Bonjour Paris !

The story of using Data Science for moving

Thomas Francillette

IBM Data Science – Applied Data Science Capstone Project
June 2020

Table of Contents

I. Introduction

In the end of 2019, a new coronavirus-related disease started to hit China. Months later, the spread was global, and the disease declared as pandemic by the WHO. Numerous countries took drastic measures to contain the spread, such as quarantines.

As a French engineer working for a pharmaceutical company in Denmark, I have been living for about two years now in Copenhagen. When the coronavirus disease hit the country, and that the Danish government responded by setting a one-month confinement, I started to reflect on my situation and decided that I should go back working in France when everything will be over. At the same time, I had really enjoyed living in Copenhagen, therefore I took the decision not to lose in terms of life quality by moving to Paris.

This project is firstly aiming to leverage Data Science tool in order to provide the best location for my moving from Copenhagen to Paris. In a second part, Data Science will also be used to help me find my feet in a city I left years ago. It is to be noticed that the story used for this project is purely fictional.

II. Data Sourcing

In order to address the problematic, diverse data sources will be used in this project. All data types with associated sources are gathered Table 1 below.

TABLE 1 – PROJECT DATA SOURCING SUMMARY

Data Type	Sources
Copenhagen Address Coordinates	Google Maps
Copenhagen & Paris Venues	Foursquare
Paris Neighborhoods Coordinates	Paris City Hall Website
Paris Transportation Stations Coordinates	RATP (Paris Transportation) Website

i. Copenhagen Address Coordinates

Starting address is know therefore it is easy to find the associated coordinates by directly typing the address on Google Maps. No cleaning is needed for this data.

ii. Copenhagen and Paris Venues

Requests are passed through Foursquare API and a json file gathering information about the venues around the location is obtained. Only the name and categories of the venues were retained.

The results are dataframes of 59 rows x 2 columns and 5200 rows x 7 columns for respectively Copenhagen and Paris (see Figure 1 and Figure 2 below).

	Name	Category
0	Sound Station	Music Store
1	Hart Bageri	Bakery
2	Juul's Vin og Spiritus	Wine Shop
3	Falernum	Wine Bar
4	Pizzicato	Pizza Place
5	Ganni	Women's Store
6	Meyers Deli	Deli / Bodega
7	Social Foodies	Ice Cream Shop
8	Ipsen & Co	Café
9	Vinstue 90	Bar
10	RIST Kaffebar	Coffee Shop
11	Central Hotel & Café	Café

FIGURE 1 – COPENHAGEN (TARGET ADDRESS) VENUES LIST

	Neighborhood	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Notre-Dame-des-Champs	48.846428	2.327357	Legend Hotel	48.845316	2.325507	Hotel
1	Notre-Dame-des-Champs	48.846428	2.327357	Gilles Verot	48.847118	2.326819	Deli / Bodega
2	Notre-Dame-des-Champs	48.846428	2.327357	Sadaharu Aoki | 青木定治	48.848013	2.330366	Dessert Shop
3	Notre-Dame-des-Champs	48.846428	2.327357	Marché de Raspail	48.848807	2.327526	Market
4	Notre-Dame-des-Champs	48.846428	2.327357	Bagels & Brownies	48.846537	2.327329	Bagel Shop
...	...	...	...	...	...	...	...
5195	La Chapelle	48.894012	2.364387	New-Thaï San	48.891324	2.361265	Asian Restaurant
5196	La Chapelle	48.894012	2.364387	Carrefour City	48.889998	2.361442	Supermarket
5197	La Chapelle	48.894012	2.364387	Restaurant Tin Tin	48.891163	2.360850	Chinese Restaurant
5198	La Chapelle	48.894012	2.364387	Le Five Paris	48.896396	2.362536	Soccer Field
5199	La Chapelle	48.894012	2.364387	O'Tacos La Chapelle	48.896282	2.359561	Mexican Restaurant

5200 rows × 7 columns

Figure 2 – Paris Neighborhood Venues List

iii. Paris Neighborhoods Coordinates

The initial .csv file contains a lot of data regarding localization, perimeter or neighborhoods geometry. Only the columns containing name and coordinates were kept. For the later, both latitude and longitude were contained in the same column and the *STRIP* function was used to recover both information in separate columns.

The result is a dataframe of 80 rows and 3 columns (see Figure 3 below).

	Neighborhood	Latitude	Longitude
0	Notre-Dame-des-Champs	48.846428	2.327357
1	Petit-Montrouge	48.826653	2.326437
2	Pont-de-Flandre	48.895556	2.384777
3	Muette	48.863275	2.259936
4	Chaillot	48.868434	2.291679
...	...	...	...
75	Ternes	48.881178	2.289964
76	Val-de-Grâce	48.841684	2.343861
77	Necker	48.842711	2.310777
78	Père-Lachaise	48.863719	2.395273
79	La Chapelle	48.894012	2.364387

80 rows × 3 columns

Figure 3 – Paris Neighborhood Coordinates

iv. Paris Transportation Stations Coordinates

As for Paris neighborhoods the collected file is a .csv but more cleaning work is need. First step was to split the columns (*STRIP* function was again used) containing multiple information that were:

- Coordinates with both Latitude and Longitude of the stations
- Description with Address and Postal Code of the stations

Once the split performed, lines were filters regarding the targeted Postal Code with the *ISIN* function. The final cleaning was about the stations' names, that were randomly written in upper or lowercase and with or without hyphen for the same station name. The following steps were performed for the name's harmonization:

- All names were put in uppercase
- All accents were removed using the *ENCODE* and *DECODE* functions
- Spaces before and after hyphen were removed and hyphen were replaced by a space character

The result is a dataframe of 264 rows and 6 columns (see below).

	ID	Name	Address	Postal Code	Latitude	Longitude
0	1927	DUROC	bd du Montparnasse	75107	48.846849	2.316937
1	2253	DUROC	bd du Montparnasse	75107	48.846993	2.316542
2	3343757	ECOLE MILITAIRE	Avenue Duquesne	75107	48.854261	2.305440
3	3749897	ECOLE MILITAIRE	FACE 3 PLACE JOFFRE	75107	48.854090	2.304963
4	3765248	SAINT GUILLAUME	183-185 BOULEVARD SAINT GERMAIN	75107	48.854624	2.329478
...	...	...	...	...	...	...
259	3813045	CHAMP DE MARS	AVENUE JOSEPH BOUVARD	75107	48.855076	2.296028
260	1638	VARENNE	13 boulevard des Invalides	75107	48.856393	2.314754
261	4009626	SEVRES BABYLONE	39 BOULEVARD RASPAIL	75107	48.851910	2.326836
262	4022886	ASSEMBLEE NATIONALE	241 BOULEVARD SAINT-GERMAIN	75107	48.861544	2.320037
263	3813125	VAUBAN HOTEL DES INVALIDES	1 AVENUE DE TOURVILLE	75107	48.853444	2.311269

264 rows × 6 columns

FIGURE 4 – PARIS TRANSPORTATION STATIONS COORDINATES

III. Methodology

Two different machine learnings were used:

- **Recommender Systems** for using its property to match an output with a user input profile, in our case the venues from the initial location in Copenhagen compared to Paris neighborhoods.
- **K-Means Clustering** for its property of finding patterns in a set of data based on criteria similarity, in our case the similarity of different venues categories around transportation stations.

IV. Results and Discussion

There is two parts in the project:

- **First Part** is about finding the best location in Paris
- **Second Part** is about exploring the surroundings

i. Finding the Best Location in Paris

For the first part, a recommender system approach will be performed in order to match the initial neighborhood with Parisian ones and find the one that matches the most.

User profile dataframe is established by screening the venues around the address in Copenhagen and normalize the counts of each venues. A section of the resulting table is given below, and it can be seen that Bakery and French Restaurant are important factors of the initial location in Copenhagen (already a bit of France!).

	Category	Score
0	Asian Restaurant	0.333333
1	Bakery	1.000000
2	Bar	0.000000
3	Bookstore	0.333333
4	Burger Joint	0.000000
5	Café	0.666667
6	Cheese Shop	0.000000
7	Cocktail Bar	0.333333
8	Coffee Shop	0.666667
9	Deli / Bodega	0.333333
10	Food Service	0.000000
11	French Restaurant	1.000000
12	Furniture / Home Store	0.000000
13	Gift Shop	0.000000

FIGURE 5 – USER PROFILE SCORE CARD

Next step is to take the Paris neighborhood venue table and multiply each venue by the corresponding user profile score. If the venue category is not in the user profile, the score will be zero for the corresponding venue. After that, all venues scores are summed by neighborhood to obtain each neighborhood score. Representation of the final results are given Figure 6 below.

By looking at the results, the “Gros Caillou” neighborhood is clearly highlighted by the analysis, with more than 10 points than the second neighborhood. Gros Caillou is therefore selected for the second part of the project.

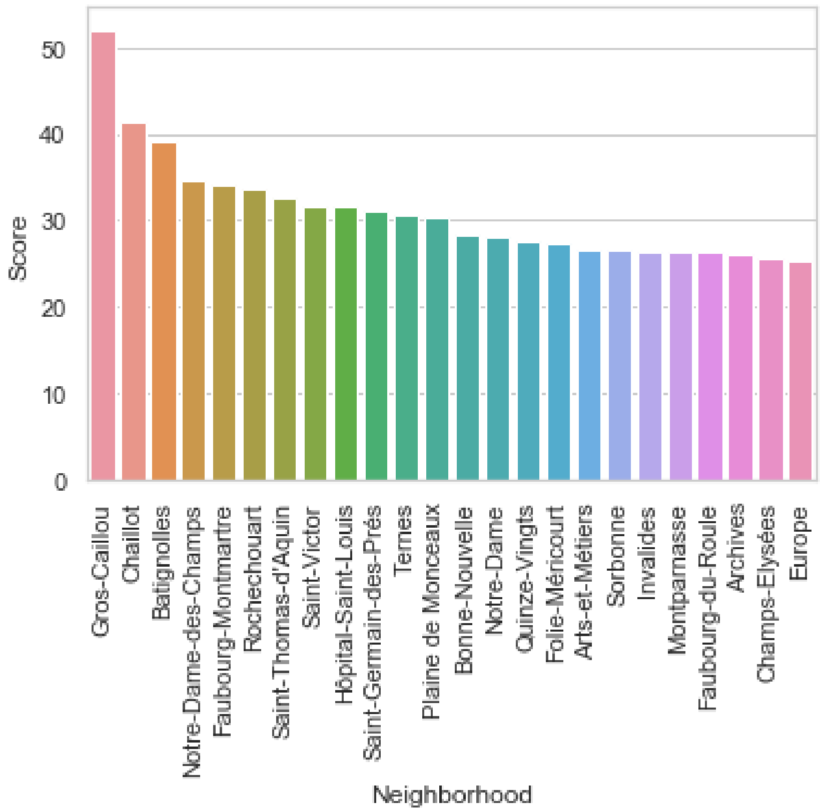

Figure 6 – Paris Neighborhoods Ranking

ii. Exploring the Surroundings

Gros Caillou was selected as target Parisian neighborhood. All transportation stations locations in the neighborhood are checked in terms of surrounding venues. Then, K-means clustering approach will be used in order to find patterns in Gros Caillou venues.

First thing is to collect the venues for each transportation station and results are stored in a database where the five most common venues are ranked (see Figure 7 below).

	Metro	1st	2nd	3rd	4th	5th
0	1638	Café	Garden	Vietnamese Restaurant	Diner	Clothing Store
1	1666	Cocktail Bar	Hotel	Garden	Vietnamese Restaurant	Diner
2	1667	Chocolate Shop	Garden	Hotel	Tailor Shop	French Restaurant
3	1669	Coffee Shop	French Restaurant	Café	Salad Place	Bakery
4	1690	French Restaurant	Hotel	Bakery	Supermarket	Italian Restaurant
...	...	...	...	...	...	...
245	7653685	Café	Coffee Shop	Salad Place	Bakery	French Restaurant
246	7653686	Bus Stop	Food Truck	Vietnamese Restaurant	Diner	Clothing Store
247	7653687	Bus Stop	Vietnamese Restaurant	Diner	Clothing Store	Cocktail Bar
248	7653688	French Restaurant	Bus Stop	Vietnamese Restaurant	Diner	Clothing Store
249	7653719	French Restaurant	Vietnamese Restaurant	Diner	Clothing Store	Cocktail Bar

250 rows × 6 columns

Figure 7 – Most Common Venues of Gros Caillou Transportation Stations

In parallel, K-Means clustering (with k = 6) is performed on the stations surrounding venues counts. Most common venues and clusters are joined together in a single table (see Figure 8 below).

	ID	Name	Address	Postal Code	Latitude	Longitude	Cluster	1st	2nd	3rd	4th	5th
0	1927	DUROC	bd du Montparnasse	75107	48.846849	2.316937	4.0	Salon / Barbershop	Bakery	Vietnamese Restaurant	Diner	Clothing Store
1	2253	DUROC	bd du Montparnasse	75107	48.846993	2.316542	4.0	Salon / Barbershop	Bakery	Vietnamese Restaurant	Diner	Clothing Store
2	3343757	ECOLE MILITAIRE	Avenue Duquesne	75107	48.854261	2.305440	5.0	French Restaurant	Plaza	Hotel	Dessert Shop	Chocolate Shop
3	3749897	ECOLE MILITAIRE	FACE 3 PLACE JOFFRE	75107	48.854090	2.304963	5.0	French Restaurant	Plaza	Hotel	Dessert Shop	Chocolate Shop
4	3765248	SAINT GUILLAUME	183-185 BOULEVARD SAINT GERMAIN	75107	48.854624	2.329478	1.0	Italian Restaurant	Tailor Shop	Sandwich Place	Diner	Clothing Store
...	...	...	...	...	...	...	...	...	...	...	...	...
259	3813045	CHAMP DE MARS	AVENUE JOSEPH BOUVARD	75107	48.855076	2.296028	1.0	Café	Bakery	French Restaurant	Pizza Place	Vietnamese Restaurant
260	1638	VARENNE	13 boulevard des Invalides	75107	48.856393	2.314754	0.0	Café	Garden	Vietnamese Restaurant	Diner	Clothing Store
261	4009626	SEVRES BABYLONE	39 BOULEVARD RASPAIL	75107	48.851910	2.326836	1.0	Chocolate Shop	Art Gallery	Garden	Hotel	Tailor Shop
262	4022886	ASSEMBLEE NATIONALE	241 BOULEVARD SAINT-GERMAIN	75107	48.861544	2.320037	5.0	French Restaurant	Bus Stop	Vietnamese Restaurant	Diner	Clothing Store
263	3813125	VAUBAN HOTEL DES INVALIDES	1 AVENUE DE TOURVILLE	75107	48.853444	2.311269	1.0	Plaza	Restaurant	Garden	Diner	Vietnamese Restaurant

250 rows × 12 columns

FIGURE 8 – GROS CAILLOU STATIONS AND ASSOCIATED VENUES AND CLUSTERS

There are 6 different clusters in total, by looking at their relative abundance (see Figure 9 below), it can be seen that one cluster is in majority: Cluster 1. This can be seen also on the Folium map rendering (see Figure 10 – same clusters color code as Figure 9). Another thing that is noticeable is that there is no geographical logic in the clustering, meaning that whatever is inside the clusters, and Cluster 1 for instance, it is spread across the neighborhood.

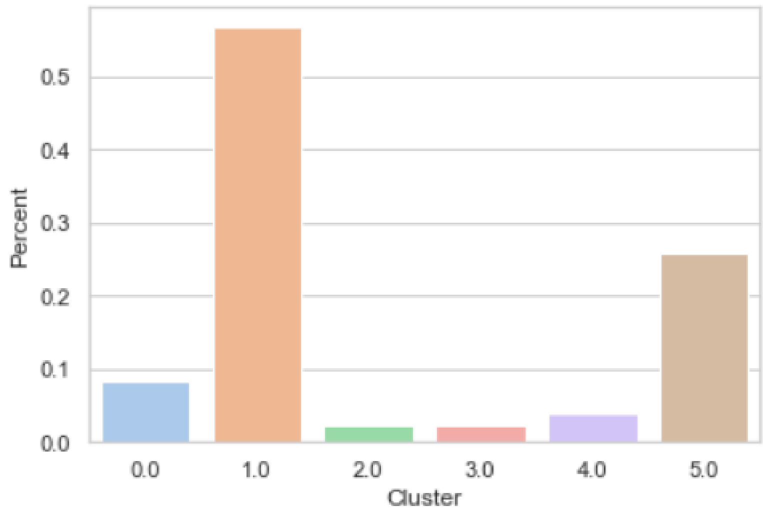

FIGURE 9 – GROS CAILLOU CLUSTERS RELATIVE ABUNDANCE

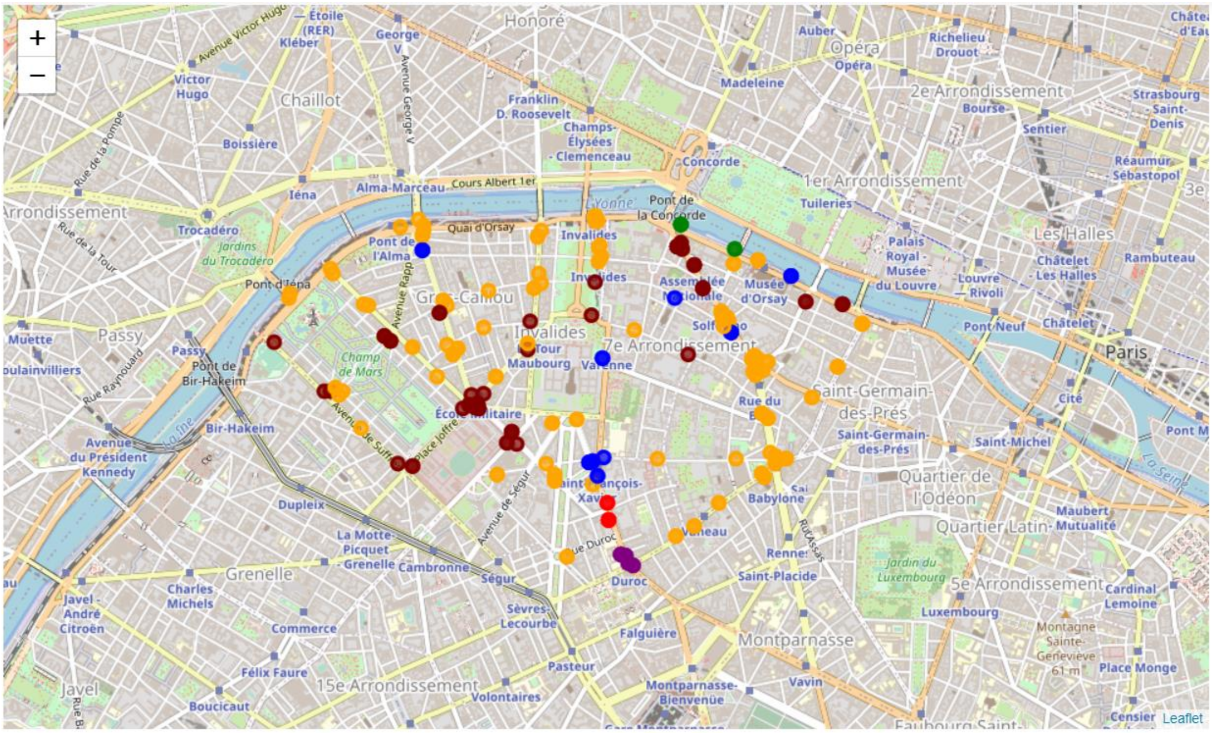

FIGURE 10 – GROS CAILLOU STATIONS MAP AND ASSOCIATED CLUSTERS

To better apprehend what behind each cluster, most common venues are plotted for each cluster (see Figure 11). Low abundance clusters are correlated with low venue categories diversity (which is probably what helped the algorithm to separate these clusters). These clusters are highlighting the Bus Stop, Gym and Salon/Barbershop commodities (which is important to locate when moving to a new city).

By analyzing the most abundant cluster (see Figure 11), it can be seen that there is a majority of French restaurant, plaza and cafe. This is showing that Gros Caillou neighborhood is probably a combination of a historical neighborhood with highly developed tourism, as majority of venues are very "French".

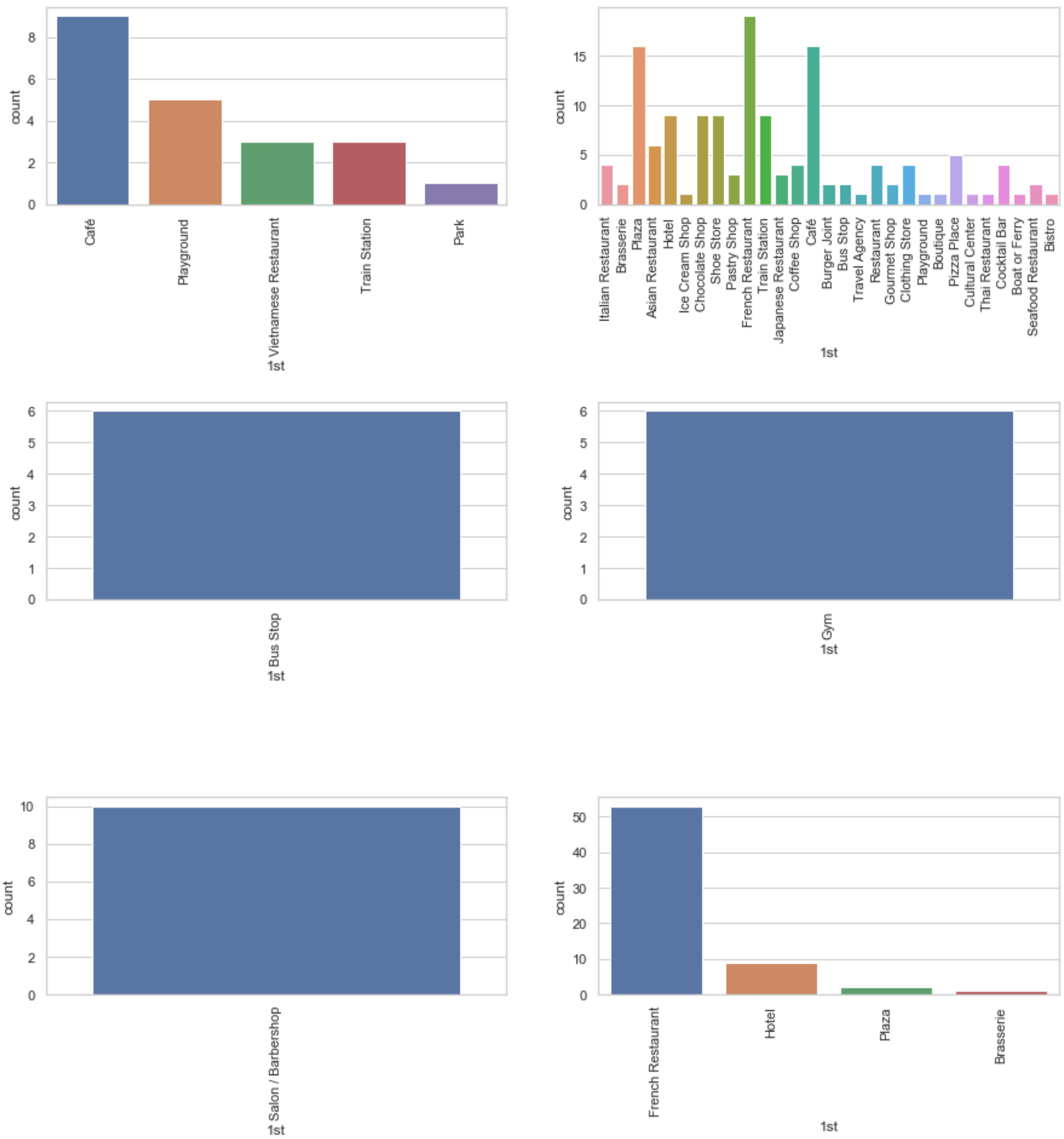

FIGURE 11 – GROS CAILLOU CLUSTERS MOST IMPORTANT VENUES

(Top Left = Cluster 0, Top Right = Cluster 1, Middle Left = Cluster 2, Middle Right = Cluster 3, Down Left = Cluster 4, Down Right = Cluster 5)

V. Conclusions

By combining recommender system and clustering approach, we were able to find a neighborhood in Paris similar to the initial neighborhood of Copenhagen, and to further explore this neighborhood to better understand its dynamic.

This approach was extremely easy to implement and is easy transposable to any city in any country. Moreover, this approach is not limited to the moving case study and can be implemented in the case of searching for a place to set up a shop for instance.