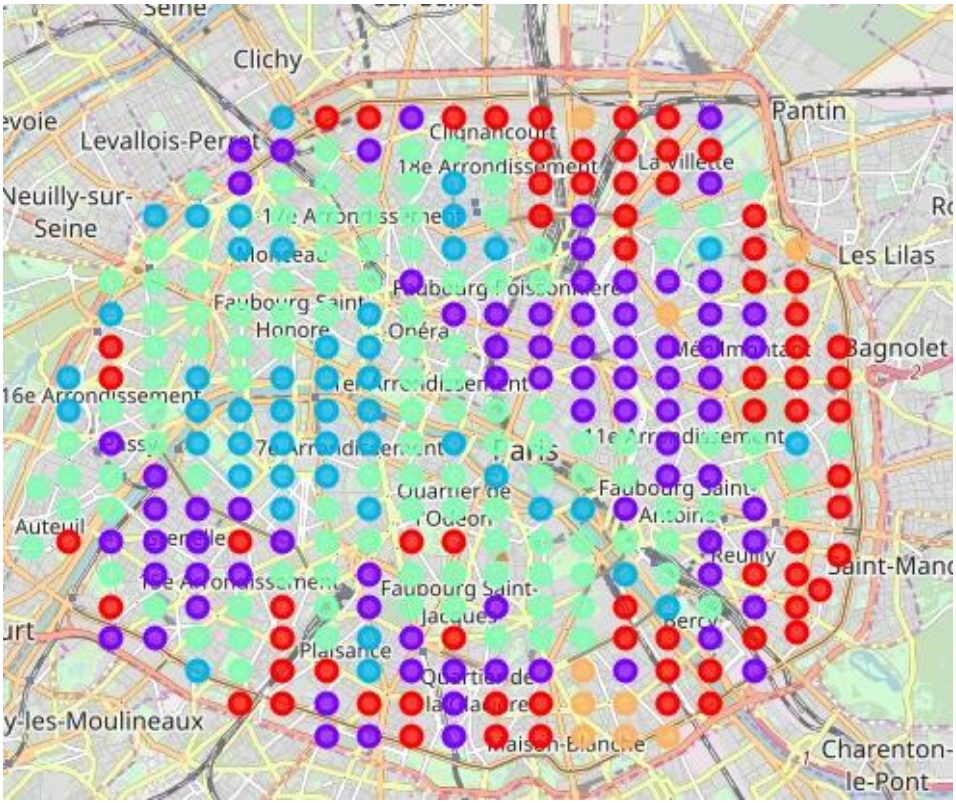




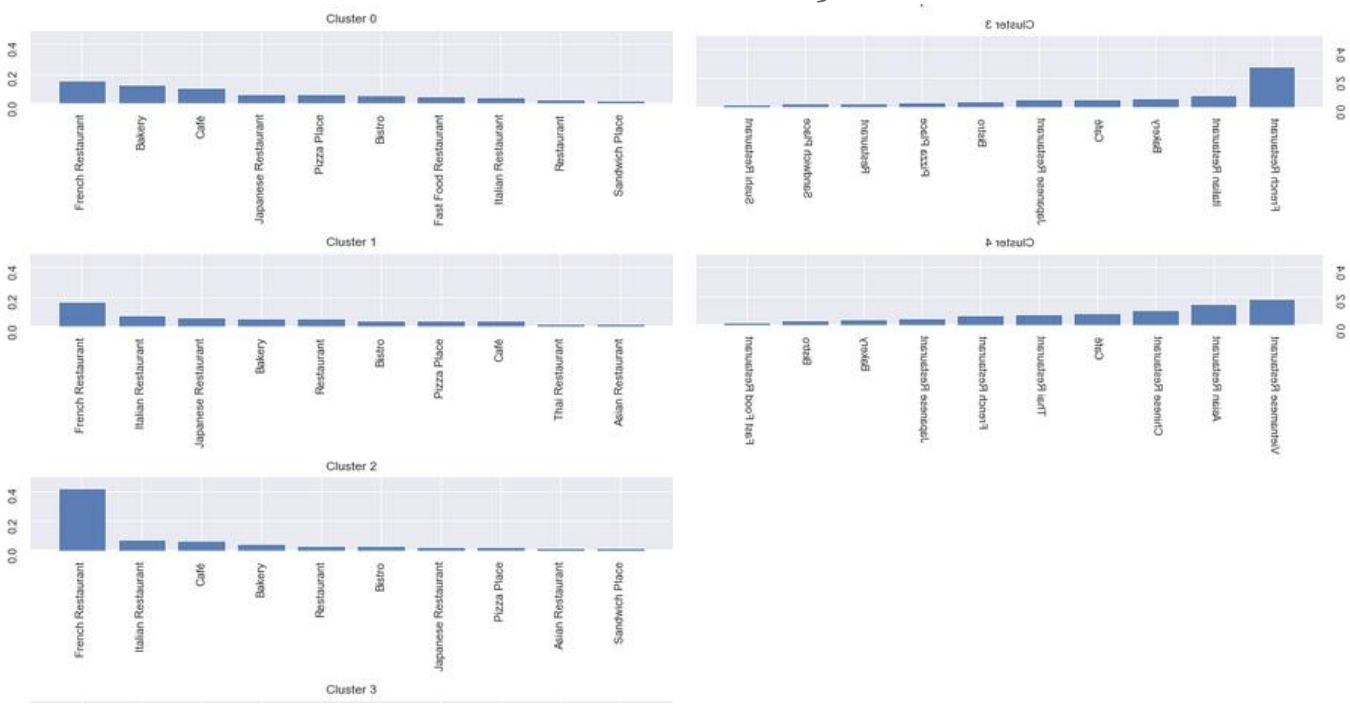
k-Means clusters for **k** = 4



Note that we have to drill to 5 clusters to see appear the Chinatown of the 13th arrondissement



k-Means clusters for **k = 5**



For instance, let's discover south east cluster 4 located in the 13th borough:

Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
48.830752	4	Vietnamese Restaurant	French Restaurant	Chinese Restaurant	Pizza Place	Asian Restaurant
48.82694	4	Vietnamese Restaurant	Asian Restaurant	Thai Restaurant	Chinese Restaurant	French Restaurant
48.823128	4	Vietnamese Restaurant	Japanese Restaurant	Café	Chinese Restaurant	Restaurant
48.880308	4	Vietnamese Restaurant	Bistro	Café	Asian Restaurant	Empanada Restaurant
48.823128	4	Vietnamese Restaurant	Asian Restaurant	Chinese Restaurant	Bakery	Café

Results

We can see that the results of a study highly depend on metric used to solve the problem.

The results given by the method 1 in which we maximize the grade reflecting diversity, proximity to monuments and population density are as follow:
We advise to search for a place to open a restaurant located less than 300 meter away from:

Address	Latitude	Longitude	Borough	Score /5
Rue de l'Abreuvoir, 75018 Paris, France	48.887932	2.338778	75018	5
30 Rue de Laghouat, 75018 Paris, France	48.887932	2.354034	75018	5
10 Rue des Abbesses, 75018 Paris, France	48.88412	2.338778	75018	5
4 Rue d'Orsel, 75018 Paris, France	48.88412	2.346406	75018	5
6 Passage Cottin, 75018 Paris, France	48.887932	2.346406	75018	5

The results given by the method 2 in which we group neighborhoods by restaurant types proportions give more informations and freedom to the stakeholder to choose what type of restaurants and where to open it. For instance , a Pizza place in one of the neighborhoods colored in red in last grid, as this type of restaurants is not widespread, if the budget allocated to the project is modest. Why not a Sandwich place or a Sushi Restaurant near the “Avenue des Champs-Élysées” located in cluster 3.

Discussion section

When you walk around Paris like in the 5th borough you discover that a restaurant can be very crowded in the same time as the next one is empty, so the question is hard to answer, and we can only base the results on assumptions like we did.

Note that shops and restaurants in Paris open and close very often, so our data may not be up to date. Many restaurants that actually do exist where not mentioned in our restaurants Dataset.

The two methods we used give complete different results as they did not take same assumptions.

Our grid could have been much denser, and we could not find any datas related to profit of each restaurant or types of restaurants, which would result in much more accurate conclusions.

More generally, the more characteristics we can find on our records, the more correlation we can involve in our solutions.

Conclusion

The reader will easily see that almost any kind of problem can scientifically apprehended. While browsing through our maps, one can have an idea of "where to open a restaurant in Paris", depending on what kind of restaurant he wishes to open, knowing that a lot more characteristics could be take into consideration.
In our problem, no quantitative information about our restaurant was the targeted object, so no supervised learning algorithms like logistic regression or neural networks could be implemented. We could only have implemented them if we had dataset of known "good" or "bad" places for a restaurant, or informations on their benefits.

The I Python Notebook can be found here:
https://github.com/KBouchek/Coursera_Capstone/blob/master/week5.ipynb

This report is the final report of last course from IBM Data Science Professional Certificate
https://www.coursera.org/specializations/ibm-data-science-professional-certificate

Many examples of Python codes can be found on:
https://github.com/KBouchek or https://codekarim.com/

Do not hesitate to contact me:
https://www.linkedin.com/in/karim-bouchekoura-84238572/