A new Jazz Club in Paris – Capstone Project

by François MORENNE - 27/07/2020

Table of contents

I. Introduction

I.a) Background

Jazz is a music genre that originated in the African-American communities of New Orleans, United States, in the late 19th and early 20th centuries, with its roots in blues and ragtime. Since the 1920s Jazz Age, it has been recognized as a major form of musical expression in traditional and popular music, linked by the common bonds of African-American and European-American musical parentage.

Jazz is characterized by swing and blue notes, call and response vocals, polyrhythms and improvisation. Jazz has roots in West African cultural and musical expression, and in African-American music traditions.

The genre landed in Paris with the sulfurous American Josephine Baker, before invading the cabarets of Pigalle with her swing during the Twenties.

Nowadays we can find great Jazz Clubs in Paris:

- the Sunset/Sunside
- the Baiser Salé
- the Duc des Lombards
- the Caveau de la Huchette
- ...

As a huge Jazz fan, I would love to open a club in Paris. The purpose of this work is to study Paris neighborhoods and determine which area would be best suited for our new club.

I.b) Problem

Data that might contribute to determining the best place for our Jazz Club include restaurants, bars/ nightlife spots, Cinema, public transport (parking, bus stop, metro station, etc.). Meanwhile, we have to make sure the other jazz clubs are far away from ours, to avoid competition.

Thanks to this analysis, we are hoping to identify the neighborhoods best suited for our Frenchy Jazz club.

I.c) Interest

As a future investment I am obviously very interested to know what is the best place for my Jazz Club.

I need to find the right neighborhoods where people are most likely to go if they want to hang out and listen to some good jazz music. Also, I want to make sure that I won't have to compete with other Jazz places.

II. Data acquisition and Cleaning

II.a) Data Sources

For this study, we will use the following datasets:

- **Foursquare API** : to retrieve location data, help us identify music venues, as well as grouping each neighbourhood into clusters.
- **Wikipedia article about Paris neighborhoods** : https://en.wikipedia.org/wiki/Quarters_of_Paris
- **GeoJSON file containing the shape of the neighborhoods in Paris**: https://france-geojson.gregoiredavid.fr/GeoJSON

II.b) Data Cleaning

I first had to scrape Wikipedia page in order to retrieve useful info such as "Arrondissement", "Quartiers", and "Name". To do so, I used BeautifulSoup.

Then in order to generate a map, I had to be able to make use of GeoJSON file. To do so, I needed to retrieve the postal code of each arrondissement. I created a new dataframe and integrated the postal code consisting of : "75"+"1xx" (xx: identification number of the arrondissement). To print the name of the neighborhoods on the map, I had to retrieve their latitude and longitude information. I made use of geopy library.

During my experimentations, I realized that cutting Paris into neighborhoods was not enough. I needed more granularity. I so decided to use another geoJSON file to split into "quartiers". The difficulty was that quartier name was not always exactly the same depending on the dataset (missing accents, having pronouns 'le', 'les' or not...). That's why I decided to rely on an identification number instead. I realized the geoJSON file contained a parameter called "c_quinsee", representing INSEE identification of each quartier. I so decided to add to my dataframe a new column with this INSEE identification, consisting of :

'75101' + 'xx' (xx: ID of the quartier)

I iterated over the whole data frame and added this information manually.

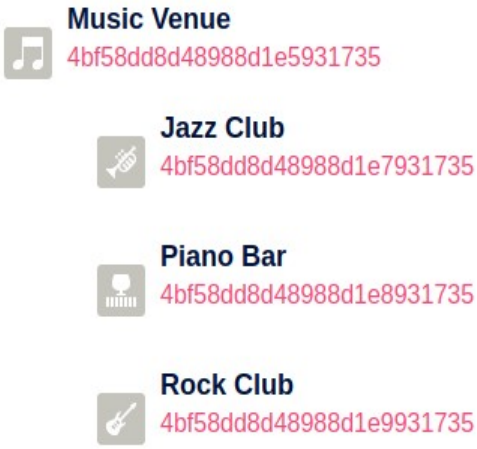
At this point, I was able to generate a choropleth map with a view by qartier.

III. Exploratory Data Analysis & Methodology

III.a) Exploring Music Venues in « Latin Neighborhood »

I started my analysis by reporting all music venues in the « Latin Neighborhood » in Paris.

To do so, I created a foursquare search query with a specific ID = **4bf58dd8d48988d1e5931735** allowing to retrieve all the music venues, such as Jazz clubs, Piano bars, Rock clubs, and so on.



This query returns a raw JSON. From this JSON, I extracted venue categories, names, addresses, latitude, longitude and then put all the collected information into a Dataframe :

	name	lat	lng	formattedAddress	categories
0	Le Piano Vache	48.847267	2.347707	[8 rue Laplace, 75005 Paris, France]	Pub
1	Le Petit Journal Saint Michel	48.846467	2.340402	[71 boulevard Saint-Michel, 75005 Paris, France]	Jazz Club
2	Le Caveau des Oubliettes	48.852007	2.346736	[52 rue Galande, 75005 Paris, France]	Jazz Club
3	Aux Trois Mailletz	48.852177	2.346609	[56 rue Galande, 75005 Paris, France]	Piano Bar
4	Caveau de la Huchette	48.852785	2.346305	[5 rue de la Huchette, 75005 Paris, France]	Jazz Club
5	Le Petit Jornal	48.845281	2.340555	[Boulevard Saint Michel, Paris, France]	Jazz Club
6	Boub's	48.850926	2.339206	[France]	Jazz Club
7	Piano Bar	48.852967	2.345922	[14 rue de la Huchette, 75005 Paris, France]	Piano Bar
8	studio purple	48.847867	2.351259	[France]	Music Venue

III.b) Exploring Music Venues in Paris

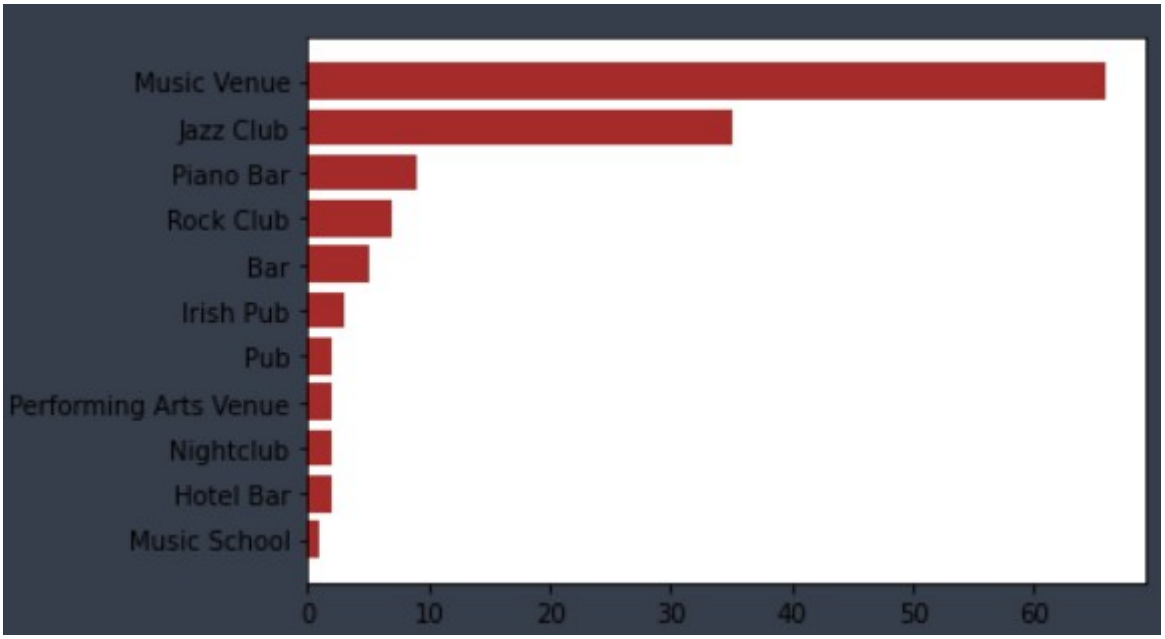
I then extended this analysis to the whole Paris and tried to extract interesting information :

III.b.i. Analysis of Top 10 Music Quartier

I first identified the Top 10 quartier for music by counting the number of music venues by quartier :

	Music Venue Count	INSEE_ID	Quartier Name_y
0	17.0	7510204	Bonne-Nouvelle
1	15.0	7510602	Odéon
2	15.0	7511101	Folie-Méricourt
3	14.0	7510102	Les Halles
4	14.0	7510601	Monnaie
5	13.0	7510401	Saint-Merri
6	12.0	7510304	Sainte-Avoye
7	12.0	7510904	Rochechouart
8	11.0	7510203	Mail
9	11.0	7511002	Porte-Saint-Denis

I then studied what was the most representative music venue in these Top 10 quartier :



The difficulty in our dataset is that « music venue » is a mix of different music categories. It’s therefore quite hard to extract meaningful information from it.

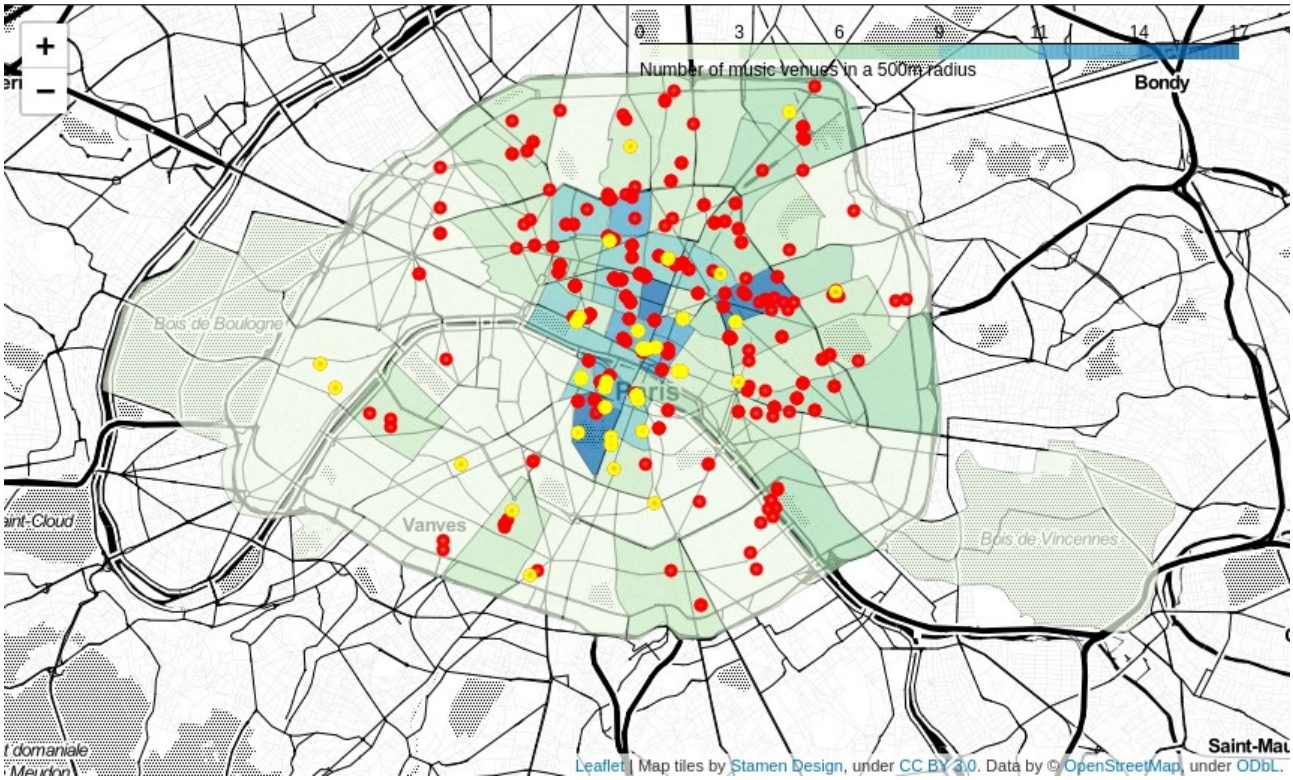
Nevertheless, we can see that Jazz Clubs and Piano Bars are very well represented in these Top 10 Music quartier.

III.b.ii. Choropleth map of Music Venues

To get a better idea of the « active » quartier in terms of music, I then created a choropleth map to analyze the number of music venues in a 500m radius by quartier.

I also added red points, representing those music venues.

I then distinguished between music venues (in red) and Jazz clubs (in yellow) :

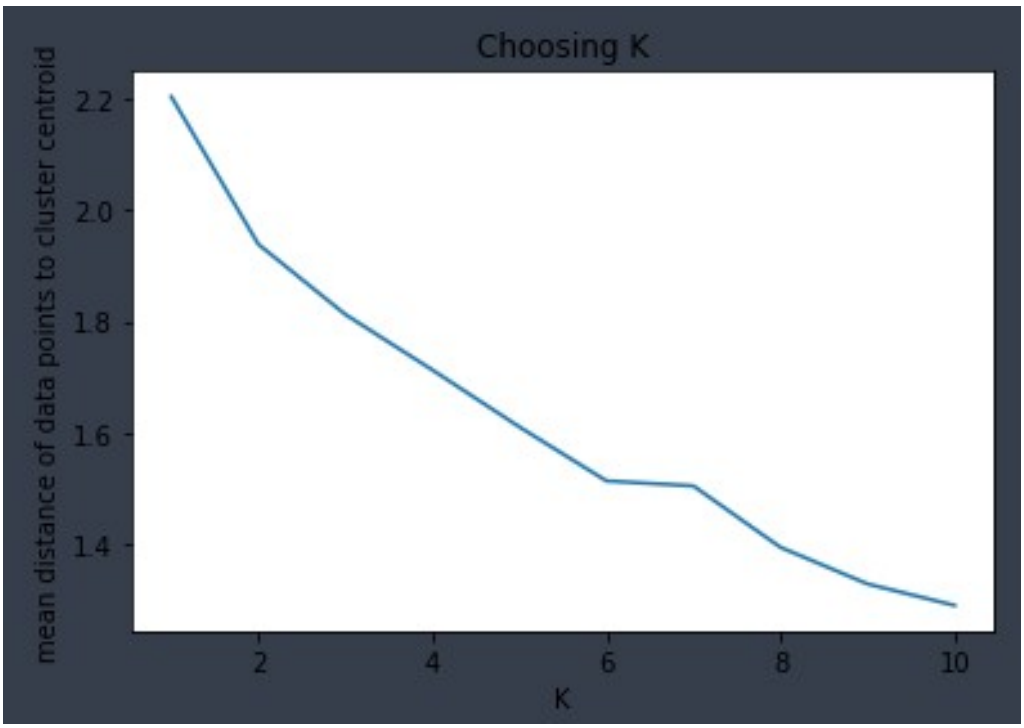


III.b.iii. Clustering the quartier

I used a machine learning alogorithm in order to cluster the different quartiers, to then identify where to open my new Jazz Club.

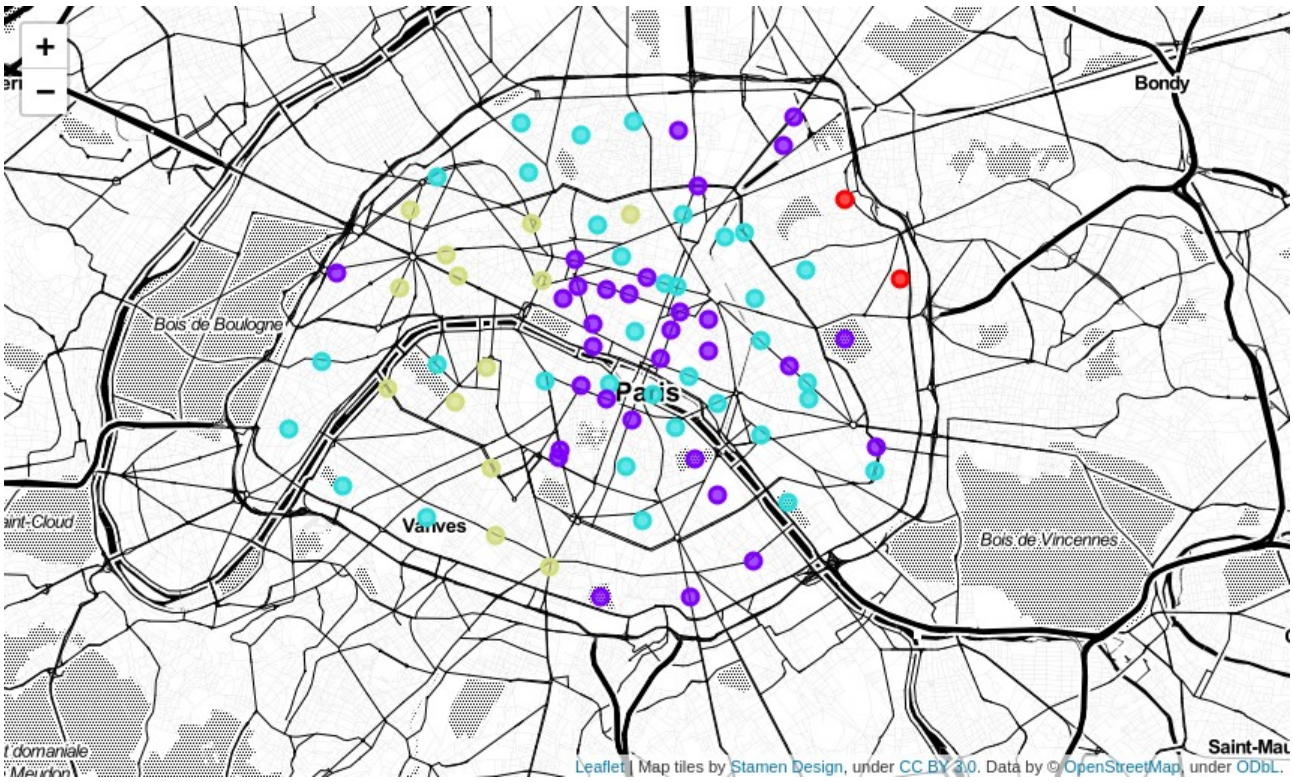
To do so, I used K-means algorithm. The first step was to choose K value, ie the number of clusters. To do so, I tried to apply the « elbow method ».

I started by running the algorithm for different K values ranging from 1 to 10 and reported the mean distance of data points to cluster centroids. The figure below shows the results obtained :



As we can see, the « elbow » is quite hard to identify… I decided to carry out some empirical studies and then realized that K=4 seemed to be a good tradeoff.

The figure below shows the different clusters on the map obtained for K=4 :



I repoted the top 10 venues from these 4 clusters :

	Cluster0 top10	Cluster1 top10	Cluster2 top10	Cluster3 top10
0	Pizza Place	French Restaurant	French Restaurant	Hotel
1	English Restaurant	Hotel	Hotel	French Restaurant
2	Plaza	Café	Italian Restaurant	Bakery
3	Supermarket	Bakery	Bakery	Italian Restaurant
4	Café	Japanese Restaurant	Bar	Japanese Restaurant
5	French Restaurant	Bistro	Bistro	Bistro
6	Bistro	Italian Restaurant	Japanese Restaurant	Café
7	Empanada Restaurant	Bar	Coffee Shop	Restaurant
8	Tram Station	Coffee Shop	Pizza Place	Coffee Shop
9	Fast Food Restaurant	Cocktail Bar	Burger Joint	Sandwich Place
10	Theater	Plaza	Restaurant	Cosmetics Shop
11	Pool	Restaurant	Wine Bar	Plaza
12	Bakery	Wine Bar	Café	Brasserie
13	Bookstore	Sandwich Place	Cocktail Bar	Gourmet Shop
14	Diner	Chinese Restaurant	Thai Restaurant	Bar
15	Ethiopian Restaurant	Theater	Plaza	Pizza Place
16	Arts & Entertainment	Pastry Shop	Bookstore	Korean Restaurant
17	Metro Station	Bookstore	Chinese Restaurant	Cocktail Bar
18	Bed & Breakfast	Supermarket	Supermarket	Chinese Restaurant
19	Exhibit	Burger Joint	Park	Clothing Store

From the table, we can see some interesting venues, possibly indicating high potential for a Jazz place, such as Bars, Cocktail Bars, and Wine Bars.

From this table, it seems likely that :

- ***Cluster 0*** *doesn't represent a good potential for a Jazz Club.*

  We have some restaurants for sure, but these type of quartier seems to be more centered on exhibits, theaters, and Arts & Entertainment.

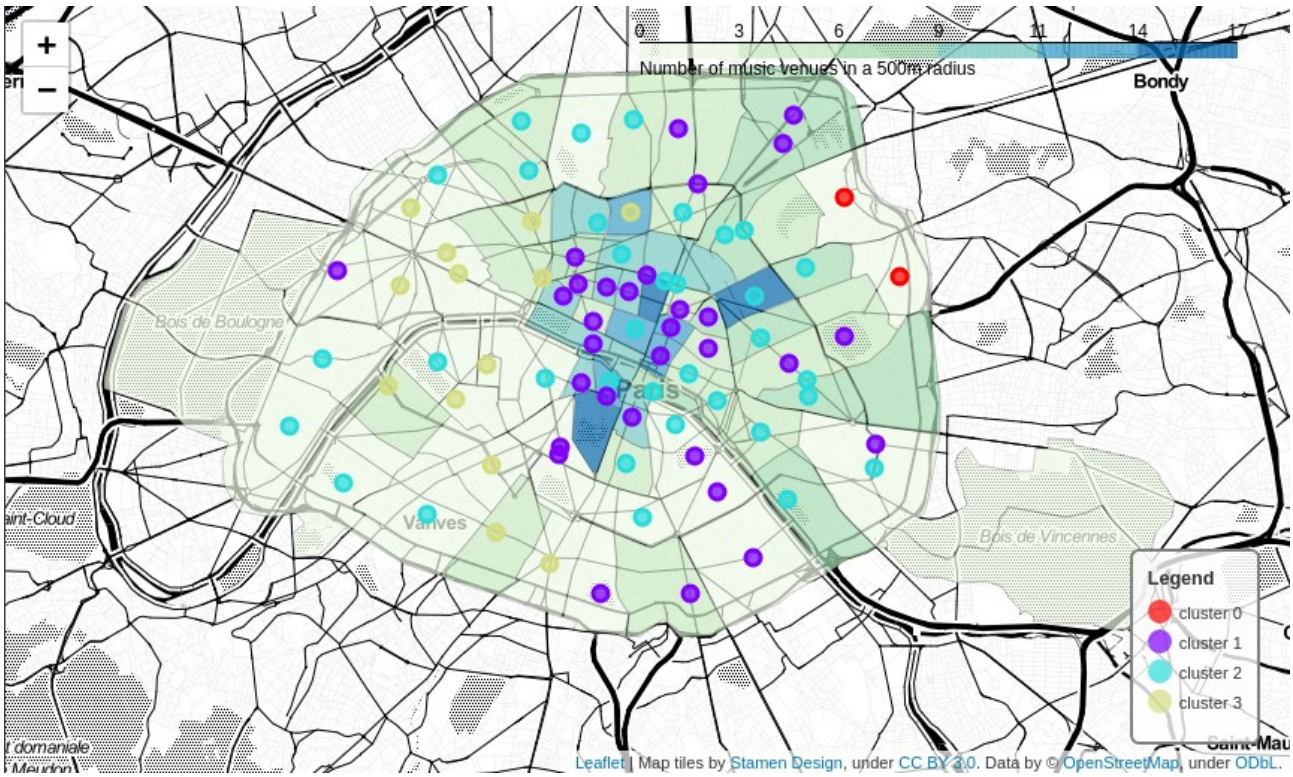
- ***Clusters 1 & 2*** *are likely to present a good opportunity.*

  There are plenty of restaurants, as well as different types of bar places (cocktail, wine...), which implies an active night life.

- ***Cluster 3*** *presents a moderate opportunity.*

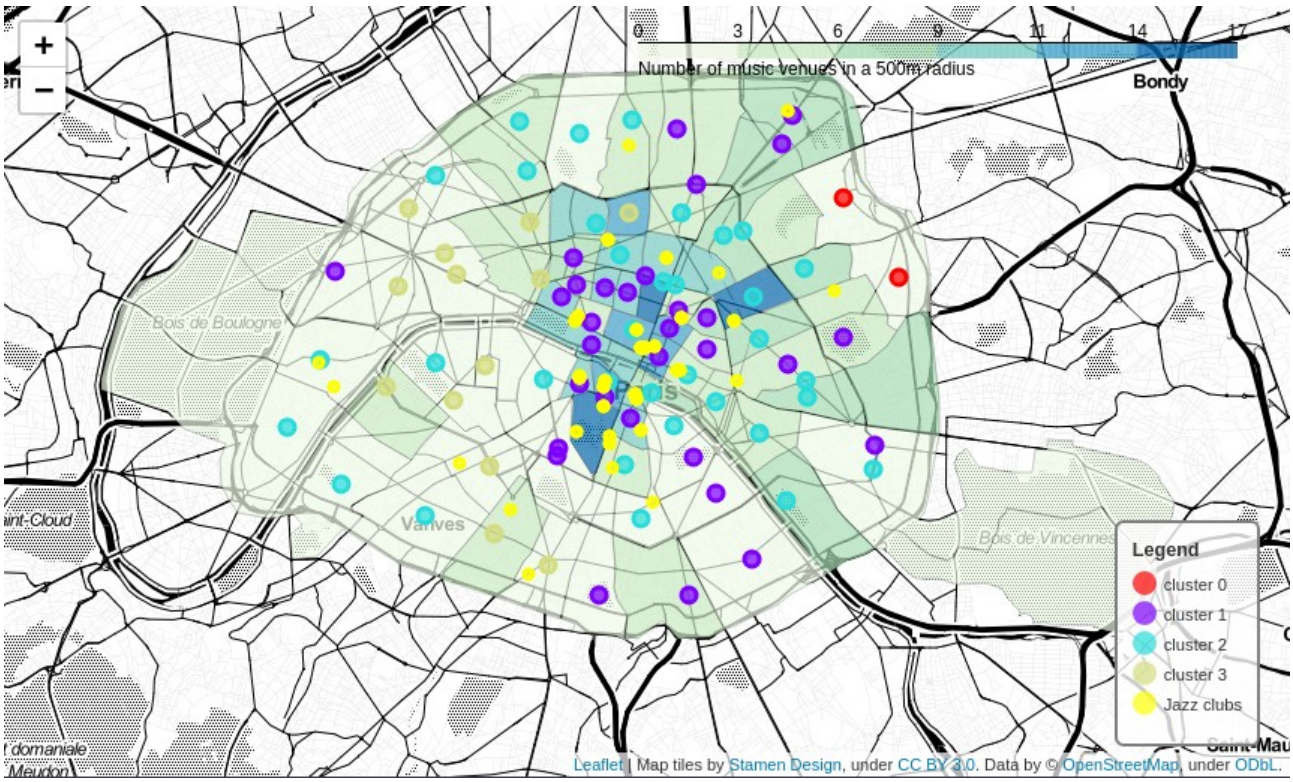
  Even though we can see there are some bars, there are fewer than in clusters 1 & 2. There are some restaurants, but also hotels, sandwich places, cosmetics shops, clothing stores, which are not really aligned with a Jazz Place's standards.

If we superpose the clusters on top of the number of music venues in a 500m radius, we see that this first asumption seems to be somehow validated :



Indeed, we see the top neighborhoods in terms of number of music venues seem to be cluster 1 and 2 mostly, and cluster 3 moderately.

If we add Jazz Clubs identified previously on the map, we see that this hypotheses is pretty accurate :



III.b.iv. Identification of most promising quartiers

I merged the information from the 2 most promising clusters (1 & 2).

Then, I added the number of jazz venues and music venues. I made the hypothesis that the best spot would be the one which minimizes the number of jazz venues, while maximizing music venue counts (which can attract music enthousiasts).

To do so, I considered the number of jazz venues and applied the following formula to normalize the values :

$$x_{new} = 1 - x_{old}/x_{max}$$

With this new metric, the less jazz venue an arrondissement has, the better the value (max value : 1, and min value : 0).

	Quartier Name	Count of Jazz Venue	1 - normalized count
0	Saint-Germain-l'Auxerrois	4.0	0.500
1	Palais-Royal	2.0	0.750
2	Place-Vendôme	2.0	0.750
3	Gaillon	1.0	0.875
4	Vivienne	0.0	1.000
5	Mail	1.0	0.875
6	Bonne-Nouvelle	1.0	0.875
7	Arts-et-Métiers	1.0	0.875
8	Enfants-Rouges	2.0	0.750
9	Archives	2.0	0.750
10	Sainte-Avoye	7.0	0.125
11	Saint-Merri	8.0	0.000
12	Jardin-des-Plantes	0.0	1.000
13	Sorbonne	6.0	0.250
14	Odéon	8.0	0.000
15	Notre-Dame-des-Champs	1.0	0.875
16	Saint-Germain-des-Prés	4.0	0.500
17	Chaussée-d'Antin	1.0	0.875
18	La Roquette	0.0	1.000
19	Picpus	0.0	1.000

I then merged this dataframe with the one containing both music venue counts and INSEE identification of the quartiers.

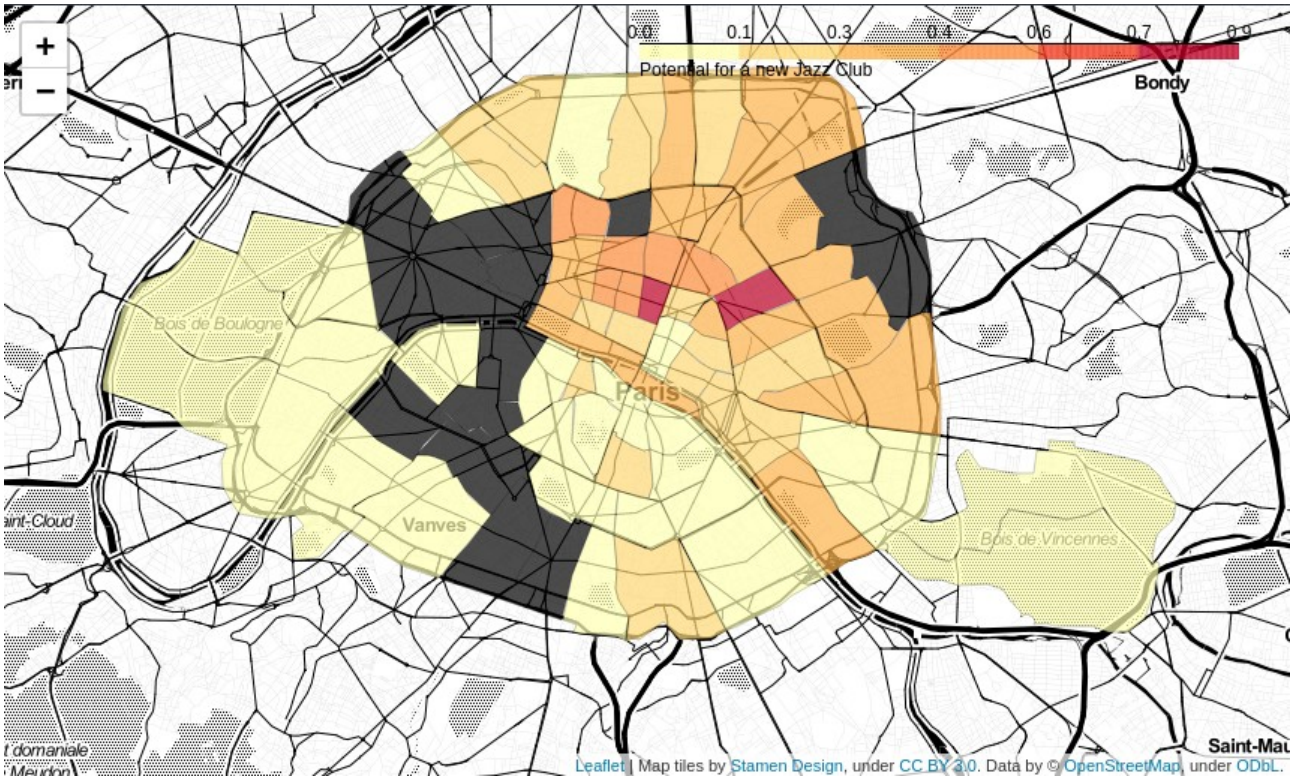
From that, I calculated a new column, called 'rating' and representing the 'potential' of each quartier to welcome our new Jazz Club.

'rating' is defined as follow :

Rating = (1-normalized count) x Music Venue Count

	Quartier Name	Count of Jazz Venue	1 - normalized count	Music Venue Count	INSEE_ID	rating
0	Saint-Germain-l'Auxerrois	4.0	0.500	0.588235	7510101	0.294118
1	Palais-Royal	2.0	0.750	0.294118	7510103	0.220588
2	Place-Vendôme	2.0	0.750	0.529412	7510104	0.397059
3	Gaillon	1.0	0.875	0.470588	7510201	0.411765
4	Vivienne	0.0	1.000	0.529412	7510202	0.529412
5	Mail	1.0	0.875	0.647059	7510203	0.566176
6	Bonne-Nouvelle	1.0	0.875	1.000000	7510204	0.875000
7	Arts-et-Métiers	1.0	0.875	0.294118	7510301	0.257353
8	Enfants-Rouges	2.0	0.750	0.470588	7510302	0.352941
9	Archives	2.0	0.750	0.294118	7510303	0.220588
10	Sainte-Avoye	7.0	0.125	0.705882	7510304	0.088235
11	Saint-Merri	8.0	0.000	0.764706	7510401	0.000000
12	Jardin-des-Plantes	0.0	1.000	0.058824	7510502	0.058824
13	Sorbonne	6.0	0.250	0.529412	7510504	0.132353
14	Odéon	8.0	0.000	0.882353	7510602	0.000000
15	Notre-Dame-des-Champs	1.0	0.875	0.117647	7510603	0.102941
16	Saint-Germain-des-Prés	4.0	0.500	0.529412	7510604	0.264706
17	Chaussée-d'Antin	1.0	0.875	0.470588	7510902	0.411765
18	La Roquette	0.0	1.000	0.235294	7511103	0.235294
19	Picpus	0.0	1.000	0.000000	7511202	0.000000

The following map presents rating metric in each quartier :



IV. Conclusion and further improvements

From this anaysis, we were able to cluster Paris quartiers into 4 distinct groups and identified 2 cluster categories with high potential.

In the last part of the study, we zoomed into these 2 clusters and identified some potential quartier in which to open our new Jazz Club :

	Quartier Name	rating
0	Bonne-Nouvelle	0.875000
1	Folie-Méricourt	0.772059
2	Mail	0.566176
3	Faubourg-Montmartre	0.566176
4	Porte-Saint-Denis	0.566176
5	Vivienne	0.529412
6	Saint-Georges	0.514706
7	Porte-Saint-Martin	0.441176
8	Charonne	0.411765
9	Gaillon	0.411765
10	Chaussée-d'Antin	0.411765
11	Bercy	0.411765
12	Place-Vendôme	0.397059
13	Hôpital-Saint-Louis	0.360294
14	Saint-Ambroise	0.360294

The model could be improved by integrating other information such as pedestrian traffic for a particular time of day (around 21-23h typically).

Moreover, Foursquare doesn't allow to do a search query for a particular time of day / local time. This analysis would need to be carried out on different day / time of day to make sure that our results are actually reproductible.

Another interesting parameter for an investor is to know how much is the price per square meter in the targeted quartier. This could be added to our model to help the investor in the decision-making process.