

What and Where Invest?

Yassin Belhareth

January 2019

1 Introduction

To start a business project, a comprehensive study must be carried out. One of the most important things to consider is, the place of the project, The regions, which varies depending on its services, there are residential regions that are characterized by specific services, and there are those with a greater vitality where services could be high. Moreover, there are investors who are hesitant about the type of commercial project. In this study, the objective is to extract the adequate information and facilitate it for the investor in order to choose the right place and type for his project. It will be analysis the neighborhoods at the level of the business places that give service for people, to do this, it will be relying on a clustering algorithm in order to segment the neighborhoods in similar clusters, more details in the methodology section. The approach will be applied to the city of Paris, which is the most important city in terms of culture and economy. Tourism attendance reached a record high in the first half of 2018 ¹, with a total of 17.1 million hotel arrivals. Therefore, this could allow investors to invest in it.

2 Data Description

It need in this study for some data about venues in the neighborhoods, there are many providers could give data about venues, for example Yelp², Google

¹<http://www.lefigaro.fr/conjoncture/2018/08/27/20002-2\0180827ARTFIG00245-tourisme-paris-bat-tous-les-records.php>

²Yelp: <https://www.yelp.com/developers>

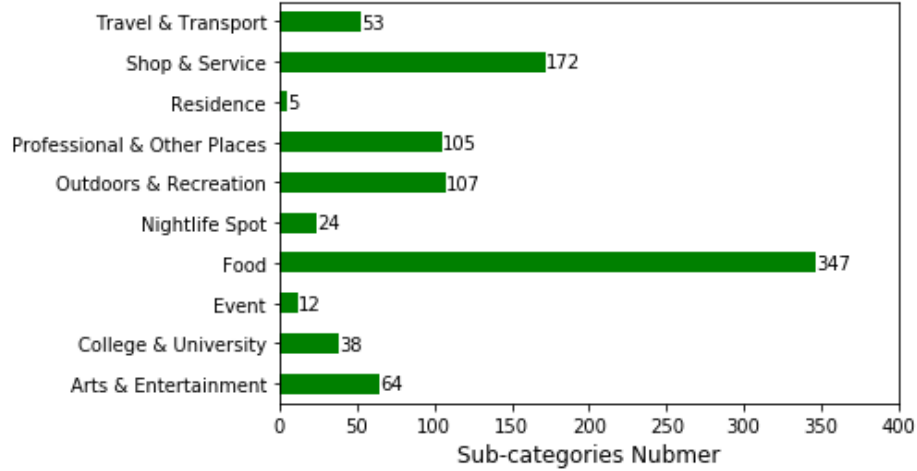


Figure 1: Number of Sub-categories by Categories in Foursquare

Places³ and Foursquare⁴.it were chosen the Foursquare API because it is quite straightforward and easy to use. It organizes its venues in categories and sub-categories, the bar chart 1 shows the number of sub-categories for each category, the food category contains the maximum number, followed by the shop&service category, the sub-categories represent all types of categories in the world (e.g. the category food contains: French Restaurant, Japanese Restaurant, pizza place, etc.). In other hand, I should collect the neighborhoods names of the city. Wikipedia page [2] contains the required data, but the name of the neighborhood is not enough to collect venues data, so we must have the latitude and longitude coordinate for each location, fortunately, if you click on any location in the Wikipedia page, you will have detailed information on it, among them its geographic coordinate (the figure 2 shows an example).

³Google Places :<https://cloud.google.com/maps-platform/places/>

⁴Foursquare : <https://developer.foursquare.com/places-api>


Arrondissement ^{1, n 1}	Quartiers	Population en 1999 (hab.) ²	Superficie (ha) ²
1 ^{er} arrondissement dit « du Louvre »	1 ^{er} Saint-Germain-l'Auxerrois	1 672	86,9
	2 ^e Halles	8 984	41,2
	3 ^e Palais-Royal	3 195	27,4
	4 ^e Place Vendôme	3 044	26,9
2 ^e arrondissement dit « de la Bourse »	5 ^e Gaitan	1 345	18,8
	6 ^e Vivienne	2 917	24,4
	7 ^e Mail	5 783	27,8
3 ^e arrondissement dit « du Temple »	48° 51' 45" nord, 2° 20' 41" est	9 595	28,2
	responsants.	9 560	31,8
		8 562	27,2
		8 609	36,8
	Quartier des Halles	7 501	21,3
		6 523	31,3

Figure 2: Example of London Neighborhoods and How to Get Geographic Coordinate

3 Methodology

3.1 Data Analysis

As indicated in the data description section, we collected data on the city of Paris. The collection is divided into 2 tasks:

- The first task is to collect the names of the city’s neighbourhoods and their geographical coordinates. The city is composed of 20 boroughs and 80 neighbourhoods, with a population of over two million people. The data collected on the populations by neighbourhood were collected in 1999, and on the one hand, there are no more recent data, on the other hand, the population of the year 1999 is very close to the year 2019 (according to INSEE⁵), so we have decided to use it. The figure 3 shows the relation between the distance from the city centre for each neighbourhood and the population. The correlation is approximately linear(i.e. the population increases when the distance from the city centre increases too).
- The second task involves collecting the existing venues in the neighbourhoods, because of financial constraints, we have only collected 100 nearest venues for each neighbourhood. We obtained 5728 Venues, as shown in the figure 4 57% are from food category, followed by the service&shop category that represents 12%, which explains the high demand for these two types of venues. However, 50% of sub-categories exist less than 5 times, which explains the cultural diversity of the city, but there are sub-categories that are frequent, such as French Restaurant and Pizza place.

3.2 Features

To carry out any method, we must assign to each neighbourhood a set of features, the first represents the number of populations, it is an essential factor that allows a future trader to choose the appropriate region for his project.

The second feature depends on the type of place, which will be represented by its number of occurrences in each neighborhood. As we have seen in

⁵INSEE: National Institute of Statistics and Economic Studies

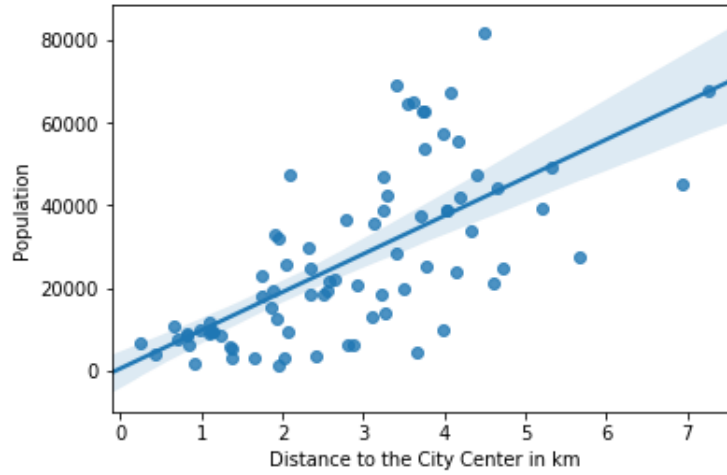


Figure 3: The relationship between Distance to the City Center and the Population

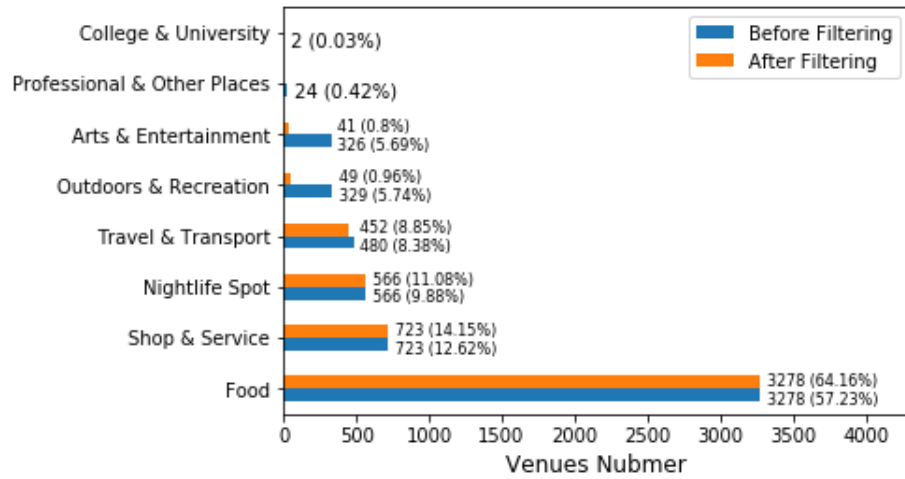


Figure 4: Venues Number according to each Category

Figure 1, there are commercial categories and non-commercial categories. So we have tried to filter the categories of non-commercial venues, table 4 shows the categories of remaining venues (it is possible that we have not put all the commercial categories). After filtering the category types, we deleted the locations obtained by the Foursquare API, which does not belong to Table 1.

Categories	Sub-Categories
Food	All Sub-Categories.
Arts & Entertainment	Bowling Alley, 'Casino, Circus, Country Dance Club, General Entertainment, Laser Tag, Movie Theater, Drive-in Theater, Indie Movie Theater and Multiplex.
Outdoors & Recreation	Gym / Fitness Center, Boxing Gym, Climbing Gym, Gymnastics Gym, Gym and Outdoor Gym.
Travel & Transport	Hotel, Bed & Breakfast, Boarding House, Hostel, Hotel Pool, Inn, Motel, Resort and Vacation Rental.
Nightlife Spot	All Sub-Categories.
Shop&Service	All Sub-Categories.

Table 1: List of Remaining Sub-Categories

3.3 Method

There are several types of segmentation algorithms, but the most important are: Density Based Clustering, Hierarchical Clustering and Partitioned based Clustering.

- The first, it is based on the location of high density region and the separation of outliers.
- The second, it considers that each observation forms a class, and with a measure of similarity between observations, it tries to reduce the number of classes through the grouping of similar observations into a single class, this type is used for a small number of observations.
- The third, it divides the data into k groups, the k is chosen by the user, each group contains the most similar observations between them according to its characteristics.

According to the objectives and explanations of 3 types indicated above, Partitioned based Clustering is the appropriate type for our approach. One of the most used algorithms in this type is k-means [1], it is relatively efficient and used for medium and large data.

4 Result

Before obtaining the segmentation result, we must specify the number of clusters. There are several methods, but the most used is the Elbow method, the figure 5 shows a curve where the x-axis represents the number of k and the y-axis is the sum of squared errors(see Equation (1), with k is the number of clusters, C the set of objects in a cluster, m the centroid point of a cluster.), the best point is the elbow point with stable SSE values for the following points, therefore, the point with k equals to 4 is the best point.

$$SSE = \sum_{i=1}^k \sum_{p \in C_i} (p - C_i) \quad (1)$$

Picture 7 shows the different neighborhoods of the city of Paris, and each location is colored according to its corresponding cluster. we notice that the clusters are distributed around the city center in an orderly manner.

To understand the characteristics of each cluster, we must analyze them at the level of the number of populations on the one hand, and at the level of categories (sub-categories) in terms of number and type, on the other hand. Figure 6 shows the distribution of the number of populations for each cluster, Figure 8 shows the number of locations by category for each cluster, according to the two figures, clusters can be defined as follows:

- Cluster 0: very low population density, very high service density.
- Cluster 1: very high population density, low service density.
- Cluster 2: high population density, very low service density.
- Cluster 3: low population density, high service density.

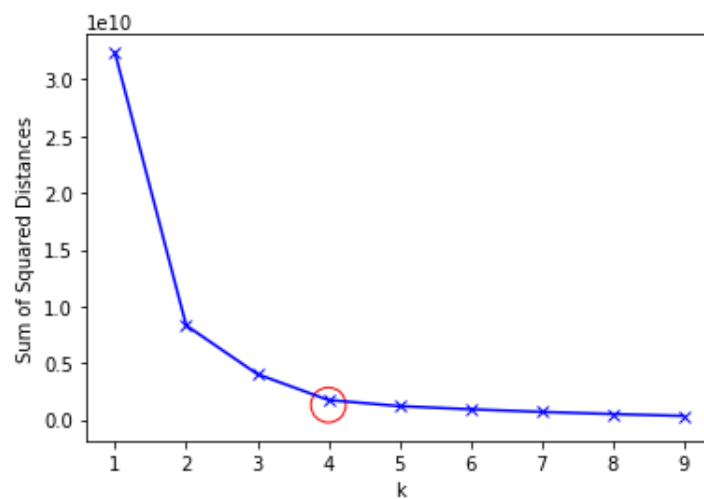


Figure 5: Elbow Method For Optimal k

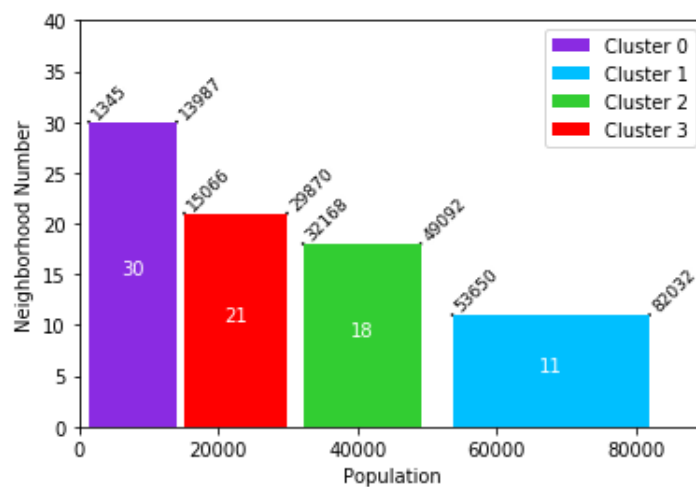


Figure 6: Population Histogram by Clusters

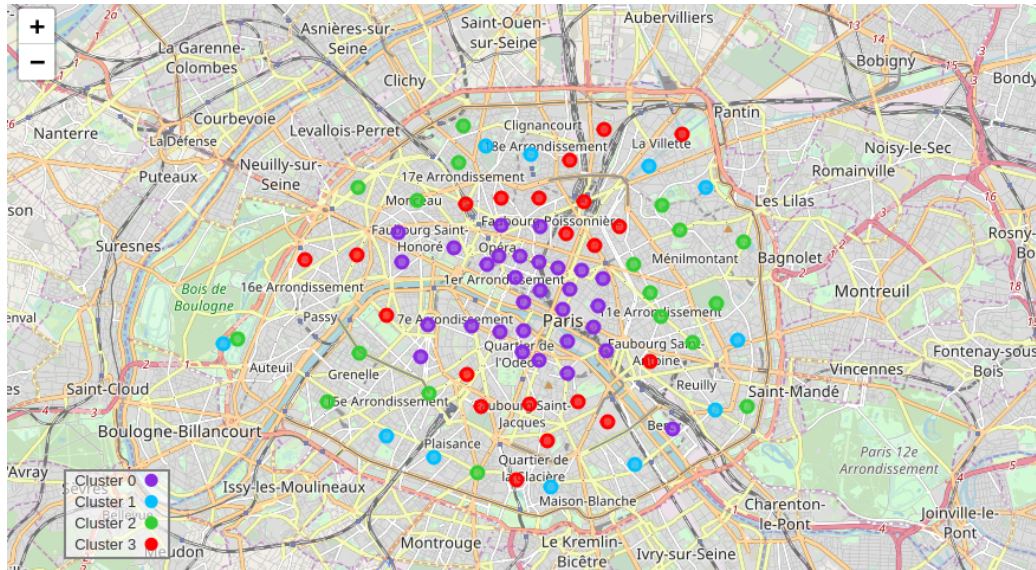


Figure 7: Map of Paris with the Colourful Neighbourhoods in Clusters

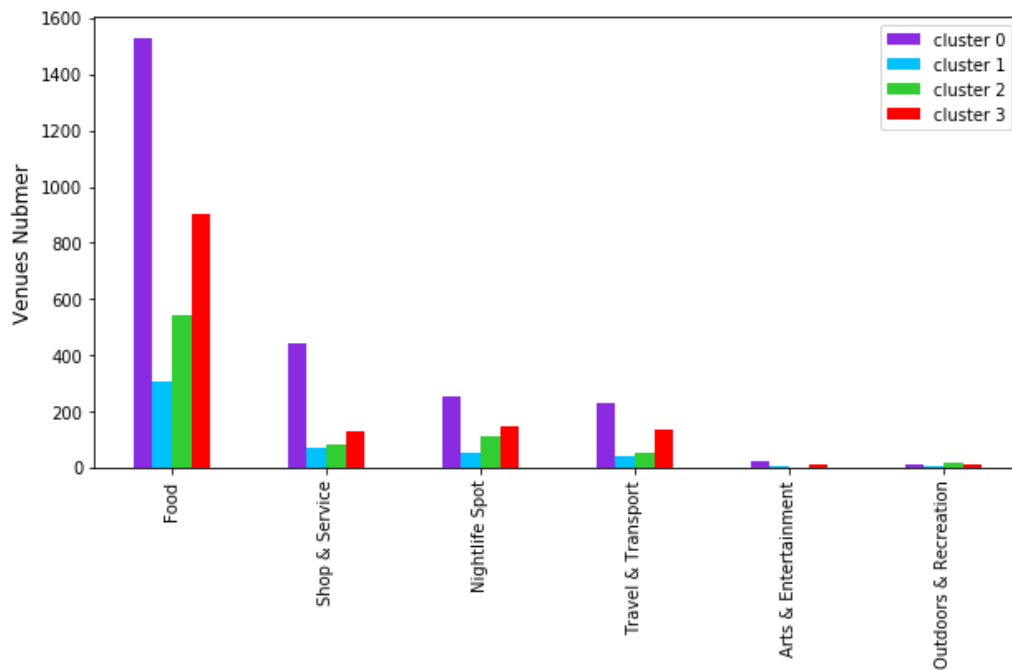


Figure 8: Venues Number by Category for each Cluster

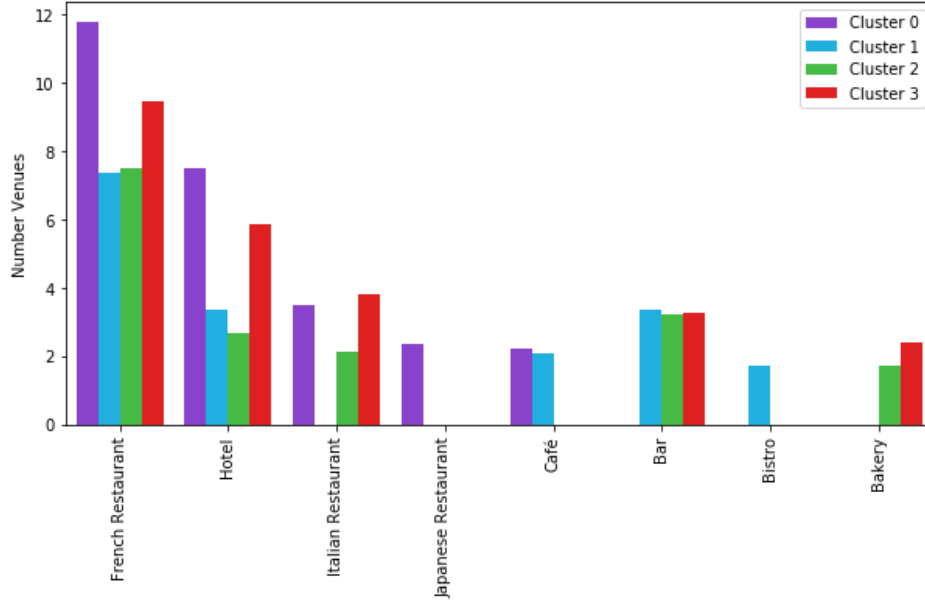


Figure 9: Venues Number by top 5 sub-categories for each Cluster

5 Discussion

Now, what and where invest? To answer this question, we could extract the top sub-categories venues in each cluster, which they are in high demand. Figure 9 shows the top 5 sub-categories for each cluster, we notice that French restaurant category is the most frequent type in all clusters, followed by the Hotel category for cluster 0 and 3 and Bar category for cluster 1 and 2. On the other hand, if a future trader wants to open a hotel, then we advise him by cluster 1 or cluster 2, and if he wants to open a Japanese restaurant, we recommend him by cluster 0 since the demand is high in this cluster. But this study remains approximate, since we have treated just 100 places for each neighbourhood. The analysis could go further, by collecting information on the available places on the one hand, and on the other hand a collection of information on the economic performance of each neighbourhood.

6 Conclusion

We have segmented the districts of the city of Paris into groups, the segmentation is done by the K-means segmentation algorithm. Following the Elbow method, we divided the districts into 4 groups. As we have seen in the results section, the segmentation were done on two levels: population density and service density. the result obtained gives us residential and other commercial clusters in varying proportions. in addition, this study gives an idea of the types of frequent and infrequent venues and their geographical coordinates.

References

- [1] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108, 1979.
- [2] Wikipedia. Liste des quartiers administratifs de Paris — Wikipedia, the free encyclopedia. <http://fr.wikipedia.org/w/index.php?title=Liste%20des%20quartiers%20administratifs%20de%20Paris&oldid=156283007>, 2019. [Online; accessed 19-March-2019].