WHAT KIND OF VENUE TO OPEN IN PARIS AND WHERE TO OPEN IT ?

Introduction

General Overview

Paris is the city of lights and love, and attracts more than 50 millions tourists a year.
With its museums, monuments, rsetaurants and shops, Paris is the most visited capital in the world. For a weekend or for a week, for lovers or for families and friends visiting, there are a lot of venues that can be of interest for tourists.

Business Problem

With a city with that much tourists, there sure are a lot of business that can be opened to be useful to tourists. This report is to be used by business person looking to open a business in Paris, to determine what kind of business to open, and where to open it to attract tourists, with few competition.

We will use data to first identify the most popular business to open, then we'll focus on the neighborhoods where it would be suitable to open the business, in lights of the two constraints above : attracting tourists, and few competition.

Data to be used

To handle our business problem, we used the following data :

- Neighborhood data for Paris
- Coordinates of the neighborhoods of Paris
- The data relative to the venues in Paris

Neighborhood Data

To retrieve the neighborhood data for Paris, we used the wikipedia page listing the neighborhood in Paris, and scraped its content using the pandas method pd.read_html, that can be easily used since Wikipedia use html tables. The url is the following : https://en.wikipedia.org/wiki/Quarters_of_Paris
The extracted data resulted in a dataframe listing for each District ('Arrondissement') in Paris the neighborhood in said districts.
There are 20 districts in Paris, and 4 neighborhoods per district, which gives a total of 80 neighborhoods in Paris.

Coordinates of neighborhoods

To retrieve the coordinates for each neighborhood, which list we previously scraped, we used the geocoder library.
The coordinates are useful for later visualization on a map, using the Folium library.

Venues data
The venues data are extracted using Foursquare API, to get the venues for each neighborhood, in a radius of 500m around the center of the neighborhood.
For each venue, we have :

- The name of the venue
- The coordinates of the venue
- The venue category

The venue category data will be particularly useful to determine which kind of venue are the most frequent.

Methodology

After scrapping the data on the wikipedia page, we re-arranged the dataframe for further use.

	District	Quarter n°	Quarter	Population in 1999	Area in hectare
0	1st arrondissement(Called "du Louvre")	1st	Saint-Germain-l'Auxerrois	1672	86.9
1	1st arrondissement(Called "du Louvre")	2nd	Les Halles	8984	41.2
2	1st arrondissement(Called "du Louvre")	3rd	Palais-Royal	3195	27.4
3	1st arrondissement(Called "du Louvre")	4th	Place-Vendôme	3044	26.9
4	2nd arrondissement(Called "de la Bourse")	5th	Gaillon	1345	18.8

Figure 1 Data extracted from wikipedia on a dataframe format

The columns that were not necessary were dropped, and we rearranged the name of the columns.

Data description

We used the describe method to explore the neighborhood of Paris.

	District	Quarter n°	Quarter	Population in 1999	Area in hectare
count	80	80	80	80.000000	80.000000
unique	20	80	80	NaN	NaN
top	7th arrondissement(Called "du Palais-Bourbon")	49th	Bel-Air	NaN	NaN
freq	4	1	1	NaN	NaN
mean	NaN	NaN	NaN	26573.137500	121.398750
std	NaN	NaN	NaN	20235.804847	132.880318
min	NaN	NaN	NaN	1345.000000	18.800000
25%	NaN	NaN	NaN	9413.750000	49.750000
50%	NaN	NaN	NaN	21418.500000	92.850000
75%	NaN	NaN	NaN	39025.250000	144.800000
max	NaN	NaN	NaN	82032.000000	1126.000000

Figure 2 Describe method applied to the neighborhood dataframe

From this, coupled with the valuue_counts() method, we managed to determine that there were neighborhood per district, so those neighborhood are evenly distributed. The next step was to visualize those neighborhood on a map.

Data visualization

Using the geocoder method and the dataframe containing the name of the neighborhoods, we retrieved the coordinates of the neighborhood, for further visualization on a map, using Folium.

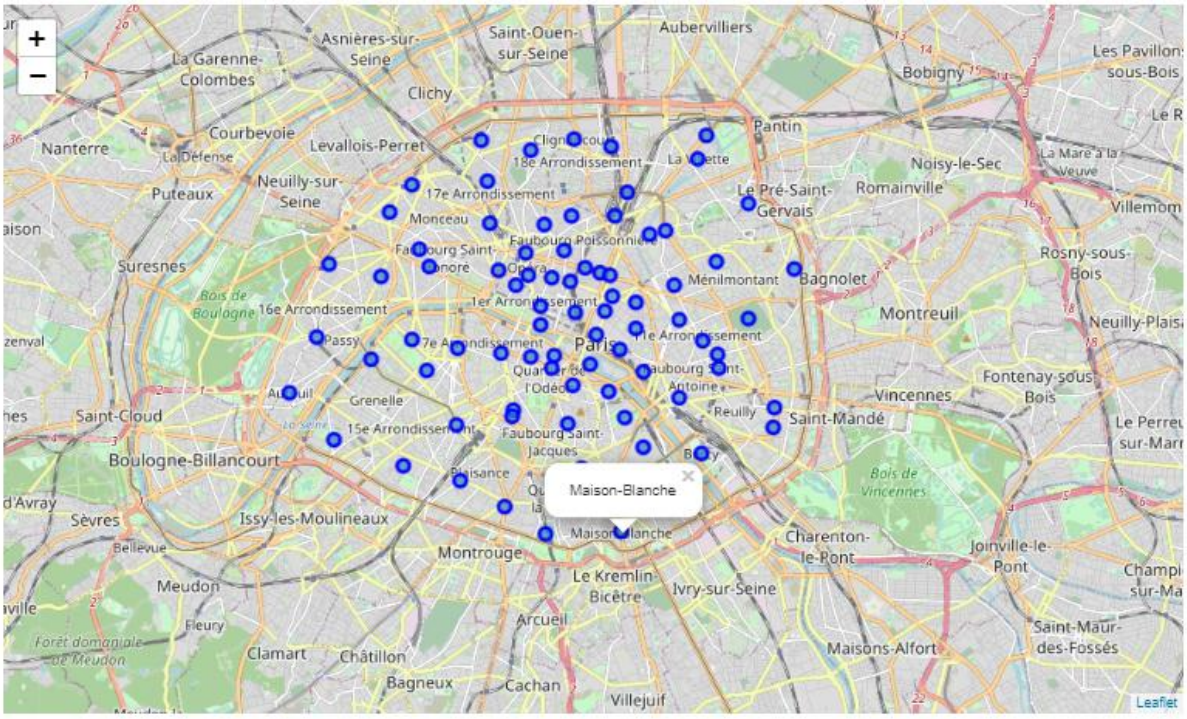

Figure 3 The 80 neighborhoods of Paris

The neighborhood of Paris are pretty spread out, even if we see that they are more concentrated towards the city center.

Venue exploration

Using the Foursquare API, we extracted all the venues in a radius of 500m around the center of each neighborhood, to list all the venues of the neighorhood.

500m is a good enough radius to include all the venues of the neighborhood, without overlapping on the neighboring neighborhoods, since they are close to each other, particularly in the center of Paris.

When we describe the type of venue, using the describe method on the "Venue Category" feature, we can sense already that restaurant might be a top choice for tourist, since they are the most frequent venue.

The feature "Venue Category" is the one that seems more important for our analysis.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
count	2322	2322.000000	2322.000000	2322	2322.000000	2322.000000	2322
unique	80	NaN	NaN	1865	NaN	NaN	231
top	Chaillot	NaN	NaN	Carrefour City	NaN	NaN	French Restaurant
freq	30	NaN	NaN	16	NaN	NaN	321
mean	NaN	48.861076	2.342284	NaN	48.861062	2.342385	NaN
std	NaN	0.017439	0.030848	NaN	0.017182	0.030654	NaN
min	NaN	48.821991	2.266738	NaN	48.818503	2.260499	NaN
25%	NaN	48.849123	2.324562	NaN	48.849750	2.324766	NaN
50%	NaN	48.861052	2.345004	NaN	48.861520	2.344926	NaN
75%	NaN	48.871727	2.362585	NaN	48.872487	2.362263	NaN
max	NaN	48.894728	2.406736	NaN	48.897619	2.408993	NaN

Figure 4 Description of the Venues in Paris

With that new information, we focused more on the restaurant venues, filtering only the restaurant venues. From this, we see that out of 2322 venues in Paris, 855 of them are restaurants.

This information encouraged us to focus more on restaurants, to confirm that in each neighborhood, restaurants are the most frequent venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
16	Saint-Germain-l'Auxerrois	48.860211	2.336299	LouLou	48.862804	2.333500	Italian Restaurant
23	Saint-Germain-l'Auxerrois	48.860211	2.336299	La Régalade Saint-Honoré	48.861620	2.341749	French Restaurant
28	Saint-Germain-l'Auxerrois	48.860211	2.336299	Boutique yam'Tcha	48.861710	2.342380	Chinese Restaurant
33	Les Halles	48.862466	2.346009	Pirouette	48.862876	2.348135	French Restaurant
36	Les Halles	48.862466	2.346009	Enza & Famiglia	48.861191	2.343449	Italian Restaurant

```
paris_restaurant.shape
```

```
(855, 7)
```

Figure 5 The venues showing only restaurants

We aggregated the data, by replacing the cuisine type (eg "Italian restaurant") by "restaurant" only, to compare to other venues.

We used One Hot Encoding to compare more easily, for each neighborhood, the most frequent venues.

	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	Amérique	Convenience Store	Pool	Café	Bistro	Supermarket	Restaurant	Bed & Breakfast	Brewery	Basketball Court	Theater
1	Archives	Restaurant	Art Gallery	Art Museum	Sandwich Place	Creperie	Clothing Store	Farmers Market	Men's Store	Cosmetics Shop	Cheese Shop
2	Arsenal	Restaurant	Park	Plaza	Cocktail Bar	Pedestrian Plaza	Museum	Boat or Ferry	Pub	Bakery	Beer Store
3	Arts-et-Métiers	Restaurant	Hotel	Coffee Shop	Church	Bakery	Noodle House	Juice Bar	Beer Bar	Cocktail Bar	Bubble Tea Shop
4	Auteuil	Restaurant	Supermarket	Pizza Place	Bistro	Bike Rental / Bike Share	Gym / Fitness Center	Burger Joint	Art Museum	Market	Park
5	Batignolles	Restaurant	Wine Bar	Coffee Shop	Pizza Place	Gym / Fitness Center	BBQ Joint	Bar	Pastry Shop	Steakhouse	Noodle House

Figure 6 The most frequent venues for each neighborhood

Looking at the head of the dataframe, we can see that restaurants seems to be the most common venue in most neighborhood. Let's confirm with the describe method.

	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
count	80	80	80	80	80	80	80	80	80	80	80
unique	80	5	24	31	45	44	48	50	52	53	55
top	Enfants-Rouges	Restaurant	Hotel	Hotel	Cocktail Bar	Bistro	Pizza Place	Café	Bistro	Bakery	Park
freq	1	74	26	8	9	7	7	9	7	4	4

Figure 7 Describe method applied to the df giving the most frequent venues

This shows us that in 74 neighborhoods out of 80, restaurants are the most frequent venues.

It encourages us to advise to open a restaurant.

On the following part of this study, we'll focus only on restaurants venues, using the dataframe paris_restaurant that contains only restaurant venues.

Using the count_values(), we can confirm that French cuisine is by far the most popular

```
paris_restaurant.groupby('Venue Category').count().sort_values(by='Venue', ascending=False)['Neighborhood']
```

```
Venue Category
French Restaurant                321
Italian Restaurant                99
```

Figure 8 There is no doubt about French Cuisine

K-means clustering

We can now cluster the neighborhoods, working only with restaurant venues.

First we determined the optimal K for clustering using the Silouhette method and Silouhette Score.

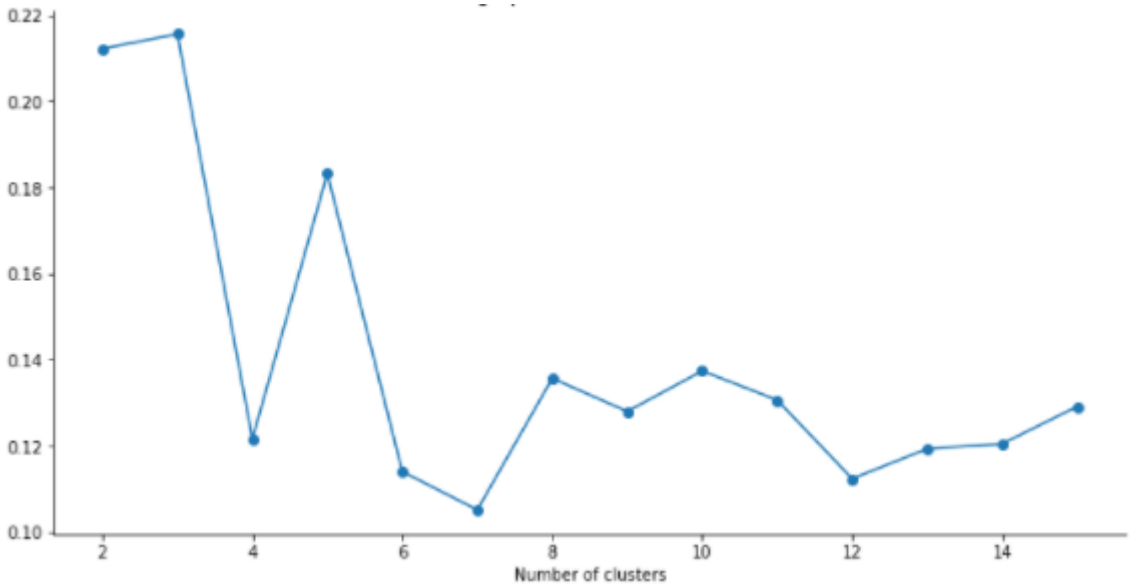

Figure 9 Silouhette Score showing optimal K

The optimal K is 3.

Clustering is performed using k = 3.

Results

The clustering model then clusters the neighborhoods and provides a label for each neighborhood which is representative of the cluster it belongs to.

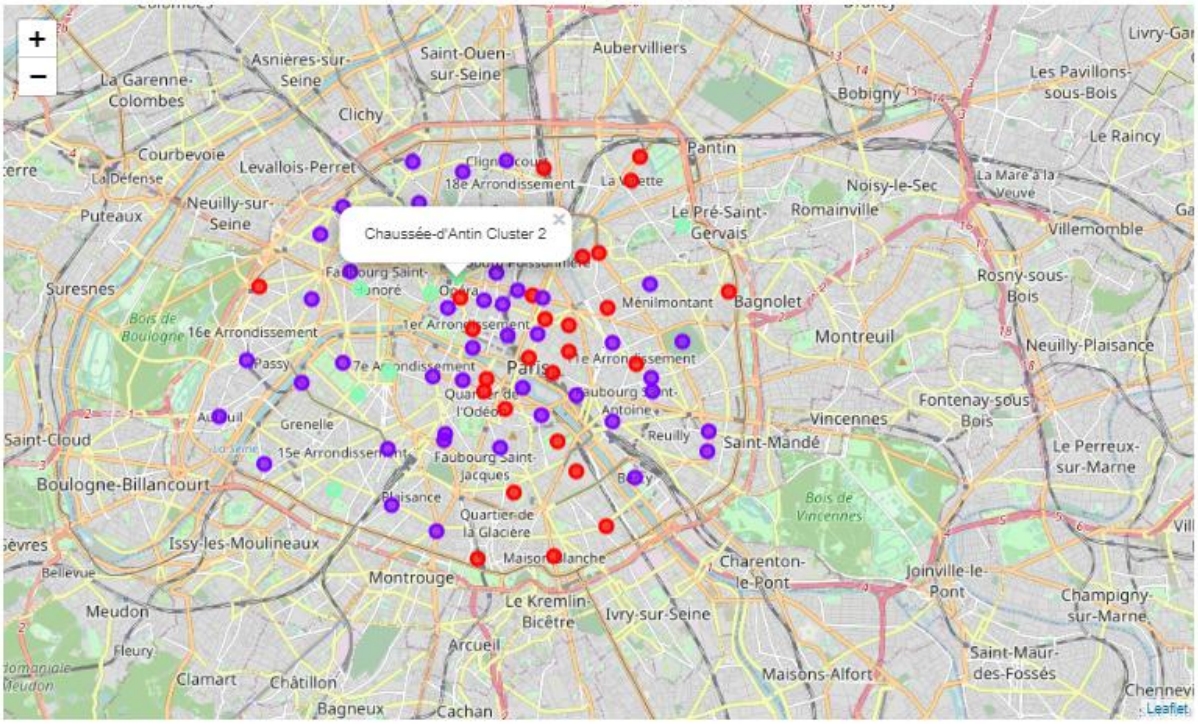

Figure 10 The 3 kind of clusters for restaurants

We see that Cluster n°0 has most italian cuisine restaurants, with french cuisine being a close 2nd.

We see that Cluster n°1 has most french cuisine restaurants, way more than the 2 other clusters, and italian cuisine being second

Lastly, in Cluster n°2, french cuisine is still the top cuisine, but with more 'around the world' cuisine restaurants.

So we'll call Cluster n°0 the "French Italian Cluster", Cluster n°1 the "French Cluster" and Cluster n°2 the "Around the world" cluster.

	Quarter n°	Latitude	Longitude	Cluster Labels	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	10 C
count	7	7.000000	7.000000	7.0	7	7	7	7	7	7	7	7	7	
unique	7	NaN	NaN	NaN	1	3	4	5	4	4	4	4	4	
top	57th	NaN	NaN	NaN	French Restaurant	Vietnamese Restaurant	Jiangxi Restaurant	Italian Restaurant	Israeli Restaurant	Indonesian Restaurant	Indian Restaurant	Greek Restaurant	Gluten-free Restaurant	Fa Re
freq	1	NaN	NaN	NaN	7	3	3	3	3	3	3	3	3	
mean	NaN	48.862680	2.324642	2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
std	NaN	0.016233	0.032829	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
min	NaN	48.834444	2.298216	2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
25%	NaN	48.853901	2.305044	2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
50%	NaN	48.870137	2.313343	2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
75%	NaN	48.871978	2.328411	2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
max	NaN	48.882424	2.394025	2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

Figure 11 Example of the 3rd cluster

Discussion

When we analyse the cluster, we see that some cluster are more suited that other for opening a restaurant. For instance, to be with less fierce competition, cluster n°2 (the 3rd one) is most suited because we see that there are less restaurants in this cluster. And most of the neighborhood of this cluster are located in the center of Paris, which is more likely to attract tourists.

Further more, we saw previously that the French cuisine is the most popular cuisine for restaurants. All points towards opening a French restaurant. To avoid competition, cluster n°2 also offers a good compromise since it is the cluster with less French cuisine restaurants.

Conclusion

So to address our business problem, if one is looking to open a business in Paris, this study would recommand to open a **restaurant** where it would be served **French Cuisine** that is the most in demand, and this business should be opened in neighborhoods pertaining to **cluster n°2**, where there are the less French restaurants, and really few restaurant at all. For instance the neighborhood of **Chaussée d'Antin** would be fine, since it attracts a lot of tourists, thanks to the nearby Galeries Lafayette, and is located in the center of Paris. The neighborhood of **La Madeleine** would be a fine choice too.