# Data science capstone – My Italian restaurant grand opening in Paris

## Table of contents

# 0 - Introduction

Italian restaurants are at the moment very trendy places in Paris: a lot of concept restaurants which propose Italian cuisine have recently opened in the city and compete to propose the most authentic experience of what an Italian "Trattoria" should be: affordable prices, Italian cooks and waiters, authentic decoration. For instance, to get a chance to dine at the well-named "Popolare" you must first queue outside for sometimes hours (even in winter).

Let's say that I am getting rid of this silly idea of working in data science and that I would like to open such a restaurant in Paris. The question is: in which area of the city should I open my restaurant to maximize my chances of success?

- As there are already many restaurants of this kind, I first want to make sure not to open my business in a saturated zone
- Yet, installing my restaurant in a residential area will not help me that much, as nobody would come by: it needs to be in some trendy area!
- A further requirement is that I would like to open just for dinner: people are stressed and in a hurry at lunch, while I want to propose some slow food experience (concepts like that sell very well)

In short, I want to find the evening-trendy place of Paris where there are the least Italian restaurants to open my restaurant.

# 1 - Description of the data

To find the perfect spot for a juicy business on the pizza planet, I intend to exploit the possibilities given by Foursquare API by following these steps:

1. Collect the venues data of the whole city (not just restaurants but all venues)
2. Collect the localization of the 80 neighborhoods of Paris
3. Classify the venues among their respective neighborhoods
4. Cluster the neighborhoods with the k-nearest method
    - Several values of k will be tested to retain the maximal value of k for which every cluster contains at least two neighborhoods
5. Analyze the data and identify the cluster which matches our targeted market ("trendy" means there are a lot of restaurants, bars, cultural life …)
6. In this cluster, select the half of the neighborhoods which concentrate the most trending venues at night
7. In this selection of neighborhoods, identify the one with the least Italian restaurants

# 2 –Methodology

The analysis declines upon the following steps. During the description of all steps, details will be given about the assumptions which were made and why.

## 2. 1 - Collection of the neighborhoods data

In a similar way that what has been done on the previous labs of the Data Science Capstone course, data about the neighborhoods of Paris were collected through the official open data source of the city of Paris at the following link:

https://opendata.paris.fr/explore/dataset/quartier_paris/information/

The data is collected as a .json file which contains much information for each of the 80 boroughs of Paris.

As already done for the previous labs, a data frame is created from the .json data.

Each line of the data frame contains:

- The borough names
- The neighborhood which is associated to the borough
- The coordinates of the center point of the borough:
    - Latitude
    - Longitude

## 2.2 - Collection of the venues data via Foursquare and duplicated removal

For each of the 80 boroughs, a request has been done through an automated function to Foursquare.

The parameters of the request were the following

- Radius: 1000 m
- Limit: 500

The radius value was deliberately set high to be sure not to miss venues relatively away from the center.

Because most of the neighborhoods are distanced from less than 1 km to their direct neighbors, it is expected that a lot of venues can be found twice (or even more) on the dataset.

To remove duplicates the following analysis has been performed:

- For each venue, the distance between the venue and the center of the assigned neighborhood was computed and stored in the data frame as a new column;
- Duplicated venues were cut off the original data frame and copied in another data frame;
- Each venue of the "duplicate" data frame is compared to the reference data frame. The one that is closest to its neighborhood center is copied in the reference data frame.

## 2.3 - Data Clustering

As we would like to class neighborhoods among different types of profiles to make sure to target the right audience for our Italian restaurant, data clustering through a k means analysis seems to be a relevant method.

**Pre-processing**

In a similar way that what has been done on the course labs:

- The occurrence of all type of venues per neighborhood are listed in a new data frame, which will be used for clustering
- To help the visualization later, the 10 most common venues per neighborhood are listed in another data frame

**Clustering**

At this point it is difficult for oneself to decide which number of clusters should be specified as an input to the k-means function.

The choice has been done to consider that we do not want clusters which only contain one borough. This decision is motivated by the consideration that the data should not be over-segmented.

The analysis was then performed starting with two clusters. If all clusters contained more than one neighborhood, the number of clusters was incremented, and the computation was re-run.

## 2.4 - Clustering results interpretation

When the process of clustering of over, we must decide which cluster should be used for the rest of our study. As already said, the perfect candidate is the one which contain a high concentration of trendy places, for instance: bars, coffee shops, restaurants…

Two possibilities to isolate our perfect match:

- Make statistics on each cluster and find out the one which has the highest rate of what we have just defined as trendy
- Perform this analysis by simply looking at the 10 most common values for each neighborhood in each cluster

The latter option was chosen, this part is highly subjective anyway so let us jump into subjectivity and trust our sense of business!

## 2.5 - Isolating trending places among the cluster

Among all the kept neighborhoods, we would now like to point out the ones in which there is some activity in the evening.

In order to find out another set of requests was made through Foursquare: this time the "trending" request was used for each of the neighborhoods of the kept cluster.

If specific venues around the specified location are particularly visited at the time of the request, they be returned as an output, if not the function returns an empty value.

There again, we would like to select relevant neighborhood but not be too sharp on the selection, as one last selection step is still ahead.

It was chosen to keep the 50% of neighborhoods in the cluster which have the highest number of trending venues.

To implement the criteria:

- The median value of trending venues per neighborhood is computed, it means that 50 % of neighborhoods have less (or equal) trending venues than the median, the 50 % remaining have more (or equal) trending venues
- Only neighborhoods which belong to the upper 50 % are kept

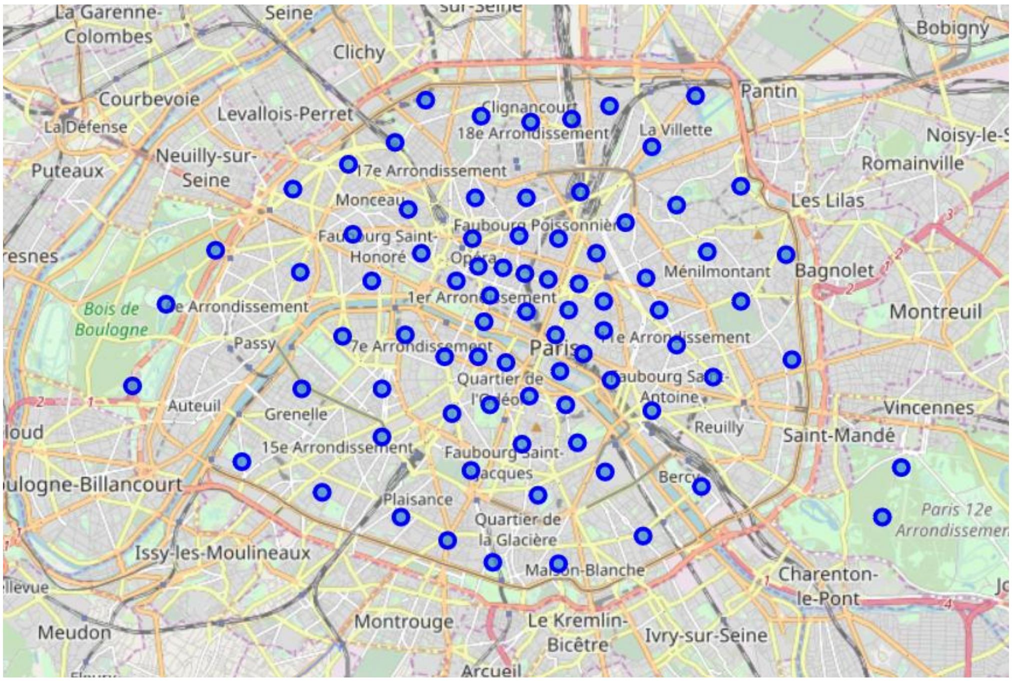
## 2.6 - Selecting the best neighborhood

To select the perfect fit for our business there is very last step to accomplish. Among the short list, we will select the neighborhood which possesses the least Italian restaurants.

- For that we compute the cumulative occurrence of both labels "Italian Restaurants" and "Pizza Place" in the corresponding data frame
- We select the minimal value among all and find out which neighborhood possesses it

# 3 - Result section

## 3.1 - Map of Paris neighborhoods



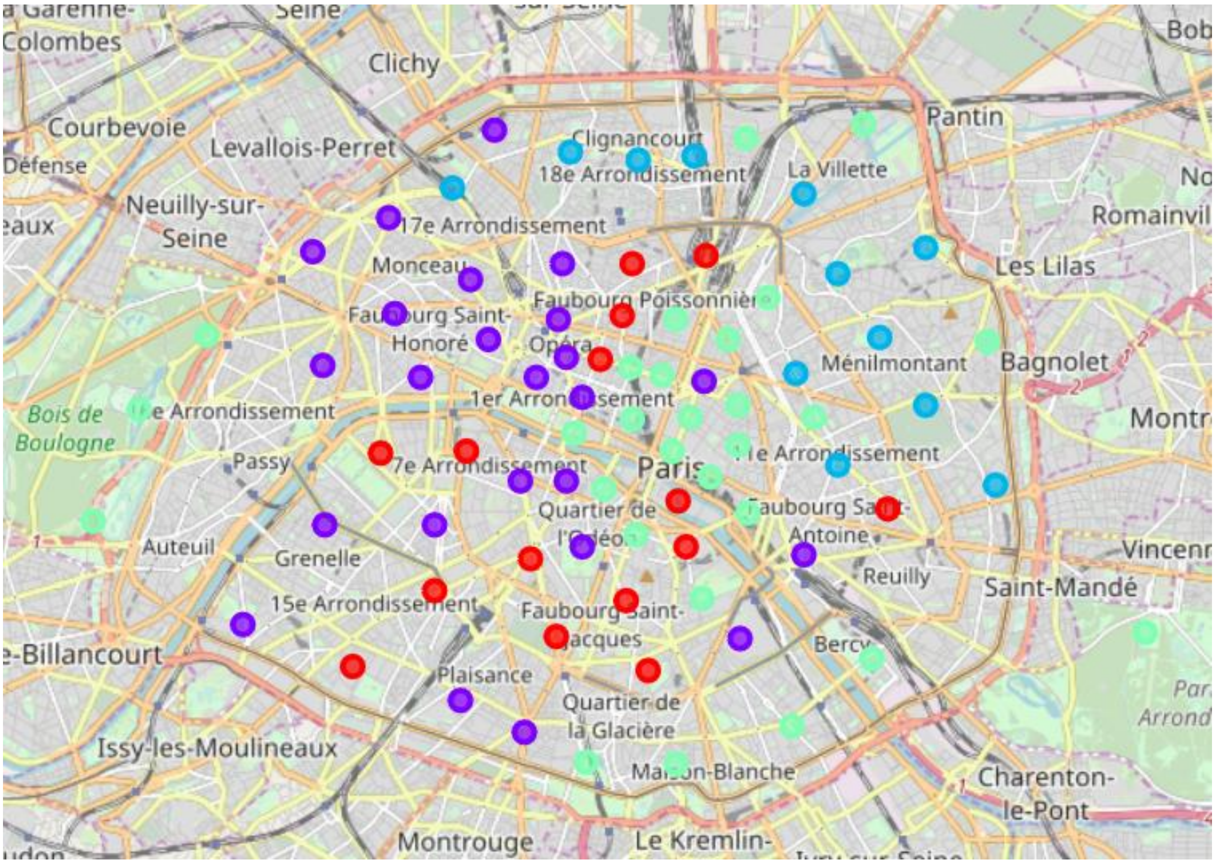
## 3.2 Collection of venues on Foursquare

A number of 7695 venues was obtained, among which 3483 were duplicates. After the data processing the final number of venues is 4212.

## 3.3 - Clustering results

**Cluster map**

The final number of clusters is 4. When k = 4, the neighborhood of "Bel-Air" constitutes a cluster to itself. The map of clusters is represented below:



Legend

Cluster 1

Cluster 2

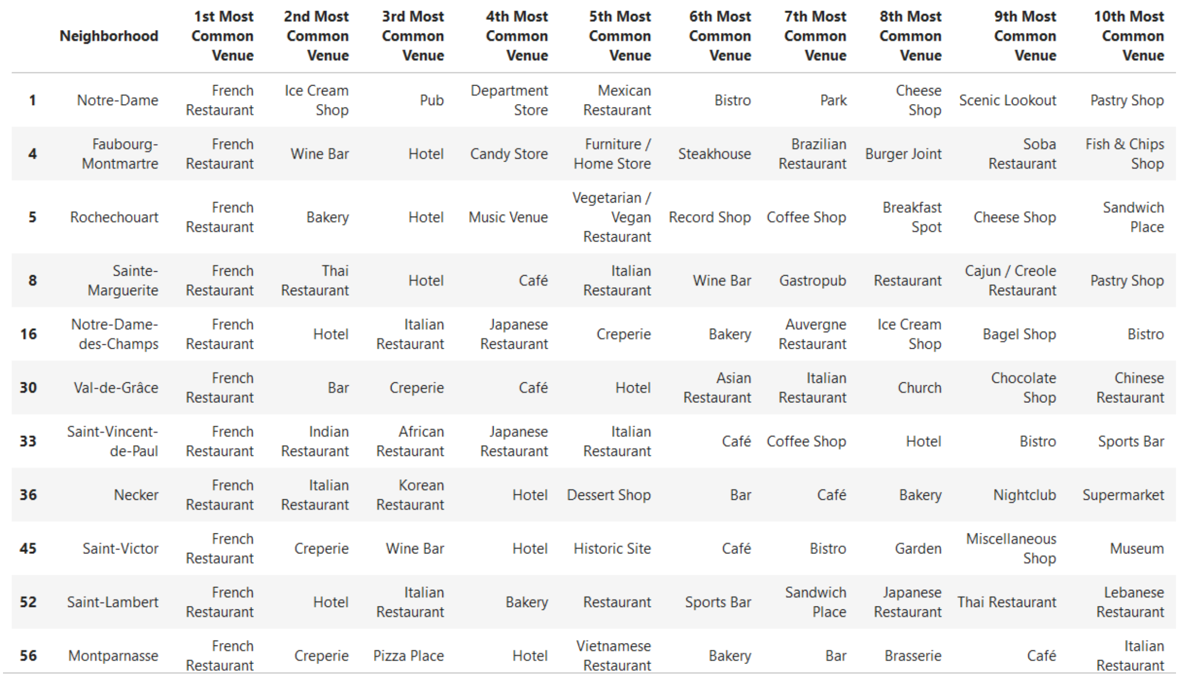
Cluster 3

Cluster 4

## Cluster analysis

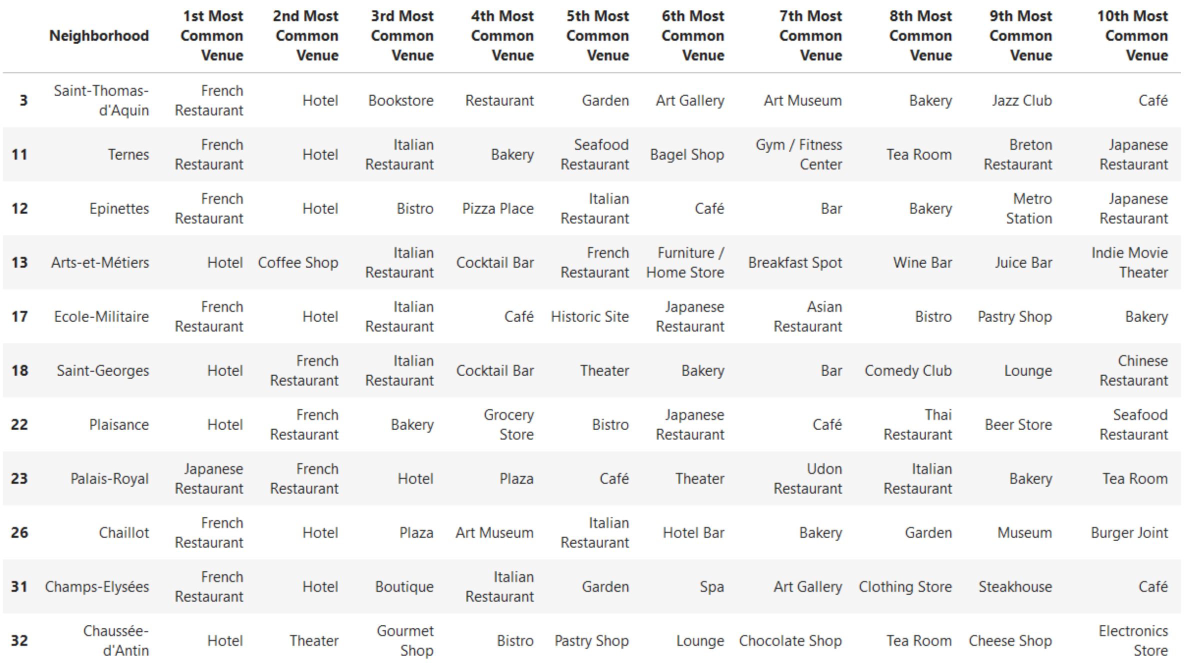
Let us represent the content of all clusters, the 10 most common venues per neighborhood are displayed below for each cluster:

### Cluster 1



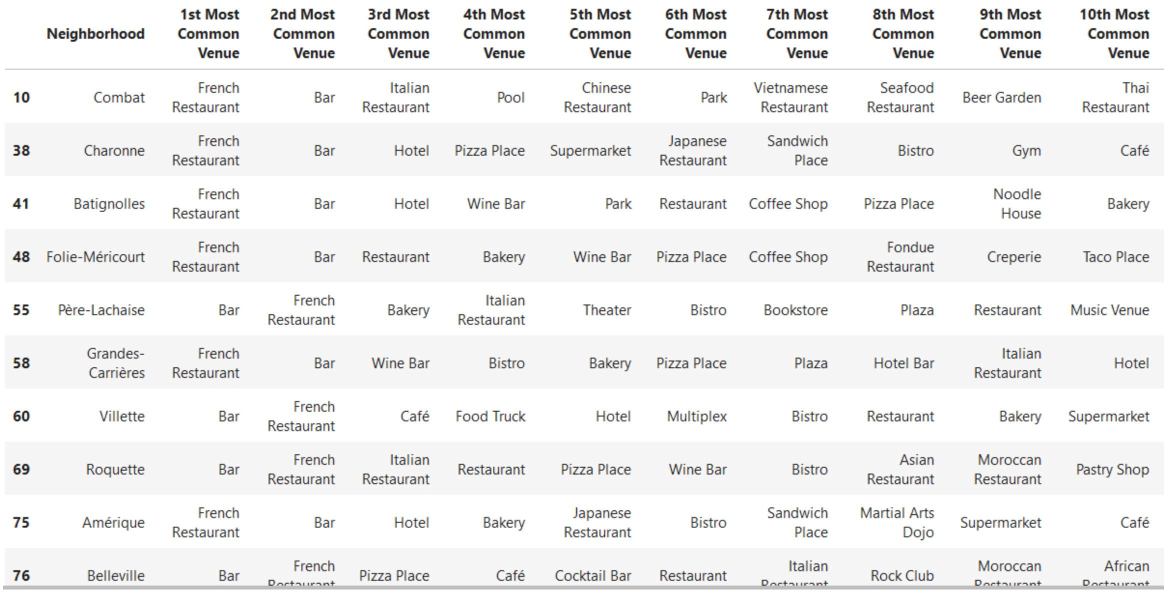
| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Notre-Dame | French Restaurant | Ice Cream Shop | Pub | Department Store | Mexican Restaurant | Bistro | Park | Cheese Shop | Scenic Lookout | Pastry Shop |
| 4 | Faubourg-Montmartre | French Restaurant | Wine Bar | Hotel | Candy Store | Furniture / Home Store | Steakhouse | Brazilian Restaurant | Burger Joint | Soba Restaurant | Fish & Chips Shop |
| 5 | Rochechouart | French Restaurant | Bakery | Hotel | Music Venue | Vegetarian / Vegan Restaurant | Record Shop | Coffee Shop | Breakfast Spot | Cheese Shop | Sandwich Place |
| 8 | Sainte-Marguerite | French Restaurant | Thai Restaurant | Hotel | Café | Italian Restaurant | Wine Bar | Gastropub | Restaurant | Cajun / Creole Restaurant | Pastry Shop |
| 16 | Notre-Dame-des-Champs | French Restaurant | Hotel | Italian Restaurant | Japanese Restaurant | Creperie | Bakery | Auvergne Restaurant | Ice Cream Shop | Bagel Shop | Bistro |
| 30 | Val-de-Grâce | French Restaurant | Bar | Creperie | Café | Hotel | Asian Restaurant | Italian Restaurant | Church | Chocolate Shop | Chinese Restaurant |
| 33 | Saint-Vincent-de-Paul | French Restaurant | Indian Restaurant | African Restaurant | Japanese Restaurant | Italian Restaurant | Café | Coffee Shop | Hotel | Bistro | Sports Bar |
| 36 | Necker | French Restaurant | Italian Restaurant | Korean Restaurant | Hotel | Dessert Shop | Bar | Café | Bakery | Nightclub | Supermarket |
| 45 | Saint-Victor | French Restaurant | Creperie | Wine Bar | Hotel | Historic Site | Café | Bistro | Garden | Miscellaneous Shop | Museum |
| 52 | Saint-Lambert | French Restaurant | Hotel | Italian Restaurant | Bakery | Restaurant | Sports Bar | Sandwich Place | Japanese Restaurant | Thai Restaurant | Lebanese Restaurant |
| 56 | Montparnasse | French Restaurant | Creperie | Pizza Place | Hotel | Vietnamese Restaurant | Bakery | Bar | Brasserie | Café | Italian Restaurant |

### Cluster 2



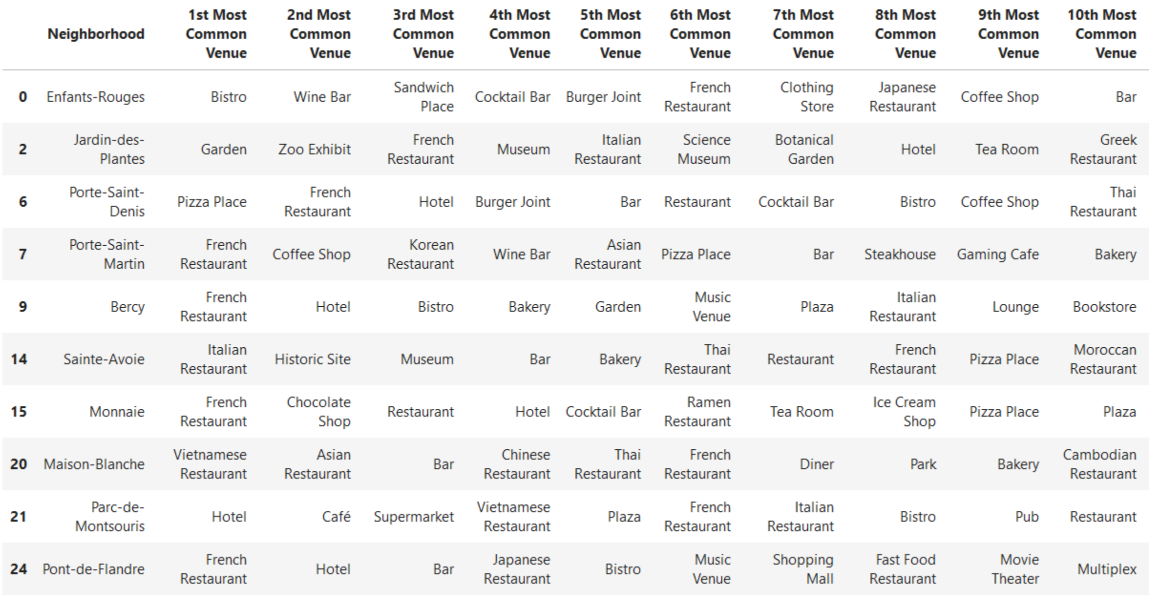
| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | Saint-Thomas-d'Aquin | French Restaurant | Hotel | Bookstore | Restaurant | Garden | Art Gallery | Art Museum | Bakery | Jazz Club | Café |
| 11 | Ternes | French Restaurant | Hotel | Italian Restaurant | Bakery | Seafood Restaurant | Bagel Shop | Gym / Fitness Center | Tea Room | Breton Restaurant | Japanese Restaurant |
| 12 | Epinettes | French Restaurant | Hotel | Bistro | Pizza Place | Italian Restaurant | Café | Bar | Bakery | Metro Station | Japanese Restaurant |
| 13 | Arts-et-Métiers | Hotel | Coffee Shop | Italian Restaurant | Cocktail Bar | French Restaurant | Furniture / Home Store | Breakfast Spot | Wine Bar | Juice Bar | Indie Movie Theater |
| 17 | Ecole-Militaire | French Restaurant | Hotel | Italian Restaurant | Café | Historic Site | Japanese Restaurant | Asian Restaurant | Bistro | Pastry Shop | Bakery |
| 18 | Saint-Georges | Hotel | French Restaurant | Italian Restaurant | Cocktail Bar | Theater | Bakery | Bar | Comedy Club | Lounge | Chinese Restaurant |
| 22 | Plaisance | Hotel | French Restaurant | Bakery | Grocery Store | Bistro | Japanese Restaurant | Café | Thai Restaurant | Beer Store | Seafood Restaurant |
| 23 | Palais-Royal | Japanese Restaurant | French Restaurant | Hotel | Plaza | Café | Theater | Udon Restaurant | Italian Restaurant | Bakery | Tea Room |
| 26 | Chaillot | French Restaurant | Hotel | Plaza | Art Museum | Italian Restaurant | Hotel Bar | Bakery | Garden | Museum | Burger Joint |
| 31 | Champs-Elysées | French Restaurant | Hotel | Boutique | Italian Restaurant | Garden | Spa | Art Gallery | Clothing Store | Steakhouse | Café |
| 32 | Chaussée-d'Antin | Hotel | Theater | Gourmet Shop | Bistro | Pastry Shop | Lounge | Chocolate Shop | Tea Room | Cheese Shop | Electronics Store |

## Cluster 3



| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | Combat | French Restaurant | Bar | Italian Restaurant | Pool | Chinese Restaurant | Park | Vietnamese Restaurant | Seafood Restaurant | Beer Garden | Thai Restaurant |
| 38 | Charonne | French Restaurant | Bar | Hotel | Pizza Place | Supermarket | Japanese Restaurant | Sandwich Place | Bistro | Gym | Café |
| 41 | Batignolles | French Restaurant | Bar | Hotel | Wine Bar | Park | Restaurant | Coffee Shop | Pizza Place | Noodle House | Bakery |
| 48 | Folie-Méricourt | French Restaurant | Bar | Restaurant | Bakery | Wine Bar | Pizza Place | Coffee Shop | Fondue Restaurant | Creperie | Taco Place |
| 55 | Père-Lachaise | Bar | French Restaurant | Bakery | Italian Restaurant | Theater | Bistro | Bookstore | Plaza | Restaurant | Music Venue |
| 58 | Grandes-Carrières | French Restaurant | Bar | Wine Bar | Bistro | Bakery | Pizza Place | Plaza | Hotel Bar | Italian Restaurant | Hotel |
| 60 | Villette | Bar | French Restaurant | Café | Food Truck | Hotel | Multiplex | Bistro | Restaurant | Bakery | Supermarket |
| 69 | Roquette | Bar | French Restaurant | Italian Restaurant | Restaurant | Pizza Place | Wine Bar | Bistro | Asian Restaurant | Moroccan Restaurant | Pastry Shop |
| 75 | Amérique | French Restaurant | Bar | Hotel | Bakery | Japanese Restaurant | Bistro | Sandwich Place | Martial Arts Dojo | Supermarket | Café |
| 76 | Belleville | Bar | French Restaurant | Pizza Place | Café | Cocktail Bar | Restaurant | Italian Restaurant | Rock Club | Moroccan Restaurant | African Restaurant |

## Cluster 4



| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Enfants-Rouges | Bistro | Wine Bar | Sandwich Place | Cocktail Bar | Burger Joint | French Restaurant | Clothing Store | Japanese Restaurant | Coffee Shop | Bar |
| 2 | Jardin-des-Plantes | Garden | Zoo Exhibit | French Restaurant | Museum | Italian Restaurant | Science Museum | Botanical Garden | Hotel | Tea Room | Greek Restaurant |
| 6 | Porte-Saint-Denis | Pizza Place | French Restaurant | Hotel | Burger Joint | Bar | Restaurant | Cocktail Bar | Bistro | Coffee Shop | Thai Restaurant |
| 7 | Porte-Saint-Martin | French Restaurant | Coffee Shop | Korean Restaurant | Wine Bar | Asian Restaurant | Pizza Place | Bar | Steakhouse | Gaming Cafe | Bakery |
| 9 | Bercy | French Restaurant | Hotel | Bistro | Bakery | Garden | Music Venue | Plaza | Italian Restaurant | Lounge | Bookstore |
| 14 | Sainte-Avoie | Italian Restaurant | Historic Site | Museum | Bar | Bakery | Thai Restaurant | Restaurant | French Restaurant | Pizza Place | Moroccan Restaurant |
| 15 | Monnaie | French Restaurant | Chocolate Shop | Restaurant | Hotel | Cocktail Bar | Ramen Restaurant | Tea Room | Ice Cream Shop | Pizza Place | Plaza |
| 20 | Maison-Blanche | Vietnamese Restaurant | Asian Restaurant | Bar | Chinese Restaurant | Thai Restaurant | French Restaurant | Diner | Park | Bakery | Cambodian Restaurant |
| 21 | Parc-de-Montsouris | Hotel | Café | Supermarket | Vietnamese Restaurant | Plaza | French Restaurant | Italian Restaurant | Bistro | Pub | Restaurant |
| 24 | Pont-de-Flandre | French Restaurant | Hotel | Bar | Japanese Restaurant | Bistro | Music Venue | Shopping Mall | Fast Food Restaurant | Movie Theater | Multiplex |

Cluster commenting

1. Touristic places : French restaurants and international restaurants
2. Cultural places & business : Museums, Historic sites, hotels …
3. Trendy places : Bars, international restaurants
4. Balanced venues : The interpretation if this last one is difficult because all kind of venues seem to be mixed

The third cluster seems interesting because it shows a concentration of evening and night life activities: bars, food Truck, international restaurants … and a good balance of other types of venues that tend to show that these are places where people live.

We choose the 3$^{rd}$ cluster for the rest of our analysis.

## 3.4 - Trending place analysis

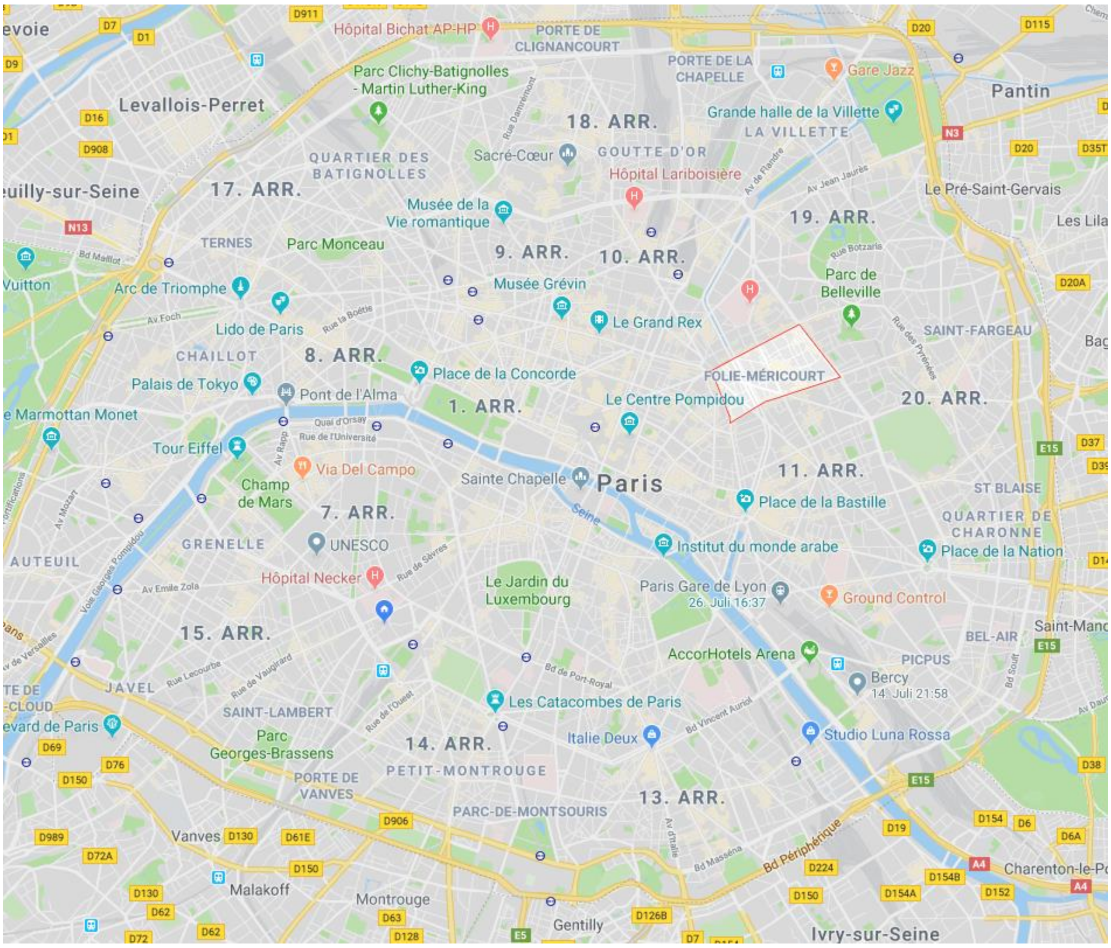
The request batch of trending places was done to foursquare at 9 pm. The hour of the request if important because foursquare returns the list of popular venues at the time of the request.

The median value of trending values for all neighborhoods is 0. It is a very surprising value which we will discuss in the next section.

Since the value is zero, we cannot select the neighborhoods which have zero or more trending venues because all of them would match … we then select the ones which have at least one trending venue.

## 3.5 - Selecting the best neighborhood

Among the last list, the chosen neighborhood is the one that has the least occurrences of venues labelled as "Italian Restaurant" and "Pizza Place". Our winner is "Folie-Méricourt" which is highlighted on the map below. Our business has the best chance to succeed if it is implanted in this area.



# 4 - Discussion

Some elements of methodology and hypothesis will be questioned in this section. It is important to look the results with some distance accordingly to these elements.

## 4.1 - Position of the neighborhoods

The position of the center of the neighborhoods is provided by an official source: the open data website of the city of Paris.

Our analysis has two main limits:

- All neighborhoods centers are not equally distanced. The superficies of some neighborhoods (especially in the periphery) is much larger than some others (i.e. in the hyper center of the city). By setting a radius value to 1km, it is very likely that the data will overlap in the center of the city (as noted in the results), while some information is missed for the largest neighborhoods. Two large parks in the periphery are included in the neighborhoods data. Some of the neighborhoods coordinates fall in these parks. In such a case, not a lot of venues can be found even within 1 km. It is a bias which has an impact on the cluster analysis: it stopped at 3 clusters because the fourth cluster is in the park. Maybe it would have been a better idea to drop these neighborhoods from the analysis to prevent these misleading effects.
- We do not consider the real geographical limits of each neighborhood. Instead of it, we considered that a venue belongs to a specific neighborhood if it is the closest with respect to the neighborhood center coordinated, which is totally different. It is highly probable that a lot if venues belong to another neighborhood that the one which was assigned. We should then not look at the real neighborhood's limitations for the implantation of our restaurant but generate a new map of neighborhoods accordingly to the delimitation which was indirectly done during the analysis. By the way this not a critical problem, it is not important to match the real neighborhoods of Paris (we could have used a made-up list), but we should keep in mind not to look at the real limits.

## 4.2 - Limitation of the numbers of trending venues

It was the noted in the results section that the global number of trending values was very low. Indeed, the median value is zero. How can that be explained?

- First reason, the radius value was kind of low (500 m) and leaves a lot of holes, we thus do not cover the total surface of Paris.
- Second reason and probably the most significant, Foursquare is not used a lot in France. Google Maps and Trip Advisor are much more used apps among the category. It is then probable that the live data is provided by tourists. This has nothing against tourist, but they do not really fall in the category of people which could be interested by our Italian restaurant as most of them are looking for a French experience. It can be then partly deceptive to base our neighborhood selection on trending venues data which to not completely match our targeted audience.

# Conclusion

In conclusion we have found a place which seems to be a very good candidate for the opening of our Italian restaurant: the neighborhood of "Folie-Mericourt".

After gathering the data of more than 4000 venues in Paris, we performed a k-means cluster analysis to identify groups of neighborhood which share common characteristics. The analysis yielded four clusters, whose respective content was analyzed and interpreted. One cluster has withdrawn our attention, as the most popular venues of the neighborhood members mainly consisted of trendy type of places: bars, international food, street food … It is the kind of area which is interesting for our application.

Among this cluster, we have found out which neighborhoods are the most frequented in the evening. From this last list, we kept the one neighborhood which has the least Italian restaurants.

Despite some methodology limitations, such as the non-exact mapping of neighborhoods and the limited number of trending venues, we can have a decent solution to our initial problem.

Even if some instability was noted about the results yielded by the analysis (Foursquare does not give exactly the same venues list all the time), the categorization of the clusters does not change radically, which proves a certain robustness of the results.

Here are a few ideas to reach an upper degree of precision in the results:

- Mapping in a more precise way the real geographical limits of each neighborhood
- Getting an exhaustive list of all venues in Paris
- Getting a richer mapping of trending venues from Foursquare