Finding the right neighbourhood to move to in Paris

Francesco Siciliano

July 13th 2019

"Quality of life actually begins at home - it's in your street, around your community."

- Charles Kennedy -

1. Introduction

1.1. Background

Finding a house in Paris, whether to buy or to rent, is by itself a daunting task. This gets even trickier when considering the diversity presented by each of its neighbourhoods, so-called *arrondissements*. Looking for an apartment does not only limits to the apartment itself: a neighbourhood fitting your needs is essential to live a happy life, whether you are looking for a lively place filled with bars and clubs, or a quieter place, surrounded by parks or museums.

1.2. Problem

The problem tackled by this project is to provide a tool to people looking for houses/apartments to move into in the city of Paris, specifically not meant to look for the house itself, but rather to look for some zones that satisfies the user requirements, allowing to narrow down the research to fewer zones of the city.

1.3. Interest

Developing this project to its full potential would prove to be quite handy to people in their search for their homes: especially for someone moving to a new city, using this tool will help to direct the search only towards portions of a given city that are deemed fit by the user, in term of what it was actually is looking for from a neighbourhood. The choice of an apartment involves investing substantial amounts of money, it is for this reason that deciding to move out shortly after having bought/rented it would be quite a nuisance (other than a waste of money and time). Knowing beforehand where your potential home is located is therefore essential to be more confident in your choice and avoiding troublesome situations when looking for new places where to live.

2. Data acquisition and cleaning

2.1. Data sources

The data necessary for this study come mainly from three sources:

- Paris Open Database: this website hosts several useful datasets storing many various information about Paris. For this study, two datasets have been used:
  - One for the *arrondissements*, made of 20 entries and 12 features,
  - One for the *administrative quartiers* that compose the various arrondissements. It consists of 80 entries with 10 features.

  Of the several features, the essential ones are the coordinates of both the *arrondissements* and the *administrative quartiers*, necessary to establish the distribution of the several venues throughout Paris and also to define travel distances;
- Foursquare API queries: thanks to these, it will be possible to determine the amount and more importantly the categories of the points of interest present in Paris. This data will be essential, as they will be used to perform the clustering of Paris quartiers, from which the user will then choose the more suitable for her interests;
- TomTom API queries: TomTom developer portal provides similar services to Google Maps API, but for a lower price, free if performing a limited number of requests, as in our case. It will be used for filtering the locations belonging to the selected cluster, providing as final result the neighbourhood with the shortest travelling time to reach the user workplace.

In addition to these, the average rent prices per *arrondissement* have been retrieved from this article and used to give a visual input to the user to understand which are the cheapest *arrondissements*.

2.2. Data cleaning

Luckily, the *arrondissements* and *administrative quarters* datasets are already well structured and ready to be used. The only intermediate step is to create the dataframes starting from the respective .json files. The data obtained from the query to the Foursquare API need a bit more operation, since first of all it is necessary to extract the venues and arrange them according to their categories. Finally, the output of the TomTom API request is given in form of an xml file. In order to retrieve the useful information i.e. travelling time from potential candidate and the predefined workplace and then, only for the optimal candidate, the list of coordinates of all of the points constituting the route.

2.3. Features selection

The dataframes from Paris Open Data are necessary to define the boundaries of the research and to identify the several sections of the city. The main features are the quartiers' coordinates, their name, the *arrondissement* of belonging and the geometrical shapes of the *arrondissements* themselves.

The output from Foursquare provided the venues that are then sorted and placed in their respective administrative quartier.

The output from TomTom, finally, provided the distances from the neighbourhood candidates to a possible workplace, as well as its most probable route to reach it.

3. Exploratory data analysis

During the first iterations of the research there have been one main issue: a large percentage of the venues reported by Foursquare were of the 'Restaurant' category. This led to a bias in the first clustering attempt, with a large cluster of places with restaurants as common denominator. From this derived the decision to neglect all the restaurant-type venues, as there really are plenty and it would not be a relevant factor to take into account when defining the search parameters for a neighbourhood. You fill find a nice restaurant nearby your house anywhere your house is in Paris, rest assured! Another category that has been neglected are the hotels, since the search is aimed precisely at finding a place to stay.

The choice of using neighbourhood clustering is due to the fact that a search by venue category resulted as too specific and widespread, not really resulting in narrowing down the pool of suitable user choices. It must be noted, however, that the search can be easily modified in order to do a search by venue category by skipping the clustering step and, grouping the venues by category and performing the travel time distance comparison by using the category type instead of the cluster label.

4. Neighbourhood clustering

In this project, *k-means* clustering has been used in order to partition the venues in larger groups in order to speed-up the research. It has been decided to do a 5 groups partitioning. Once the k-means clustering algorithm has been performed, each venues can be found in one and only one of these 5 clusters (thanks to the clusters being mutually exclusive); by adding the respective cluster label to each venue, it is then possible to create separate dataframes each corresponding to a different cluster, like the one shown below:

	Arrondissement	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	1	0	Café	Plaza	Exhibit	Coffee Shop	Historic Site
1	1	0	Pizza Place	Wine Bar	Café	Clothing Store	Bakery
4	2	0	Bistro	Wine Bar	Salad Place	Plaza	Creperie
13	4	0	Ice Cream Shop	Bakery	Wine Bar	Plaza	Coffee Shop
14	4	0	Plaza	Gastropub	Historic Site	Bakery	Park
19	5	0	Bakery	Bar	Wine Bar	Museum	Bistro
20	6	0	Plaza	Wine Bar	Bistro	Café	Bar
24	7	0	Café	Coffee Shop	Ice Cream Shop	Bistro	Pizza Place
25	7	0	Plaza	Café	History Museum	Art Museum	Cultural Center
35	9	0	Bakery	Bistro	Coffee Shop	Music Venue	Bar
37	10	0	Wine Shop	Coffee Shop	Theater	Grocery Store	Food & Drink Shop
42	11	0	Bar	Bakery	Wine Bar	Bistro	Coffee Shop
43	11	0	Bar	Café	Wine Bar	Bistro	Coffee Shop
44	12	0	Plaza	Playground	Sports Club	Recreation Center	Café
49	13	0	Bar	Park	Bakery	Indie Movie Theater	Gaming Cafe

Figure 1 Dataframe related to cluster 0.

The dataframes produced are then given as input in the search for the ideal neighbourhood.

The following map shows the distribution of the cluster throughout the city of Paris:

Figure 2 Map of Paris with clusters elements. By clicking on each element, the administrative quartier name and cluster label of belonging are shown.

5. Data visualisation

In order to have an immediate and clear representation of the user search results, a map showing the results is likely to be the best choice. As said in 2.1, in order to enhance the information provided to the user during her search, a heatmap featuring the rent price range has been overlapped on the map of Paris (prices considered as the cost of an apartment in terms of €/m^2 of surface):

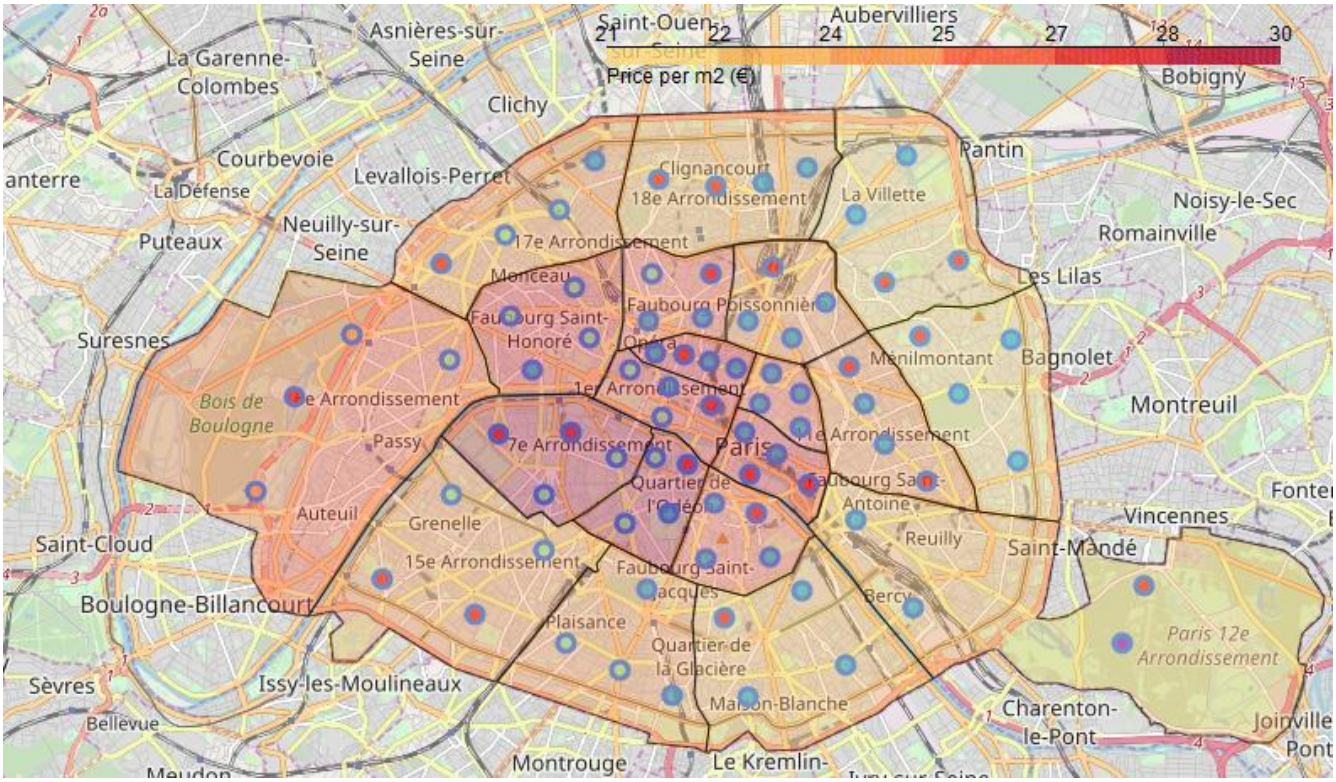

Figure 3 Cluster Map of Paris with arrondissements colored according to their price range (the price is referred to the cost an apartment in terms of €/m2).

At first sight, the map looks a bit overloaded, for this reasons it has been decided to change the *Folium* map layer with a clearer one in which it is possible to appreciate the colours i.e. the price variations more effectively:

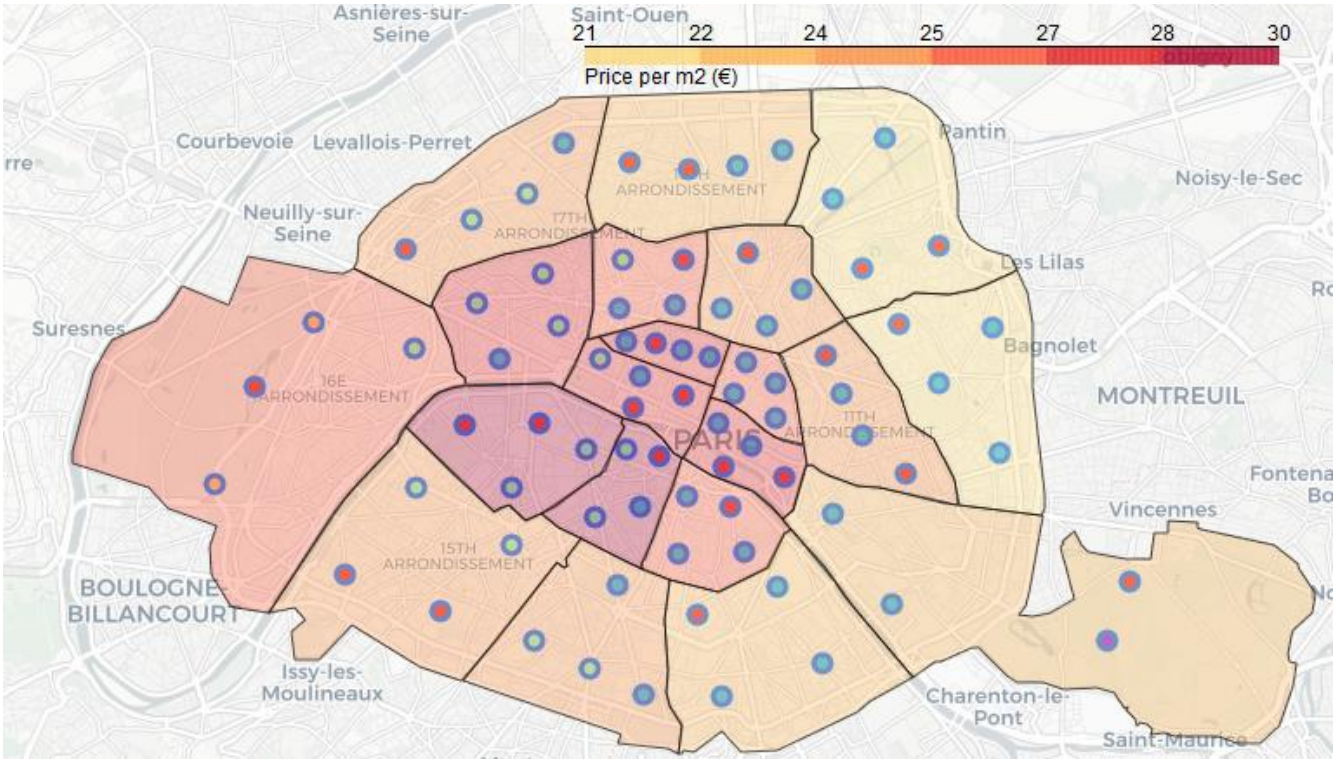

Figure 4 Cluster Map with prices using a different ('cartodbpositron' layer style) style.

Once the main map has been produced, it is now time to iterate through the cluster of preference and find the optimal neighbourhood. In order to do so, a random location around the center of Paris has been chosen, which can very likely be the an actual workplace, lot of companies having offices in the center of Paris. It has been assumed that the user is someone looking for places nearby to practice some sport and to go watch the *PSG* at the stadium on Sundays, therefore cluster 4 is the best pick for her, being characterized by the presence of tennis courts and the proximity to the sport stadium. Using TomTom API it is possible, among many other things, to find the time it takes to move through two positions *by car*, the user is therefore assumed to have a car or at least to do carpooling with a colleague/friend etc. All the cluster is therefore searched through in order to find the location with the shortest traveling time, which will be then selected as best pick. Once a location has been selected, it is also possible to retrieve from TomTom API output a set of coordinates defining the point that have to be passed through to reach the destination i.e. the most *probable route.* Simply plotting the coordinates will result in a graph like the one below:

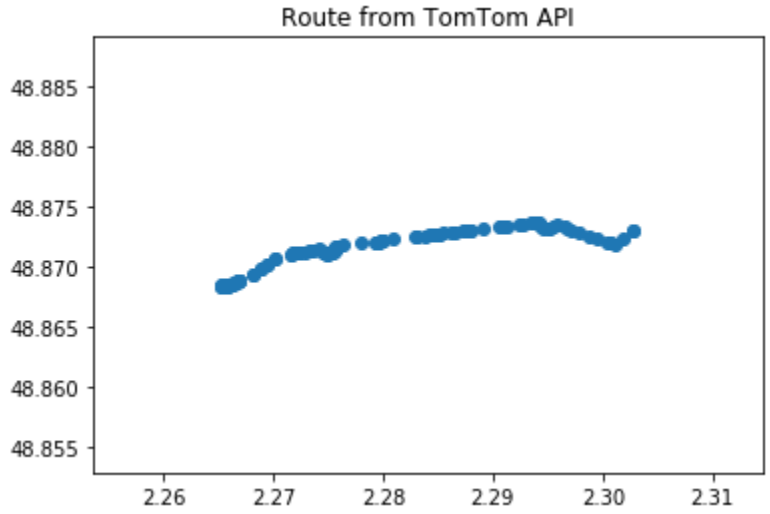

Figure 5 Route produced by TomTom API. Latitude and Longitude are used in the axis.

Finally, it is possible to visualize the route on the map by overlapping it to the previous map:

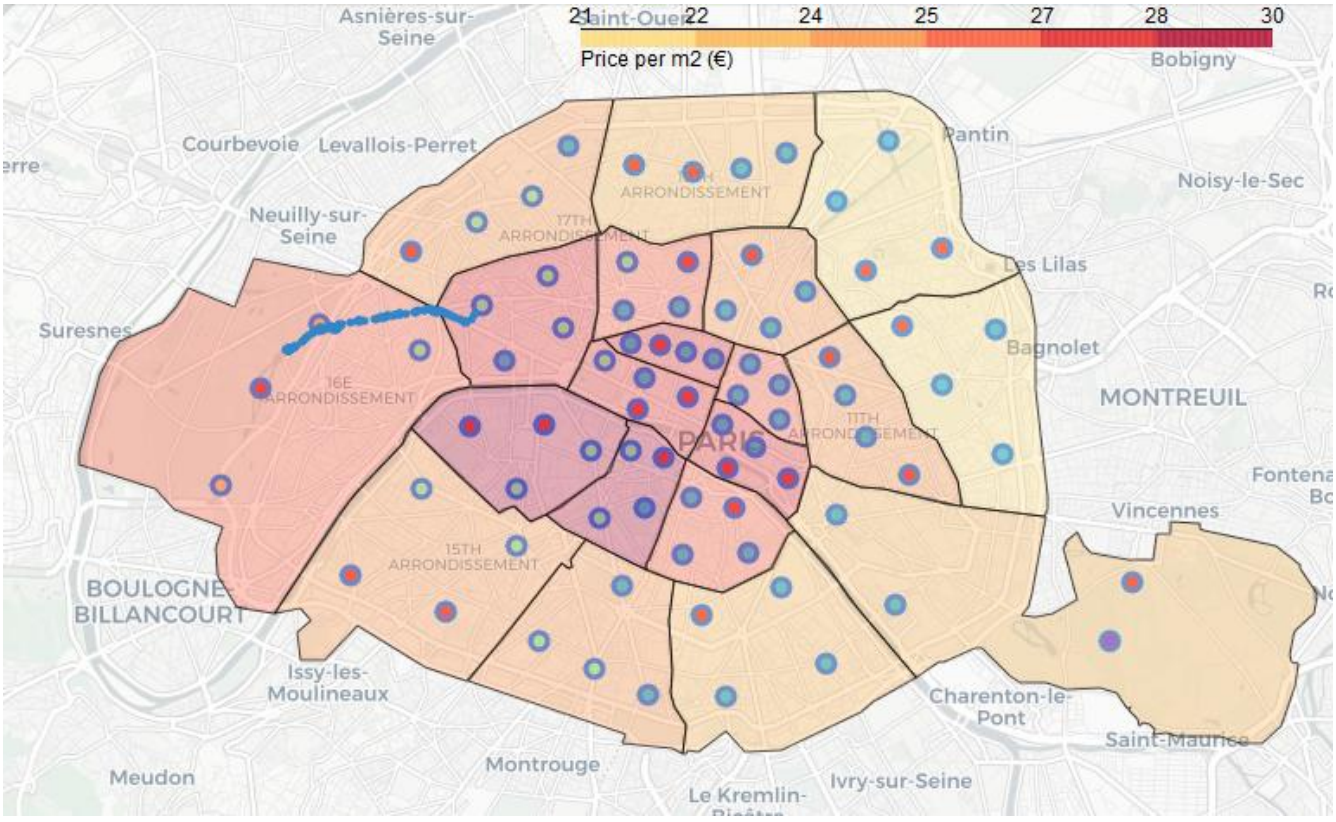

Figure 6 Cluster Map with the route just found higlighted in blue.

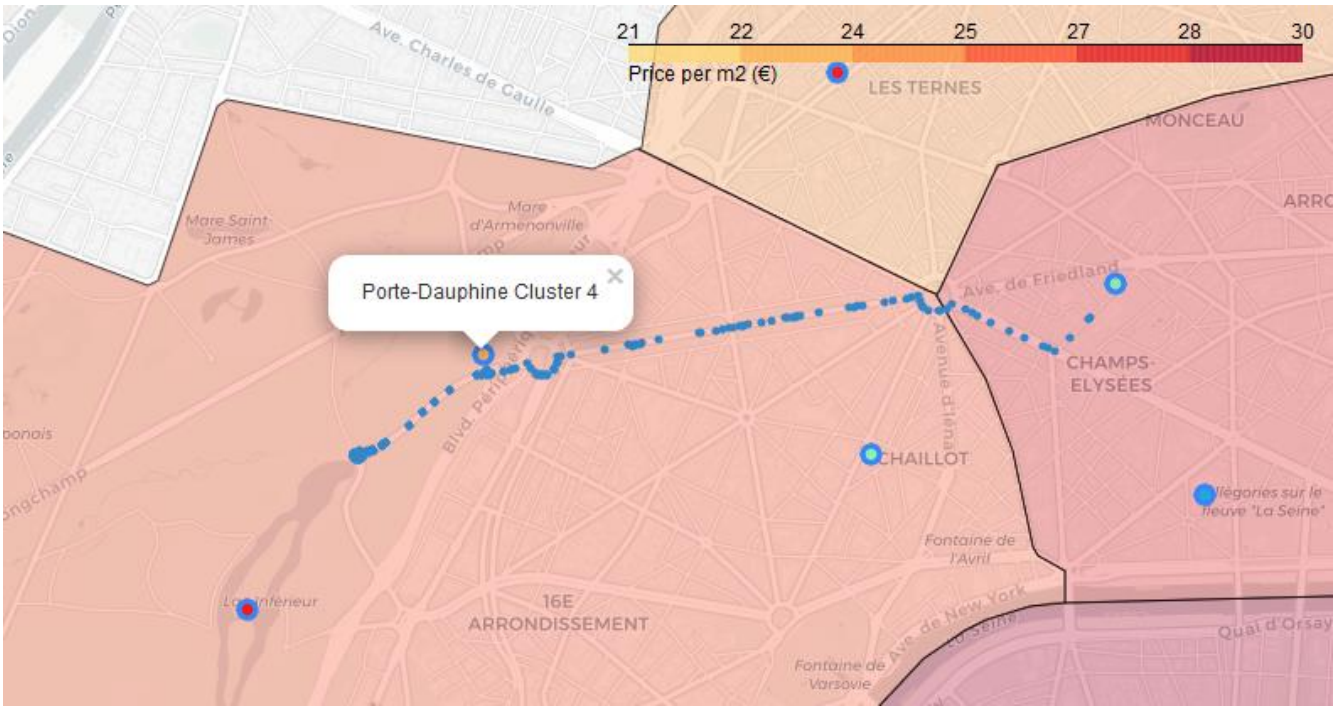

Figure 7 Zoom-in of the route and the 'winner' location.

6. Conclusions & Future directions

The research succeeded in providing a classification of the city of Paris. A potential user can define a search based on what she is looking for in terms of neighbourhood amenities and position of her workplace. As a result, she is provided with a zone that matches her criteria, with the average travel time to her workplace and a generic information about the price range in the area, in order to at least comparatively assess whether she is in a cheaper *arrondissement* or not, which in the end, is essential to finalise the investment of moving to the new house. This tool can be quite useful as *support* in a more in dept-research, since using the most common websites for looking for apartments, a user is only provided with the information about a give apartment: thanks to this project, the user can also have an initial understanding of the neighbourhood the home is located in.

The applicability of this project is higher in case of relatively large areas, where there is a high density of venues and it would be necessary to filter all those and/or find a common pattern that allows to classify a city's neighbourhoods according to such venues. Paris is a valid example of such, and I believe that many users in many big cities can benefit from the approach of this tool.

The project has many ways in which it can be developed, here a few examples:

- Provide a GUI to ease the use,
- Use machine learning algorithms to create a recommender system, in order to see the if somebody has already done a previous research similar to the user's and see if the results have been fruitful and the feedback of the choice,

- Use an API that allows to compute the traveling time not only by car, but also by other means, like public transports,
- Use multiple sources to derive the venues and create a comparison (Foursquare may be used less in Europe than in the US, for instance).