Capstone Project - The Battle of the Neighborhoods

Table of content

Introduction: Business Problem

The goal of this project is to help anybody who doesn't know Paris and would like to buy a house or an apartment with somehow the best possible value for money. Indeed, in Paris, like any big city, you have many very different neighborhoods; each of them has its advantages and disadvantages so it is hard to find the best place to live especially if you don't know already the city.

In this project, I will consider that a good place to live, is a place located near at least one subway station and as many venues as possible because I want a dynamic place with a lot of activities and transactions around it and all this for the lowest possible price.

Data

To fulfill the requirements of this project, I will need several data sources:

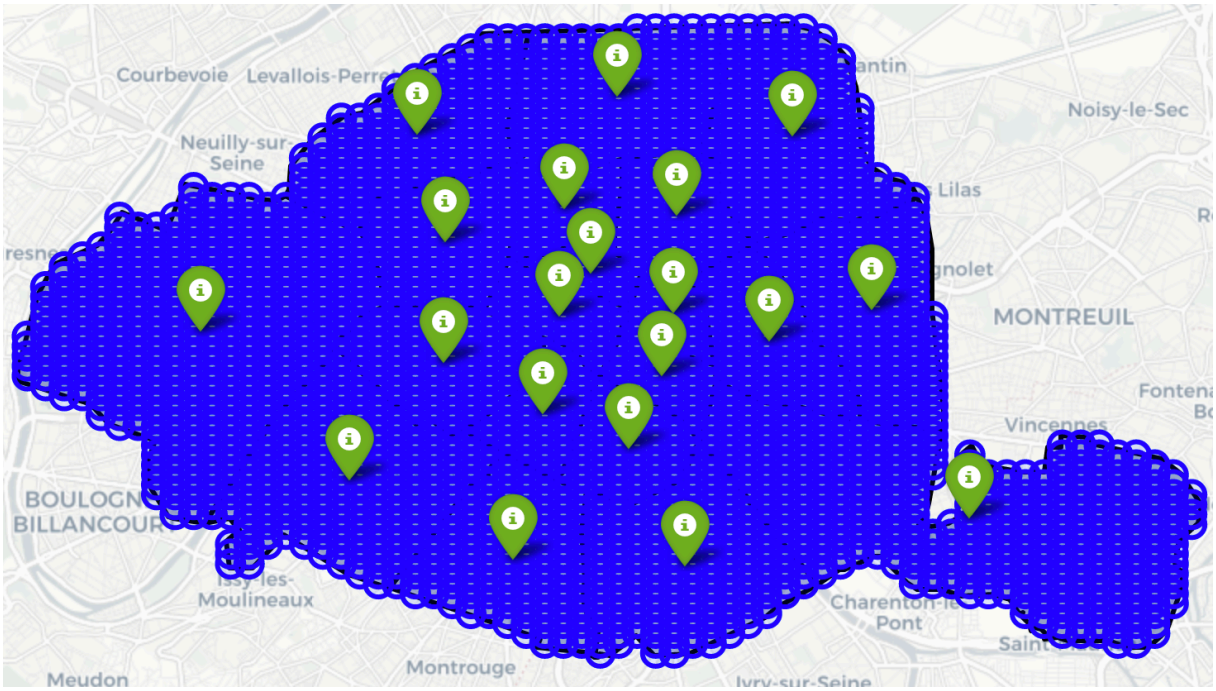
- https://opendata.paris.fr to get the shape of Paris and its boroughs, it will be used to filter and better locate the potential locations.
- http://dataratp.download.opendatasoft.com to retrieve the positions of all existing subway stations, it will be used to keep only the locations close to one of them.
- https://cadastre.data.gouv.fr to have all the transactions of houses or apartments in Paris since 2014, it will be used to get the average price by square meter and to keep potential locations where there are enough transactions.
- https://api.foursquare.com to get all the venues close to each potential location.
- https://nominatim.openstreetmap.org to retrieve the addresses corresponding to the final potential locations.

Methodology

The first thing to consider is what do we mean by “close to”? In this project, I considered that something is close to a given location if it is located within 200 meters.

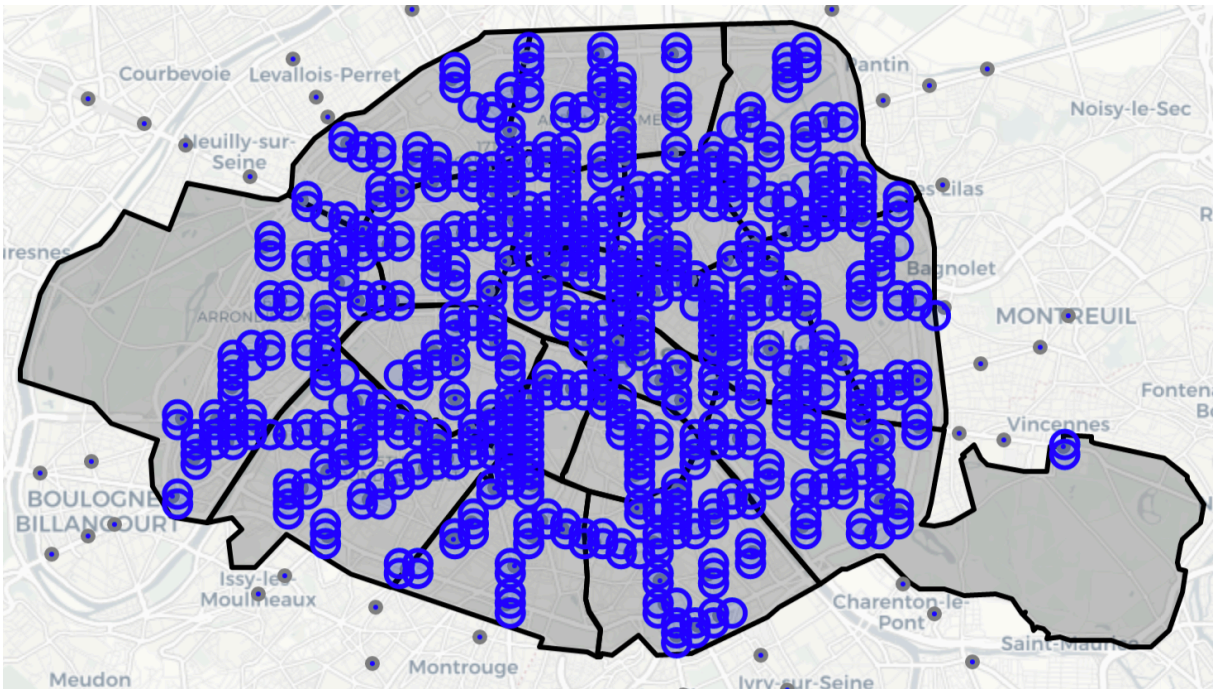
Once I retrieved the shape of Paris and its boroughs, I had to find the minimum and maximum latitude and longitude in order to define the rectangle that contains entirely Paris. Then, I covered the entire city with small circles of 200 meters; each circle is represented by its middle and is considered as a potential location. At this stage I have 2487 potential locations.

The map of Paris covered by the 2487 initial potential locations



To reduce this number, I loaded all the existing subway stations and kept only the potential locations with at least one subway station nearby, thanks to this filter, I ended up with 681 remaining potential locations.

The map of Paris with the 681 potential locations close to a subway station

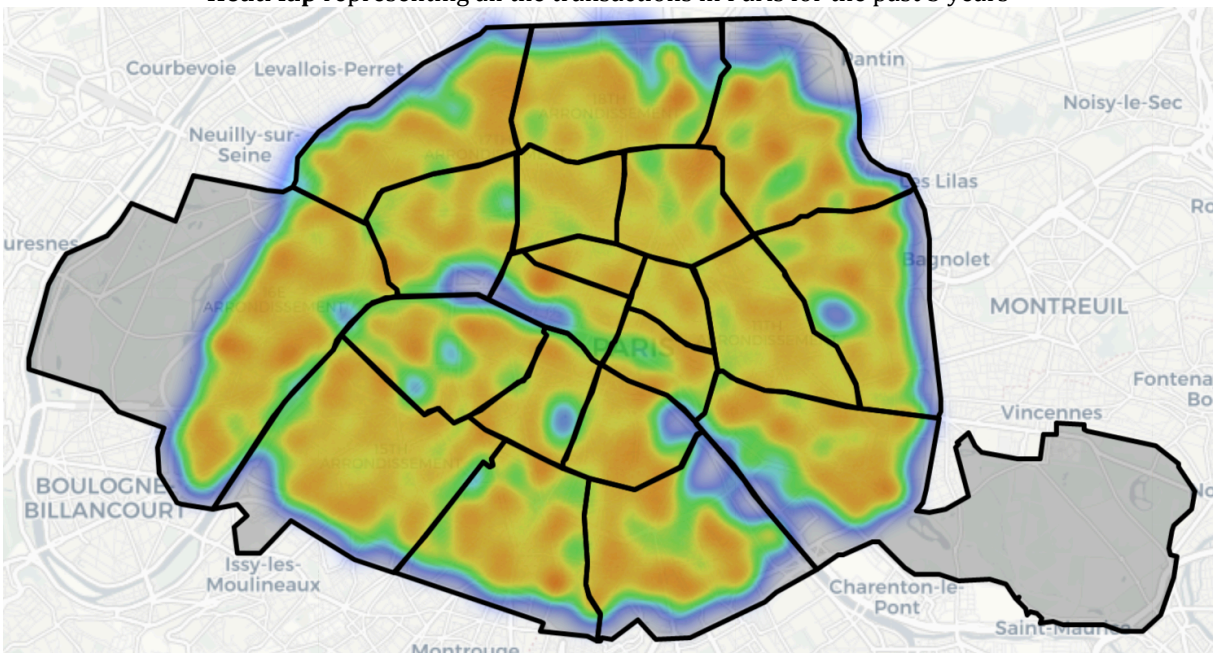


At this stage, we can still have some potential locations that don't make sense, as they could be located far from the houses.

To fix this issue, I loaded all the transactions in Paris for the past 5 years. I assumed that if no transaction has occurred near a potential location, it would mean that there is no house to buy/sell, so it could not be a good candidate.

In the transactions that I loaded, I decided to keep only the transactions with price, total amount of square meters and average price by square meters between 1 % and 99 % in order to get rid of suspicious/incorrect/invalid transactions. Indeed, I had some transactions with a very low or very high price or amount of square meters, those transactions are not representative and could be erroneous, so I had better to remove them to avoid affecting the final results.

HeatMap representing all the transactions in Paris for the past 5 years

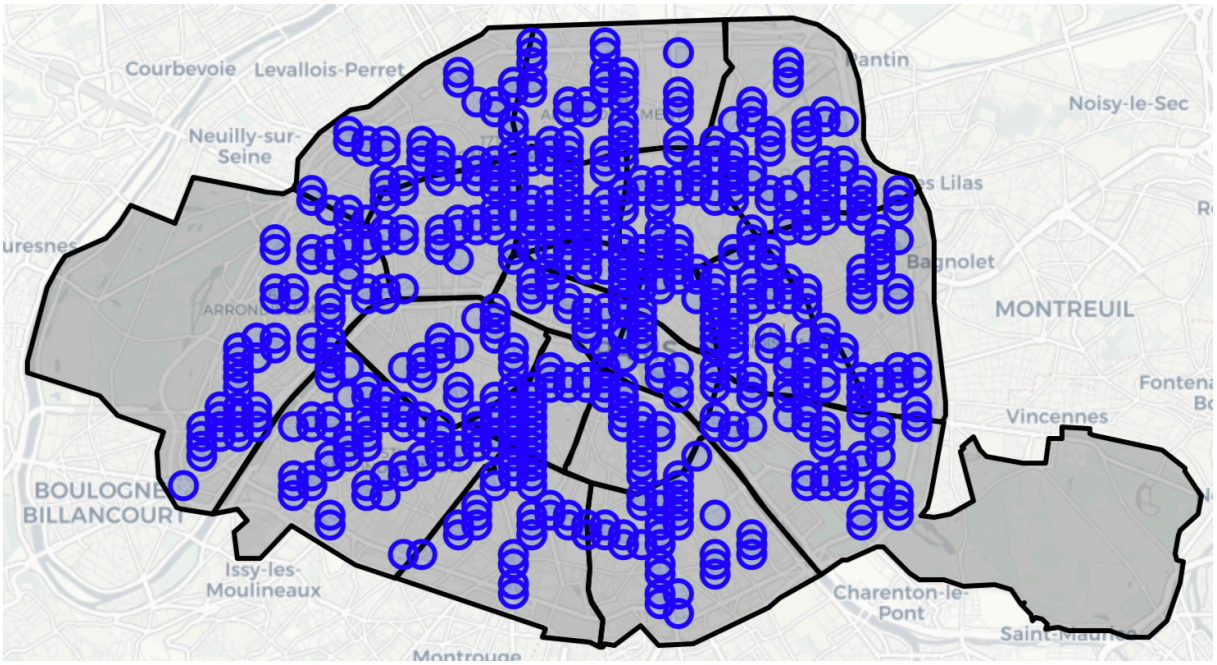


As you can see above, there are several holes where there is no transaction, they actually correspond to parks, rivers and other things that prevent having houses nearby, as they are not good candidates, I removed them from the candidate list.

I also decided to keep only the locations that have at least 10 transactions, which means that I expect more than one transaction every 6 months (2 by year in 5 years). I consider that below this threshold, it would be too hard to find a house or an apartment anyway, so no need to keep them. Once removed, I ended up with 613 potential locations.

As I reduced a lot the total amount of candidates, I could use Foursquare to retrieve all the venues close to each remaining potential location and then remove all the potential locations that don't have at least one venue since I expect a dynamic place with a lot of activities around it. Once removed, I ended up with 610 potential locations.

The map of Paris with the 610 remaining potential locations with TXN and venues

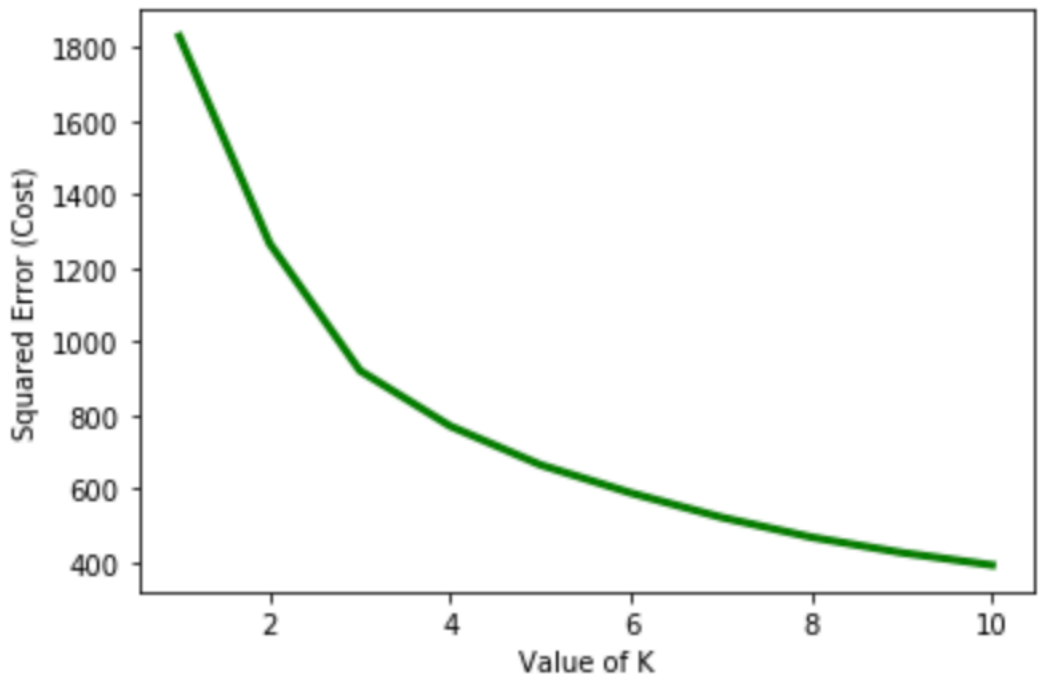


At this stage, all the remaining potential locations are real candidates so now it is time to cluster them based on the total amount of transactions, average price and the total amount of venues.

To cluster the locations, I used K-Means Clustering's approach to identify all potential locations that are similar.

The first thing to do when we want to apply the K-Means Clustering's approach, is to find the best value for K and for this, I used the elbow's approach by comparing the squared error with a value of K going from 1 to 10.

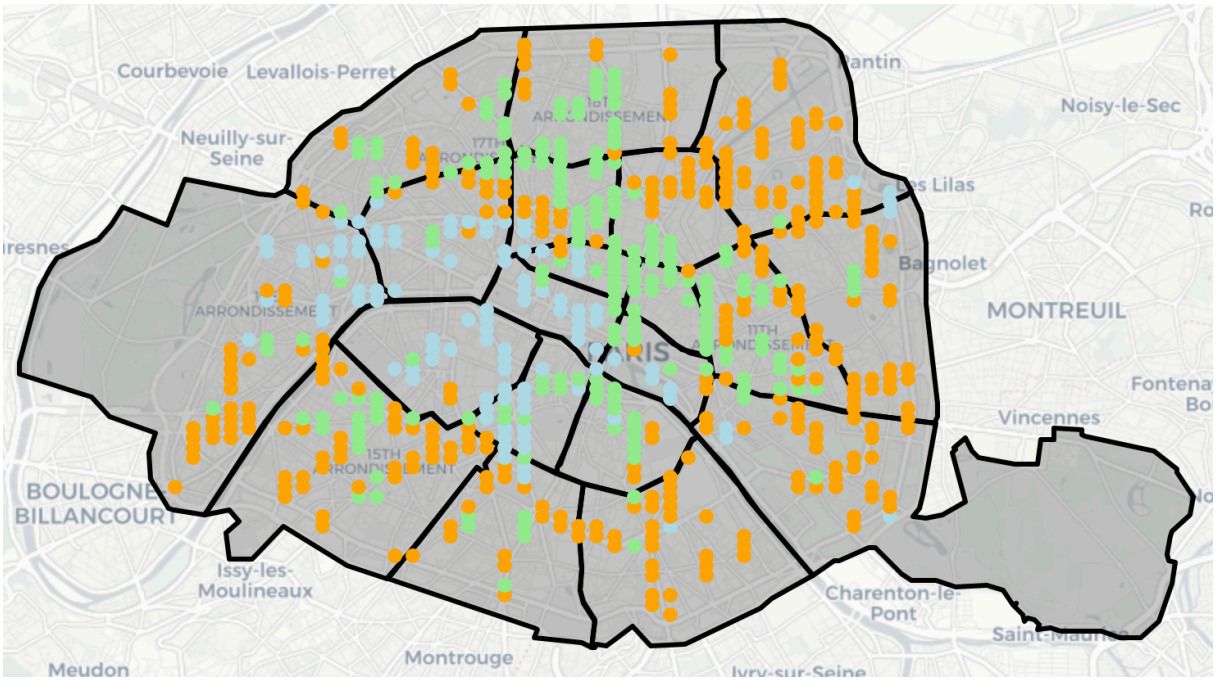
The evolution of the squared error with K between 1 and 10



According to the line chart above, the elbow point seems to be located at K = 3.

So I applied the K-Means Clustering's approach with an expected amount of cluster set to 3 and had the next distribution:

The distribution of the remaining potential locations in 3 clusters



The averages for each cluster are the following:

Label	Transactions	Price	Venues
0	53.290221	9633.529968	11.851735
1	112.417112	10815.064171	27.609626
2	41.867925	14942.330189	22.632075

According to the result above, the best possible value for money (it may depend on the stakeholder) seems to be the cluster with the label 1 as it is where there is the biggest amount of transactions which means that it should be easy to buy or sell a house or an apartment there, and the biggest amount of venues nearby indicating that there is a lot of activities, and all this for a middle average price.

The distribution of the remaining potential locations of the selected cluster

Once I have selected the cluster that fulfill the best to my requirements, I can keep only the potential locations of this particular cluster, then I sort the remaining locations by average price and total amount of venues and finally keep the first ten results.

The final result is the following:

Transactions	Price	Venues	Address
120	8060	14	18, Rue de la Chine, Père-Lachaise, 20e
102	8275	18	La Girafe et la Lune, Rue des Gâtines, Père-Lachaise, 20e
170	8285	13	Versigny Immobilier, Rue Versigny, Château Rouge, Clignancourt, 18e
167	8295	10	Le Pain en Fêtes, Rue Marcadet, Château Rouge, Clignancourt, 18e
131	8361	10	Théâtre de l'Alambic Comédie, Passage Kracher,

			Château Rouge, Clignancourt, 18e
156	8537	21	86, Rue Myrha, Goutte-d'Or, 18e
160	8585	12	109, Rue de Clignancourt, Château Rouge, Clignancourt, 18e
92	8605	28	100, Rue de la Croix Nivert, Grenelle, 15e
166	8668	23	8, Rue Custine, Château Rouge, Clignancourt, 18e
108	8691	24	Gambetta - sortie 1 - Mairie du 20ième (Hôpital), Rue Belgrand, Père-Lachaise, 20e

The distribution of the top 10 of the potential locations

Results and Discussion

In the final result, over the 10 proposals, there are 6 locations in the 18th borough and 3 locations in the 20th borough, which means that 90% of the proposals are in two boroughs. This result is actually not really surprising since they are both well-known boroughs for all the activities that you can find there for a relatively cheap price.

Even if those locations are indeed nice place to live, there is still something that has not been considered in this project which is the crime rate due to lack of data (I could not find data by borough, only for the entire city). Indeed the north/northeast part of Paris has the highest crime rate (more details about the crime rate in this report) which is probably the reason why I finally end up with those locations, if the crime rate could have been taken into account I would probably have had a different result.

Another thing to consider is the different choices that I made throughout this project that could be considered more or less objectives, those choices allowed to get those

particular locations but a stakeholder could have made different choices so we would have ended up with totally different results.

Conclusion

The purpose of this project was to find the best places to live in Paris close to at least one subway station and close to as many activities as possible for a minimum price. Thanks to the different choices I made, I ended up with 10 best places to live but I realized that the places I found where not necessarily safe places. Due to a lack of data, I could not add the total amount of crimes as a feature to consider when clustering the potential locations, I would have chosen the cluster with the biggest possible amount of transactions and venues and with the lowest possible price and amount of crimes, the result would then have been very different.

To conclude, I would say that it is important to always have all the required data and a clear idea of what the stakeholder wants to make sure that we get the best possible results.