Clustering the Paris's subway stations

Introduction

Background :

4.16 million people are using the subway of Paris every days. The system length 214 km (133mi) and count 700 trains. This is one of the biggest subway system in the world.

Purpose :

Paris is a very big city and can it get sometimes confused. It can be hard for the urbanist to understand in what purpose people are moving along the city. This project aims to classify the stations in cluster to try to understand better what is the "goal" of each neighborhood.
To identify the goal we are going to count next how many schools, bars, companies and college are close to each stations.

Interest :

Politics and business men can be interested in this classification.
It can be useful for example for a company trying to open a new restaurant. Understand what is the "purpose" of each neighborhood can help it to open the right restaurant in the right place.

Data acquisition and cleaning

Data sources :

The database of all the stations can be found here.

The position of all the places next to each stations have been found with the API foursquare.

Data cleaning :

First of all, with the database of the stations two columns were useless.

	ID	Name	Description	Coordinates	Unnamed: 4	Unnamed: 5
0	3677677.0	RUE DE LA FERME	ROND-POINT MARTIN LUTHER KING - 77258	48.8359484574	2.629903	NaN
1	3677688.0	COLLEGE LE LUZARD	COURS DES ROCHES - 77337	48.8467902961	2.613691	NaN
2	3678812.0	ZONE TECHNIQUE	ROUTE DES ANNIVERSAIRES - 95527	49.0119496186	2.533291	NaN
3	3678816.0	ENTRETIEN NORD	ROUTE DE L'ARPENTEUR - 95527	49.0114652199	2.515530	NaN
4	3682876.0	AVRON	63 BOULEVARD ALSACE-LORRAINE - 94058	48.8500655011	2.499395	NaN
5	3682892.0	PLACE DE LA RESISTANCE	FACE 52 AVENUE DU GENERAL DE GAULLE - 93050	48.8568322265	2.530116	NaN
6	3682907.0	AVENUE DES MARTYRS	49 AVENUE DU MARECHAL FOCH - 77108	48.8706786201	2.576489	NaN
7	3682970.0	CHELLES - GOURNAY RER	PISTE GARE ROUTIERE - 77108	48.8744671511	2.582937	NaN
8	3682980.0	DISPENSAIRE	FACE 39 RUE DE LA LIBERTE - 77108	48.8767496173	2.599939	NaN
9	3682998.0	NEUILLY-PLAISANCE RER	BOULEVARD GALLIENI - 93049	48.8528305972	2.514167	NaN

Figure 1: data before cleaning

We don’t need the columns “description” of the station because it’s its adress and we already have the coordinates. The columns ID is alos not usefull. Moreover we noticed that the columns “Unamed 5” only contains NaN. I decided to delete the three of them.

I renamed the columns “Coordinates” in “x” and the columns Unamed : 4 in “y”.

To finished I had to change the type of the columns “x” from object to float64.

```
df_positionGeoSta.dtypes
```

```
ID       float64
Name      object
x         object
y        float64
dtype: object
```

After that I have deleted all the station that had the same name.
After all of this step I had to reduce the size of the dataframe because I contained more than 2000 stations (bus and tram included). I choose to take the 125 last stations because it was the most famous ones (It would have of course work no matter the stations chosen).

The last step was to change the index to put it back as 1…125.

The dataframe of the stations is now clean.

	Name	x	y
0	Saint-Lazare	48.875682	2.325592
1	Stalingrad	48.884153	2.367357
2	Télégraphe	48.875522	2.398677
3	Trocadéro	48.862667	2.287283
4	Simplon	48.894123	2.347245

Figure 2: clean dataFrame

After this dataFrame I had to put in a dataframe the number of each different place each stations had close to them.

Concerning this dataframe I just change the type of the number in int and delete the duplicated but no need to delete any columns because I was just interressted in the number of lines of the dataframe.

Data completion

After having done the first data cleaning for the station I had to find the number of each places next to each station.

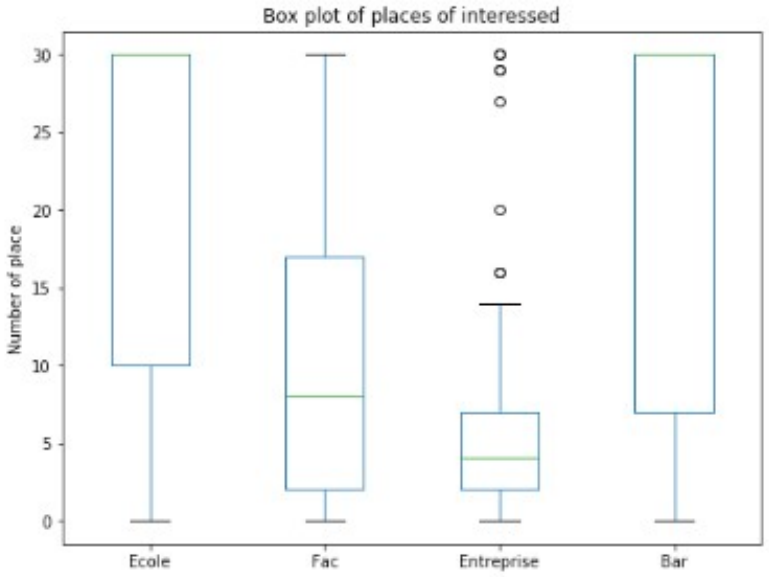
To do so I called the foursquare API for each places next to each coordinates of each stations. The place are considerate to close to a station if it is in a radius of one km. I choose this radius because we can think that someone can use a station and walk one km around.

I just had to count the number of lines of the dataframe I have got form the Foursquare API. The request for exemple to get the names of all the school in a radius of one km from the coordinates x and y and then to get the number of lines of the dataframe.

The next step is to join the two dataframe together to have the final dataframe.

	Name	x	y	Ecole	Fac	Entreprise	Bar
14	Pereire	48.885031	2.297135	30	16	8	30
15	Franklin-Roosevelt	48.868813	2.309926	30	17	20	30
16	La Hacquinière	48.895061	2.152664	3	0	1	0
17	Cergy-Saint-Christophe	49.049765	2.034345	5	0	1	0
18	Cergy-Préfecture	49.035944	2.080144	8	1	1	6

Exploratory Data Analysis



I plot the average box plot to understand more how the general places are.

I’m lacking data due to the limits of the Foursquare API.

Clustering

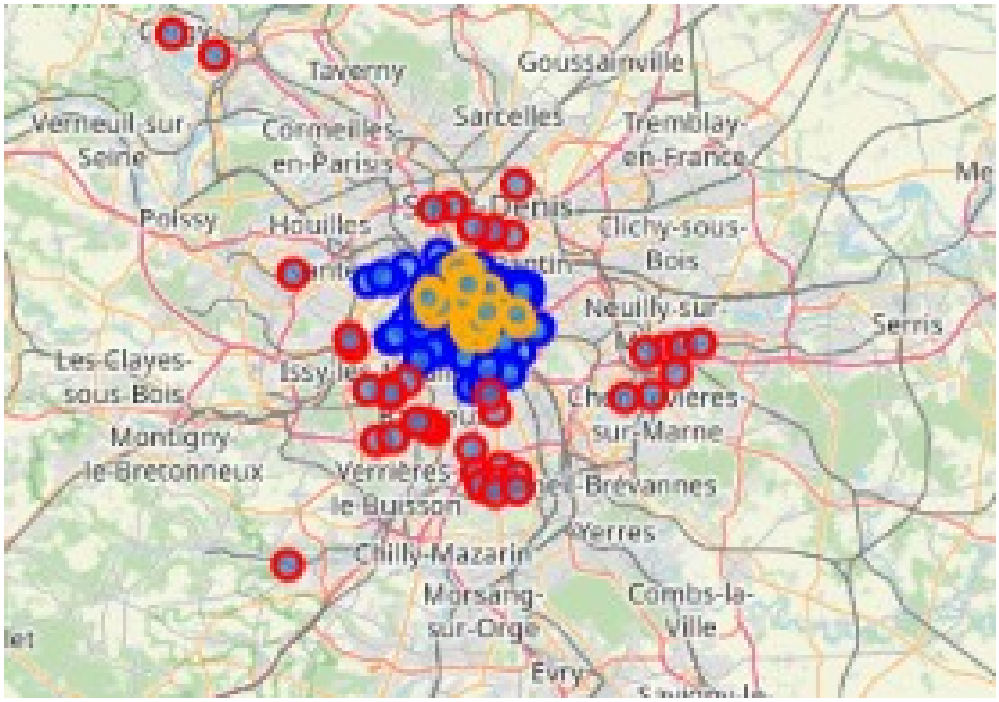
After that we want to cluster all the station thanks to the k nearest neighbor algorithm.

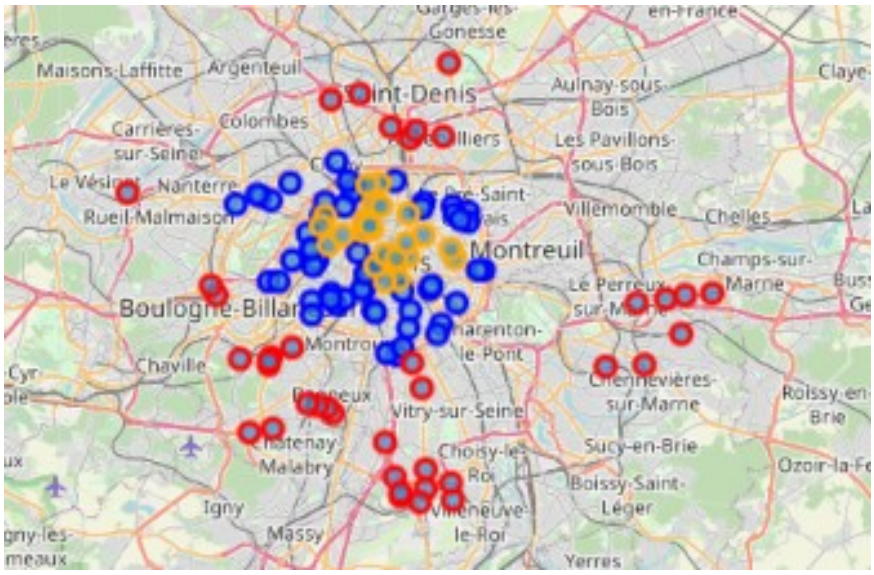
Once we have all the stations clustered we complete our dataframe with the clustered columns.

	Name	x	y	Ecole	Fac	Entreprise	Bar	Labels
0	Saint-Lazare	48.875682	2.325592	30	21	27	30	2
1	Stalingrad	48.884153	2.367357	30	14	5	30	0
2	Télégraphe	48.875522	2.398677	30	6	4	29	0
3	Trocadéro	48.862667	2.287263	30	10	5	30	0
4	Simplon	48.894123	2.347245	30	12	6	30	0

Now we have the “Labels” columns wich correspond to the cluster of each station belong.

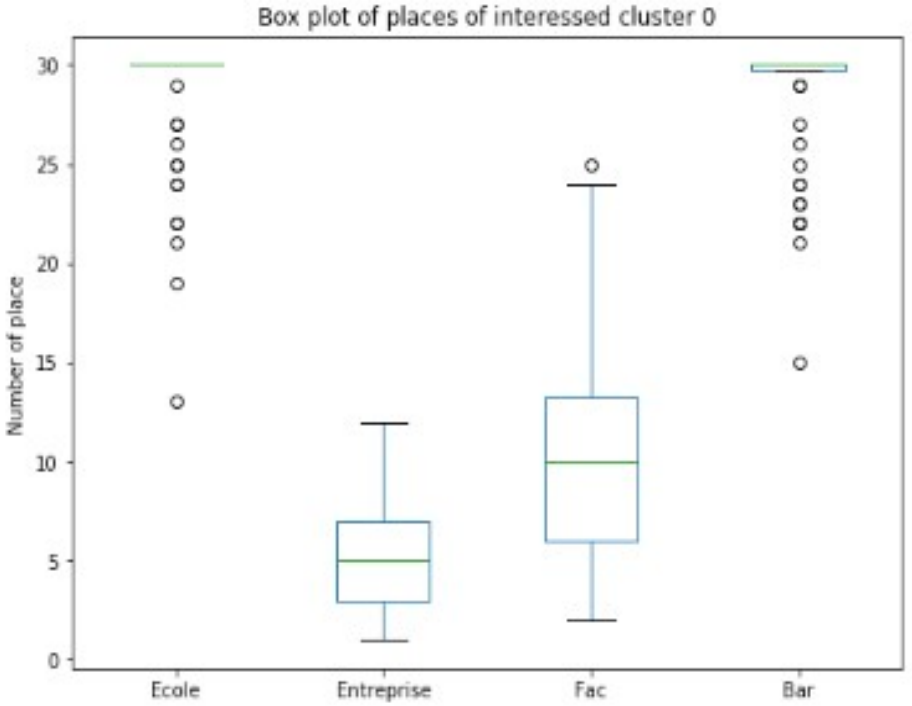
Map

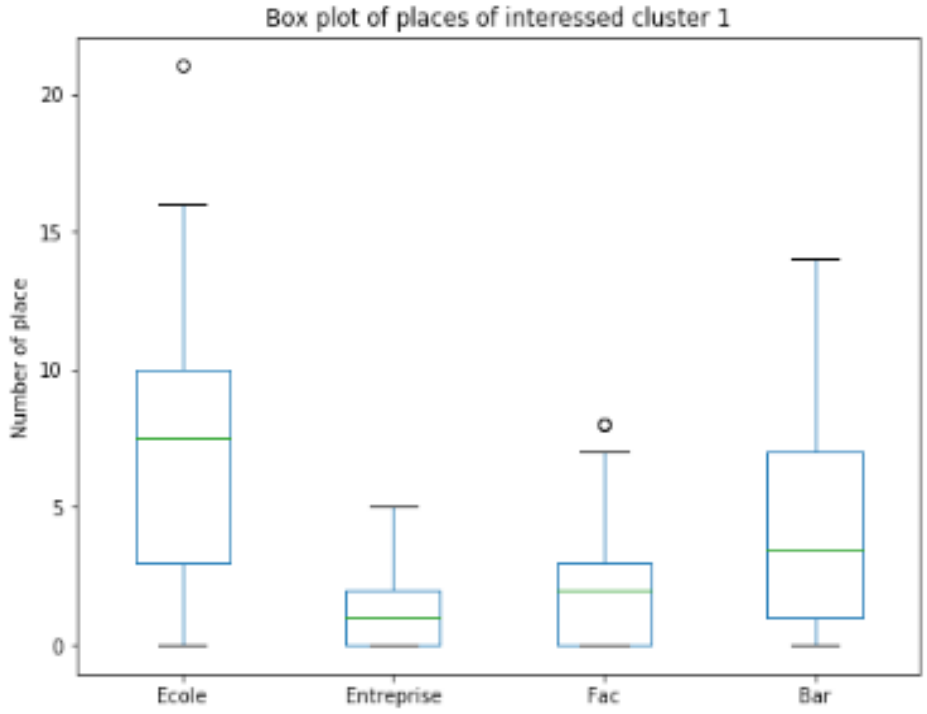


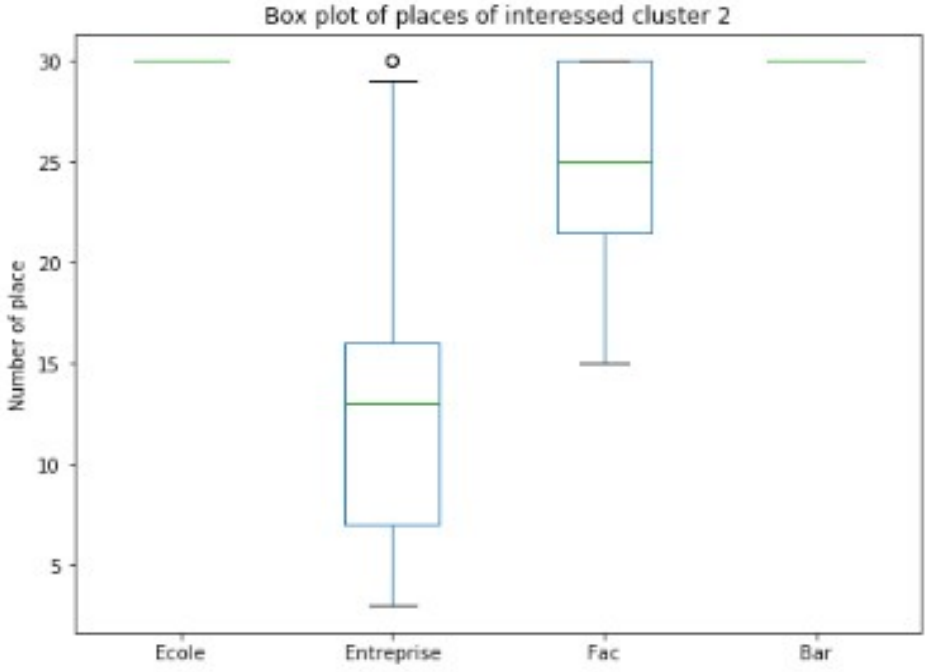


We can notice that the cluster are geographically located in “cercle”

To understand them more we are going to plot the number of each places the have each cluster.




Box plot of places of interessed cluster 1
Number of place
0
5
10
15
20
Ecole
Entreprise
Fac
Bar


Box plot of places of interessed cluster 2
Number of place
5
10
15
20
25
30
Ecole
Entreprise
Fac
Bar

We can notice that the first cluster is mostly composed off stations that are used for work and college education, the second cluster for education of children and bars. The last cluster outside of Paris is mostly composed of companies and college.

Conclusion :

The model can help us to determine the three clusters according to how people use the station.
The big problem of this study is that I have been limited by the number of call that I can make in Foursquare. The model is however ready to work with more stations.