Recommending a location for a business office in Paris

1. Introduction

In this project, we will try to find an optimal location for a new business office for a company in the city of **Paris**.

Paris is a very active city and many companies choose to locate their businesses here.
- Our first criterion will be the budget : we will be interested in looking at areas with **low to medium rent cost**.
- Second Criterion will be t**ransportation** : We want the location to be **easily accessible by the metro (subway)**, which is the most used mean of transportation in Paris
- Third and last, we will be looking at **venues** surrounding the area. More particularly, **restaurants**. Companies would want locations where their employees have diverse options for lunch break. Since a happy employee is more productive, this can benefit the company in the long term.

We will use data science to analyse the available data about Paris's neighborhoods and, based on the criteria we defined, we will recommend a list of promising neighborhoods.

2. Data

a. Data sources

Based on how we defined our problem, these are the factors that will influence our recommandation :
- Price of rent for each neighborhood
- Number of metro stations of each neighborhood
- Number of restaurants in the neighborhood

These are the data sources that we will need to extract/generate the required information:
- Paris neighborhood data : latitude and longitude of each neighborhood's center, neighborhood name, corresponding arrondissement (or borough) ... : **opendata.paris.fr**
- Parisian metro stations geo data : latitude and longitude of each station : **dataratp.opendatasoft.com**
- Rent cost per neighborhood in Paris data : **opendata.paris.fr**

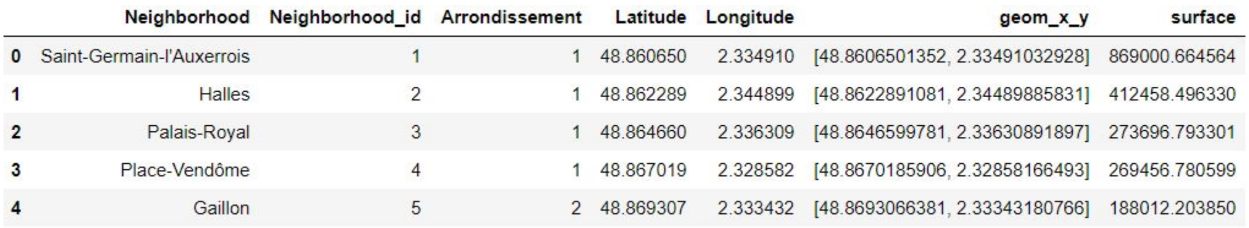
- Number of restaurants and their location in every neighborhood will be obtained using **Foursquare API**

b. Data acquisition and cleaning

i. Neighborhoods locations data

In this section, we want to extract geolocation data for each one of Paris’ 80 neighborhoods. This data is available on opendata.paris.fr in JSON format.

Data was downloaded and put in a pandas dataframe format. We then proceeded to drop the features that won’t be needed. This is the data we will be working with :



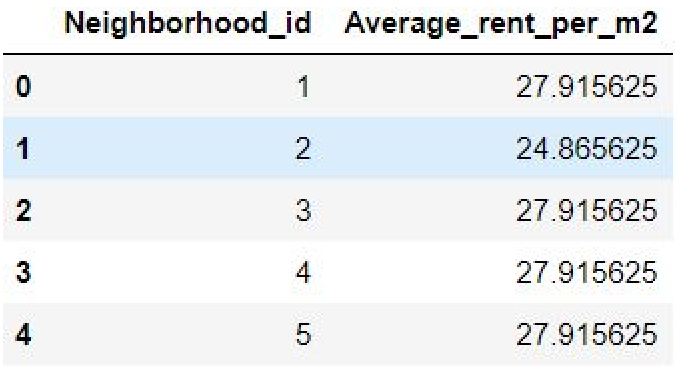
	Neighborhood	Neighborhood_id	Arrondissement	Latitude	Longitude	geom_x_y	surface
0	Saint-Germain-l'Auxerrois	1	1	48.860650	2.334910	[48.8606501352, 2.33491032928]	869000.664564
1	Halles	2	1	48.862289	2.344899	[48.8622891081, 2.34489885831]	412458.496330
2	Palais-Royal	3	1	48.864660	2.336309	[48.8646599781, 2.33630891897]	273696.793301
3	Place-Vendôme	4	1	48.867019	2.328582	[48.8670185906, 2.32858166493]	269456.780599
4	Gaillon	5	2	48.869307	2.333432	[48.8693066381, 2.33343180766]	188012.203850

ii. Rent data

The rent cost per neighborhood data was also available on opendata.paris.fr in CSV format. Data was downloaded and put into a pandas dataframe format.

Our data contained 32 rows per neighborhood. To different rent costs in different conditions (for example furnished vs not furnished). In order to simplify because we want to have 1 row per neighborhood, we grouped the data per neighborhood and kept the mean value of rent cost of all the 32 categories.

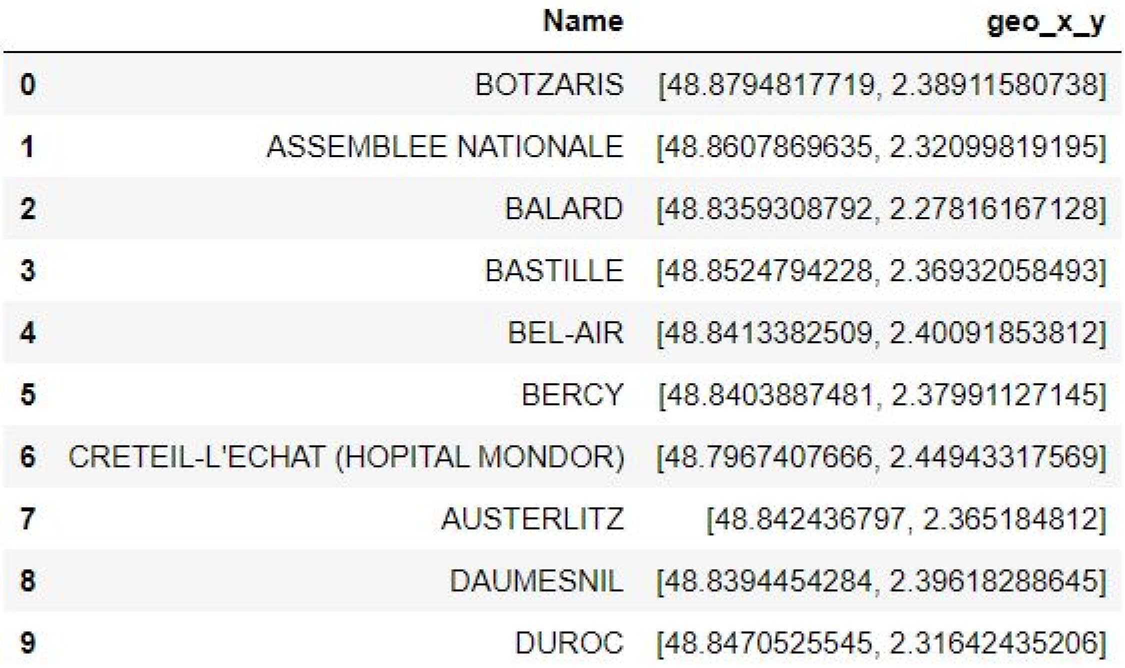
This is what our data looks like :



	Neighborhood_id	Average_rent_per_m2
0	1	27.915625
1	2	24.865625
2	3	27.915625
3	4	27.915625
4	5	27.915625

iii. Metro stations locations data

We managed to find the metro data we needed on **dataratp.opendatasoft.com** in JSON format. We proceeded to download it and only keep interesting features for our project.
This is what our data looks like after cleaning :

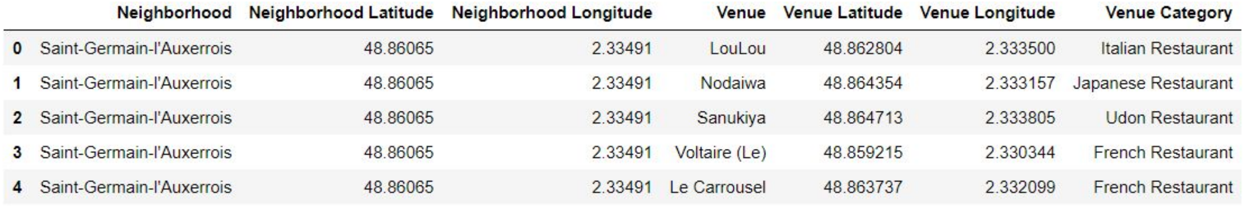


	Name	geo_x_y
0	BOTZARIS	[48.8794817719, 2.38911580738]
1	ASSEMBLEE NATIONALE	[48.8607869635, 2.32099819195]
2	BALARD	[48.8359308792, 2.27816167128]
3	BASTILLE	[48.8524794228, 2.36932058493]
4	BEL-AIR	[48.8413382509, 2.40091853812]
5	BERCY	[48.8403887481, 2.37991127145]
6	CRETEIL-L'ECHAT (HOPITAL MONDOR)	[48.7967407666, 2.44943317569]
7	AUSTERLITZ	[48.842436797, 2.365184812]
8	DAUMESNIL	[48.8394454284, 2.39618288645]
9	DUROC	[48.8470525545, 2.31642435206]

iv. Nearby restaurants data

In order to extract nearby restaurants data. We used Foursquare API's venues explore endpoint.
We extracted up to 100 venues in a radius of 500 meters around each neighborhood's center.

We then proceeded to only keep venues that fall into the restaurant category and put all this in a dataframe containing the relevant features.
This is what our restaurant data looks like :



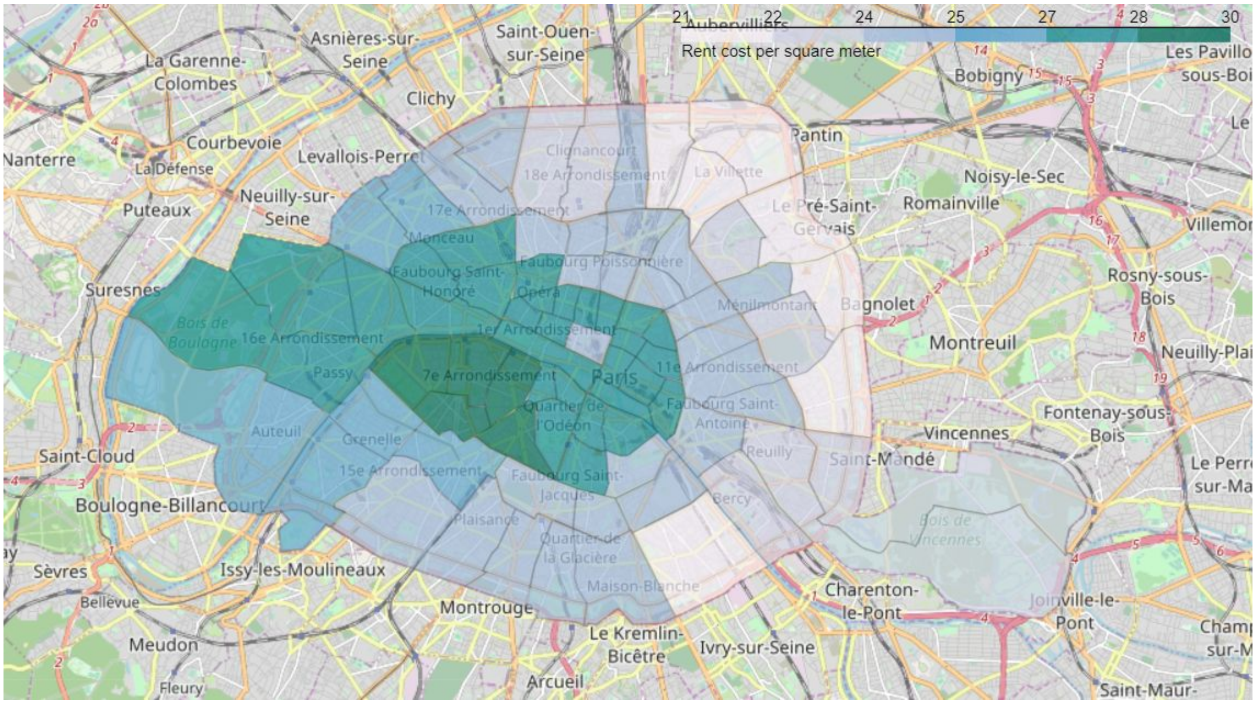
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Saint-Germain-l'Auxerrois	48.86065	2.33491	LouLou	48.862804	2.333500	Italian Restaurant
1	Saint-Germain-l'Auxerrois	48.86065	2.33491	Nodaiwa	48.864354	2.333157	Japanese Restaurant
2	Saint-Germain-l'Auxerrois	48.86065	2.33491	Sanukiya	48.864713	2.333805	Udon Restaurant
3	Saint-Germain-l'Auxerrois	48.86065	2.33491	Voltaire (Le)	48.859215	2.330344	French Restaurant
4	Saint-Germain-l'Auxerrois	48.86065	2.33491	Le Carrousel	48.863737	2.332099	French Restaurant

3. Methodology

In this project, we directed our efforts towards identifying **areas of Paris** that have : **low to medium rent cost**, a **high number of metro stations** nearby and **high restaurant density**.
Our methodology consisted of looking at our 3 criteria one by one and trying to get insights on promising areas.

c. Exploring rent cost data

We first looked at a choropleth map of Paris rent cost per neighborhood :

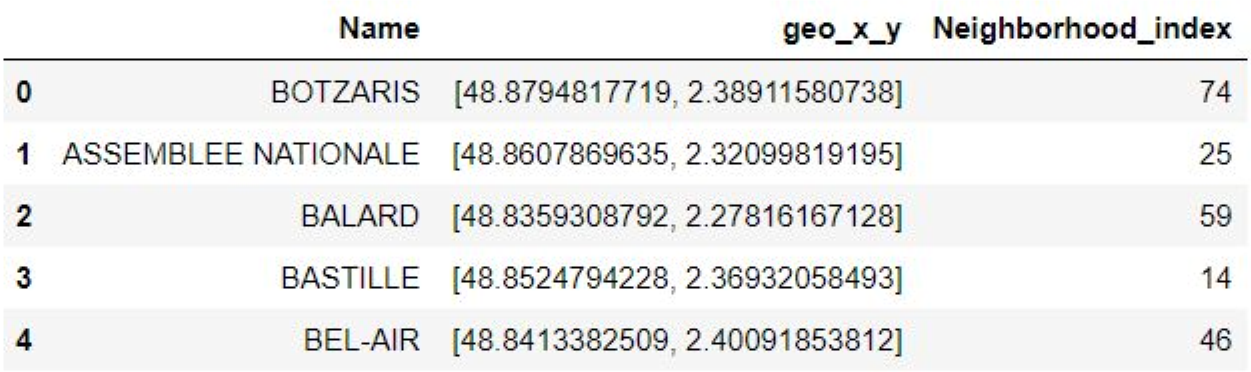


This showed us that the most interesting areas in the city are : south, south east, east, north east and north. These neighborhoods seem to have low to medium rent cost compared to the rest of the city which meets our requirements.

We then added the rent cost data to our main dataframe of Paris neighborhoods.

d. Exploring metro locations data

First step in exploring this data was to associate each station to one of the 80 neighborhoods. To achieve this, and for each station, we calculated the distance to all neighborhood centers and associated it to the closest one. This allowed us to add a column to our metro data showing the index of the closest neighborhood :

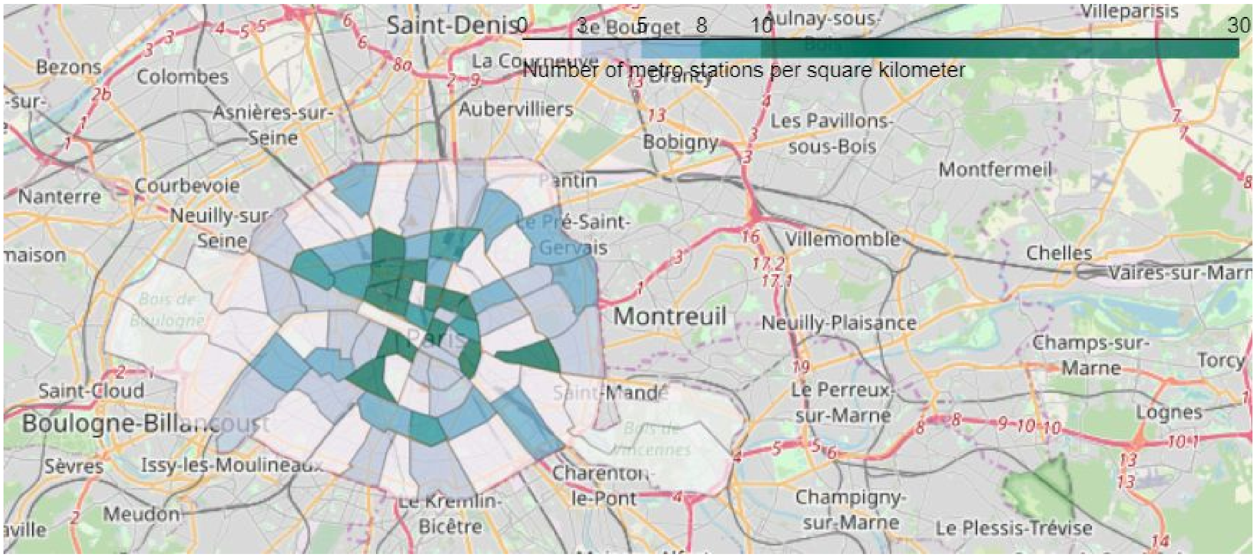


	Name	geo_x_y	Neighborhood_index
0	BOTZARIS	[48.8794817719, 2.38911580738]	74
1	ASSEMBLEE NATIONALE	[48.8607869635, 2.32099819195]	25
2	BALARD	[48.8359308792, 2.27816167128]	59
3	BASTILLE	[48.8524794228, 2.36932058493]	14
4	BEL-AIR	[48.8413382509, 2.40091853812]	46

We then proceeded to calculate the number of nearby stations for each neighborhood and added that to our main dataframe.

Dealing with a number of stations might not be relevant since neighborhoods are not the same size. That is why we proceeded to transform this feature into a density of metro stations (number per km²) :

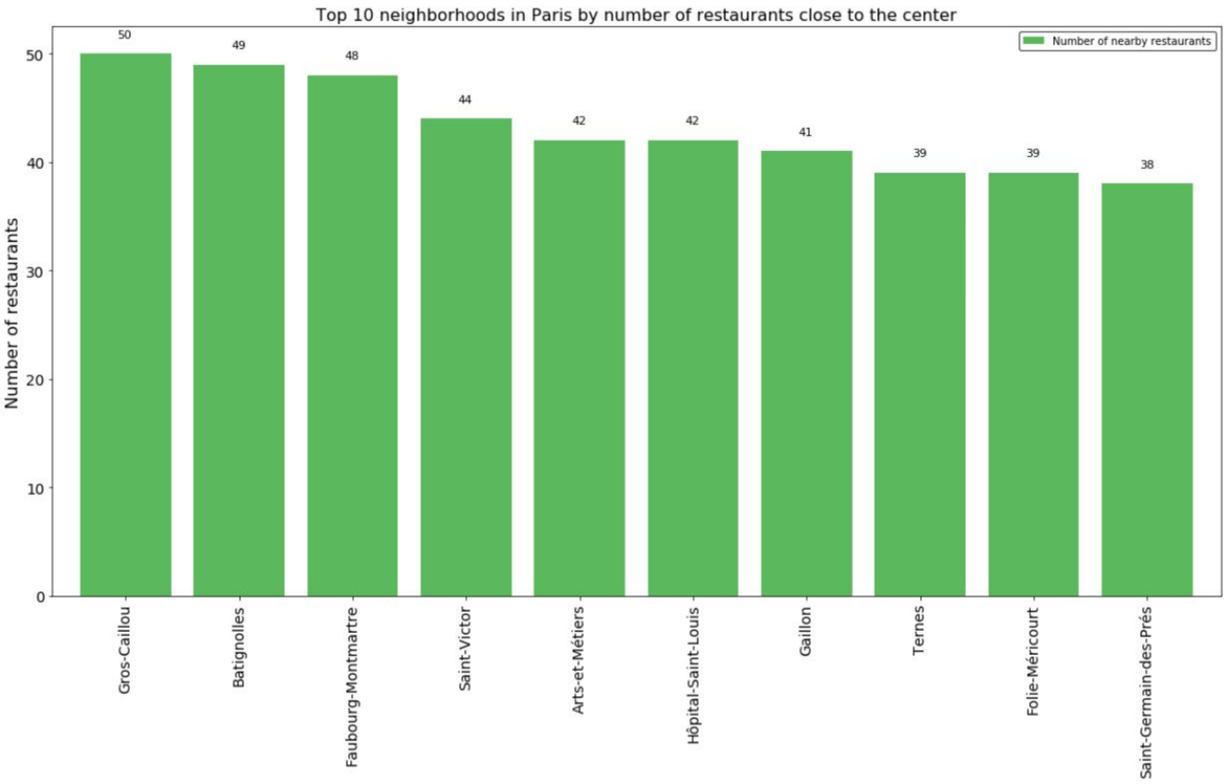
Now we can look at a choropleth map to see which areas will be interesting for our analysis :



We can see that there is no clear pattern on when neighborhoods with high metro station density are located but there seems to be many promising neighborhoods on the east side.

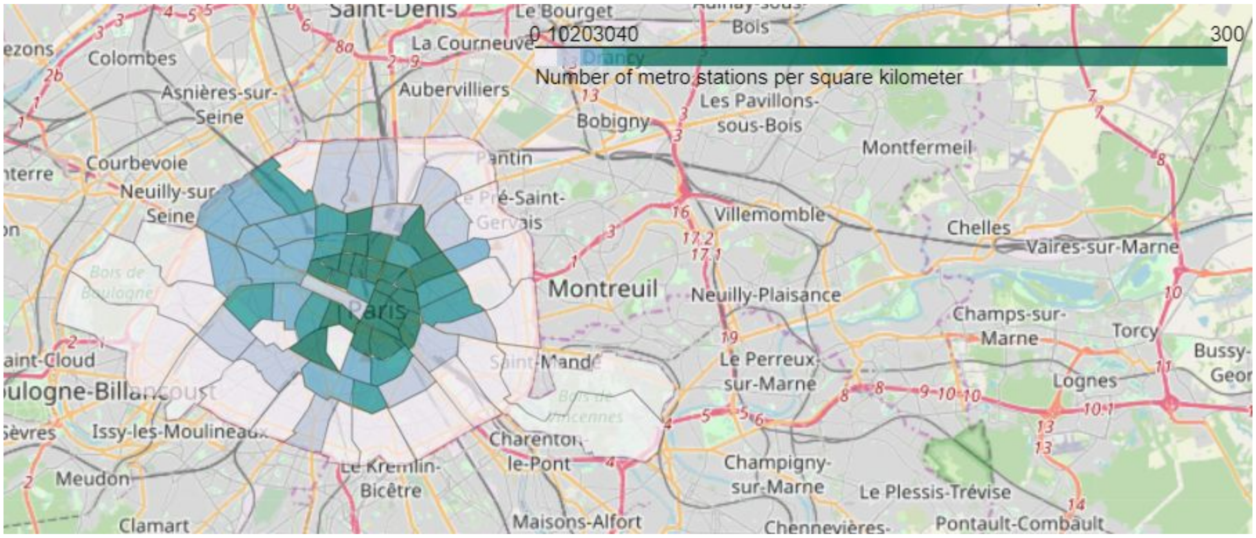
e. Exploring nearby restaurants data

Looking at our nearby restaurants data, we started by extracting the top 10 neighborhoods in Paris by number of restaurants :



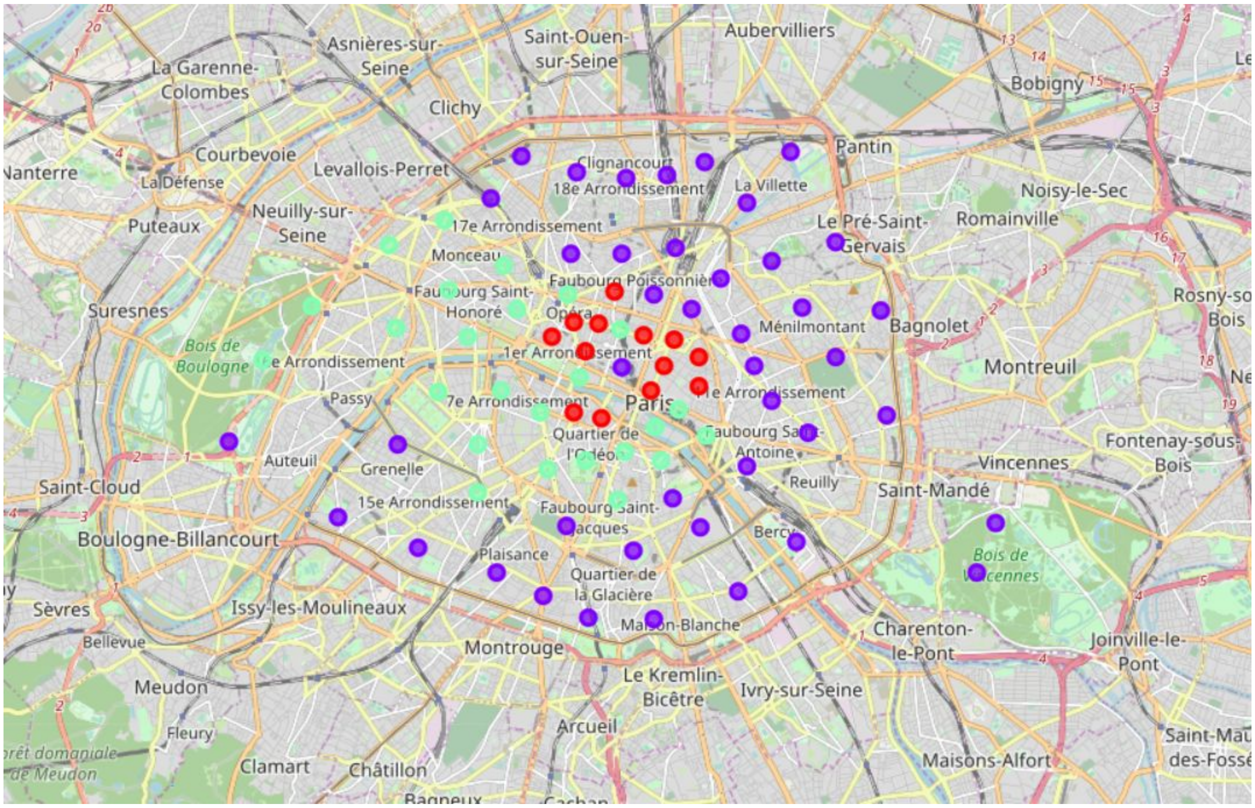
We then added this to our main dataframe. Following the same logic as with the number of metro stations, we transformed this into a density per km².

Now this is a choropleth map of nearby restaurant density per neighborhood :

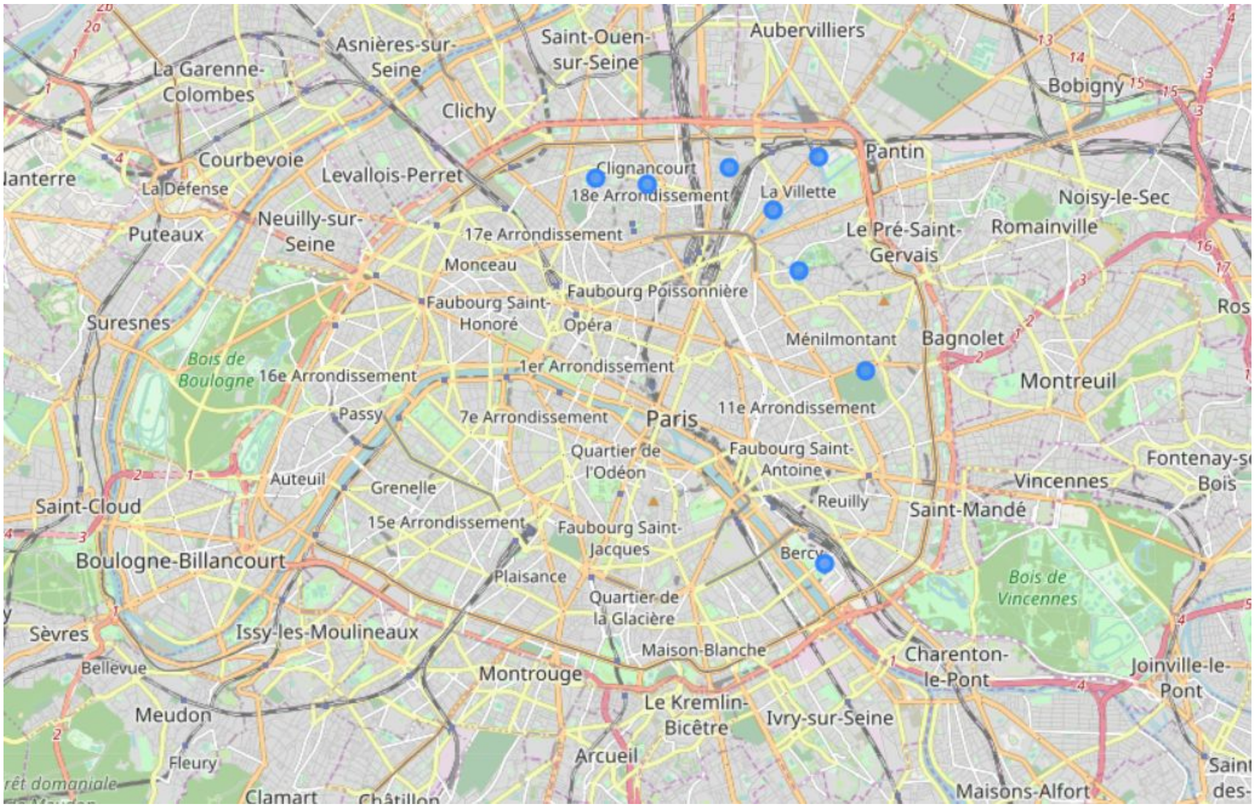


f. Clustering data

Now that we explored each dataframe and got insights about each one of our criteria and the promising areas. Let’s see if we can get more by clustering our data.
We used the K-mean clustering algorithm and defined 3 as our number of cluster, which gave us these 3 clusters :



The violet markers are the centers of our largest cluster. This seems to be the most promising cluster. We can already see on the map that these are the areas with low to medium rent and that many of these neighborhoods are easily accessible by the metro and have good restaurant density.
Yet, this cluster is quite larger and still needs refinement. Since rent cost is our first criterion, we decided to only keep neighborhoods with rent inferior to the mean value for the cluster. We also got rid of neighborhoods with very low restaurant density. These are our final results :



We obtained 7 neighborhoods situated on the north east and one south east.

4. Results

Our analysis of Paris neighborhoods was aimed at making a recommendation for a new business office location based on rent, accessibility via the metro and nearby restaurants. Exploring these criteria and analysing our data showed us that :

- Rent is higher in the center of Paris and towards its west side
- Metro stations density is very high in the center and relatively high in the east side of the city
- Paris has many areas of relatively high restaurant density, particularly north east,east and south east.

Clustering our data based on these 3 criteria confirmed our observations and gave us a cluster of neighborhoods situated more towards the east side (east, north east and south east) that contains neighborhoods that meet our requirements.

Now that we directed our attention to this cluster, we noticed that it was relatively big and that it needed some refinement. We removed neighborhoods that had high rent values and very low restaurant density values compared to the rest of the cluster.

Result of all this is 8 neighborhoods that meet the defined criteria and answer our business needs. All these neighborhoods are located north east of Paris.

5. Discussion

Although data analysis gave us satisfying results, these neighborhoods might not be optimal because of criteria that we did not define in this problem (like crime rate for example). But these results still represent a good starting point and a good first recommendation for a business office location.

6. Conclusion

Purpose of this project was to identify Paris areas that are suitable for opening a new business office in order to aid stakeholders in narrowing down the search for optimal location. By looking at rent cost in different neighborhoods, we started identifying our target areas. Then, by calculating metro stations density and restaurants density, we obtained promising results that confirm what we find out during data exploration. This allowed us to narrow down our search to certain areas. Clustering of Paris neighborhoods was then performed in order confirm our results and to create major zones of interest and locations to be used as starting points for final exploration by stakeholders.

Final decision on optimal location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.