The Biggest cities in the world

Business understanding

Everyday, businesses are looking for new opportunities to expand there businesses, to understand their markets.

This study is about defining the different cities to analyze them potential opportunities and to understand in a deeper way how each cities have a shape of ecosystems and its own proportions. The purpose at the end will be to cluster the biggest cities thank to the information about the different venues we have in each and to be able to take decision about it or to dig the analysis of a particular cluster.

Analytic Approach

The analytic approach will consist first in gathering a solid database of the biggest cities and the proportions of the venues in it.

The steps to gather the datas:

1. Wikipedia for the list of the biggest cities
2. Find their localization (with geocoder for example)
3. Get the proportions of the venues in each cities thank to the foursquare API

Then come the analysis in 3 steps:

1. Find if there is a shape in the data
2. Define the best clusters
3. Analyze the clusters to make the best out of it

Data requirements

For the data, we will need the list of the cities, their population, their localizations (coordinates), the type of venues and the number or proportions of each venues in each city.

Data collection

The collection of data take three tools:

Web scrapping with beautifullsoup4 on Wikipedia : https://en.wikipedia.org/wiki/List_of_largest_cities

	City	Country	Population city	Latitude	Longitude
0	Chongqing	China	30751600.0	29.56278	106.55278
1	Shanghai	China	24256800.0	31.22222	121.45806

Then make this datas clean to complete this datas with the coordinates thanks to geocoder.

Then complete the database with foursquare to get the proportions of the venues.

Data understanding

Biggest cities and their population

It was mainly about understand how the foursquare API would save our interest. The explore call just retrieve 100 venues maximum, the idea was to multiply this by 5 by taking 5 coordinates next to the center of every city and launch the explore call.

Taking 5 points in the center of the city

We got a list of venues. But it was only a sample, the bigger, the more representative of the city. Then we divided the number of each venues by the total number of venues in a city.

For instance, a city with 4 restaurant, 1 zoo, 4 hotels and 1 train station would end up like this:

city	restaurant	zoo	hotel	Train station
Example1	0,4	0,1	0,4	0,1

Data preparation

The preparation was to transform the number of venues into proportions to be able to compare them between each city.

	City	Country	Population city	Latitude	Longitude	ATM	Abruzzo Restaurant	Acai House	Accessories Store	Adult Boutique	...	Xinjiang Restaurant	Yakitori Restaurant
0	Chongqing	China	30751600.0	29.56278	106.55278	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0
1	Shanghai	China	24256800.0	31.22222	121.45806	0.0	0.0	0.0	0.0	0.0	...	0.003484	0.0
2	Beijing	China	21516000.0	39.90750	116.39723	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0
3	Lagos	Nigeria	16060303.0	6.45471	3.38876	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0
4	Dhaka	Bangladesh	8906039.0	23.71323	90.39957	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0

Modeling

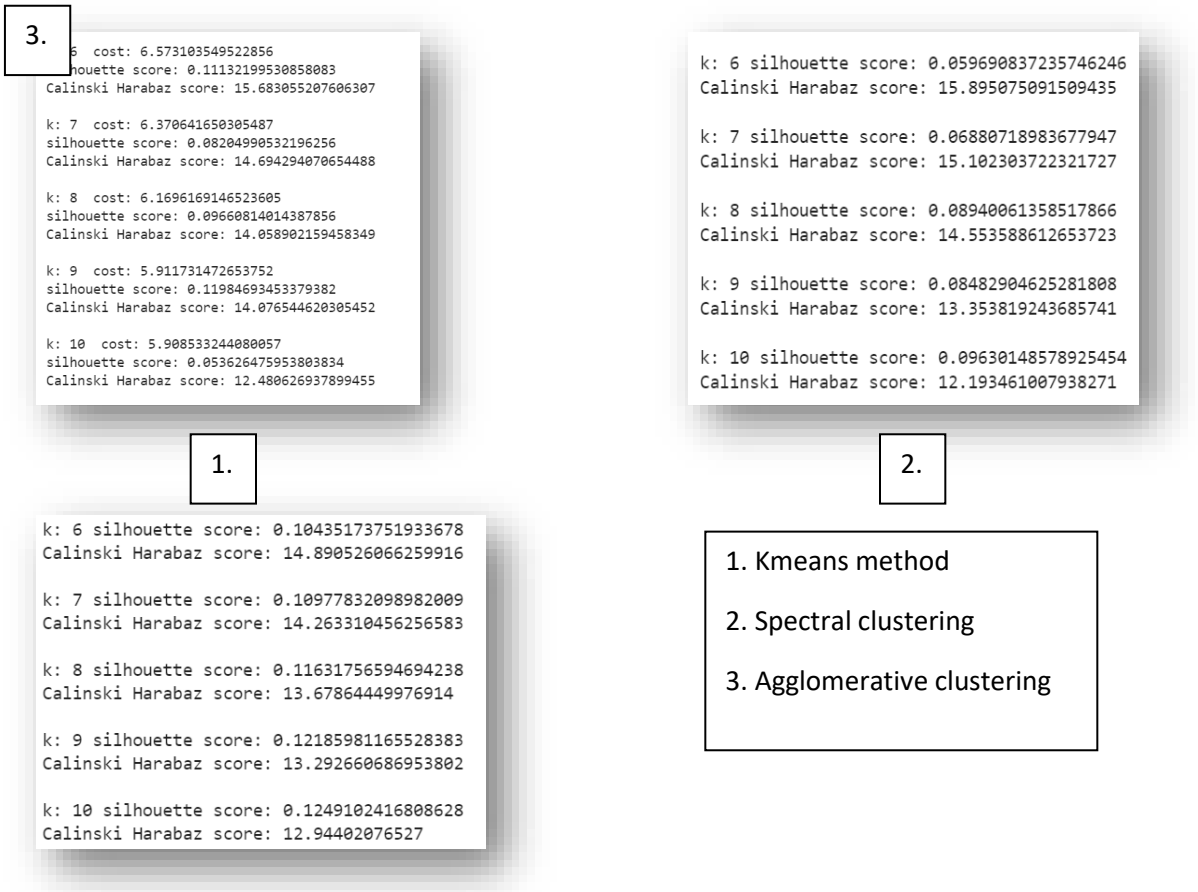
Modeling was about finding the link between the data. Yes we think at first that the cities are link by regions (Asia, Europe, etc…) but it has nothing to do with what we analyze. We analyze the city only based on the proportions of the shops in their center. To begin with, I tried a density based clustering on the cities based on the proportions of the venues. But not results. But thinking about how the datas were (a lot of zeros in a lot of columns), the KMeans clustering seemed to be a better approach. With a random k (=10) at first, I was able to recognize a shape. A cluster for India, a cluster for Japan, and one really global for the occidental countries.

Based on this, different modeling were possible: KMeans, Agglomerative Clustering and Spectral Clustering.

Evaluation

Evaluation was done in two times :

1. Find the best clustering with each method



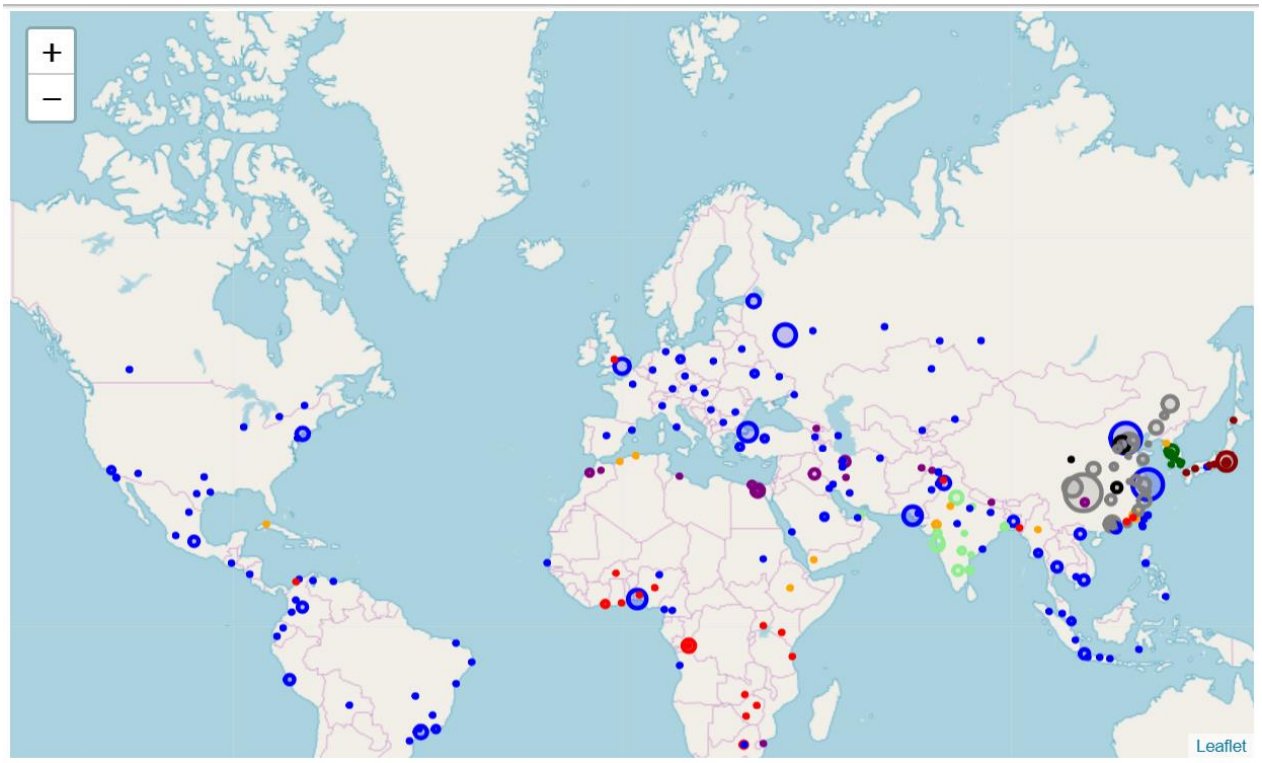
2. Find the best method

Once it was done, we had the clusters to analyze.

Method	k	Inertia	Silhouette Score	Calinski Harabaz Score
KMeans	9	5.95378	0.103405	13.767882
Spectral Clustering	8	None	0.089401	14.553589
Agglomerative Clustering	9	None	0.121860	13.292661

Deployment

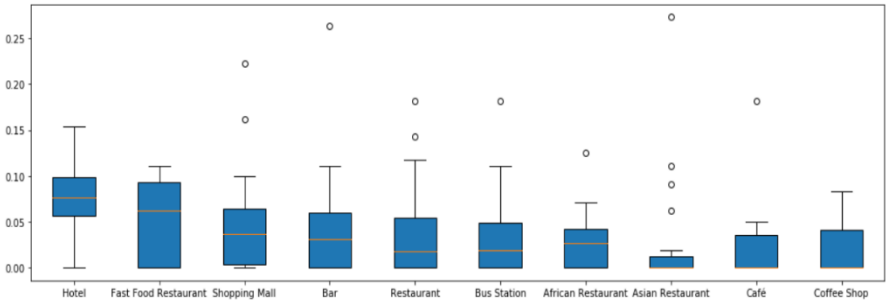
The deployment consisted in analyzing the final clusters



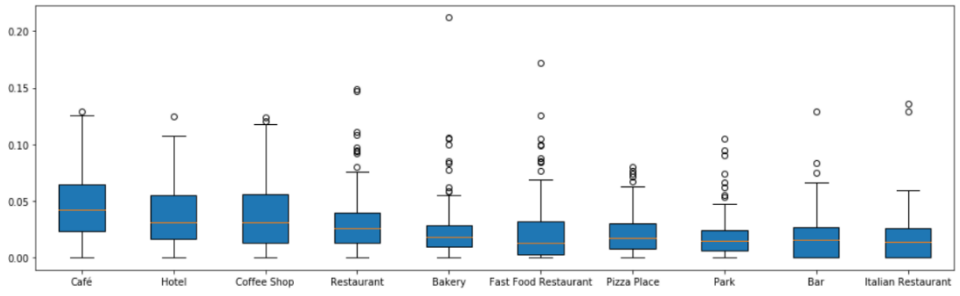
We got 9 clusters thanks to the agglomerative clustering:

Here are the results:

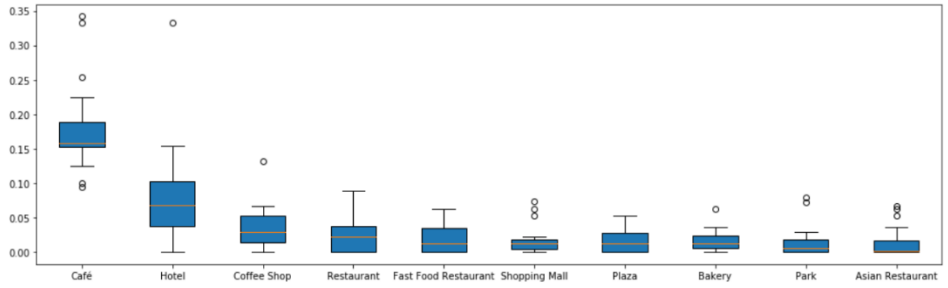
Cluster 1:



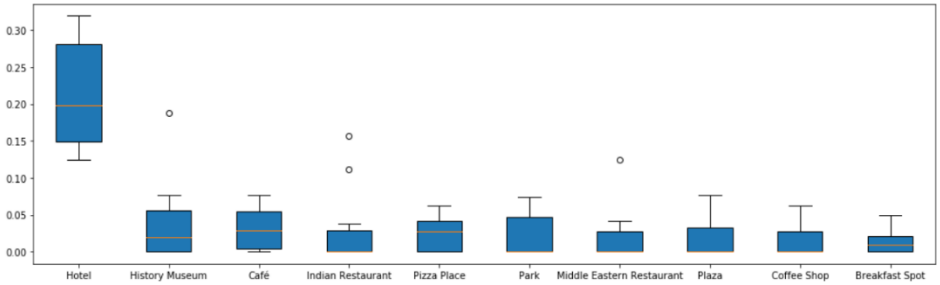
Cluster 2:



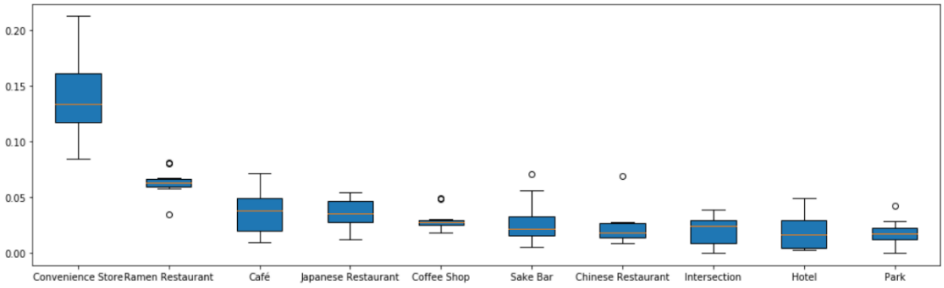
Cluster 3:



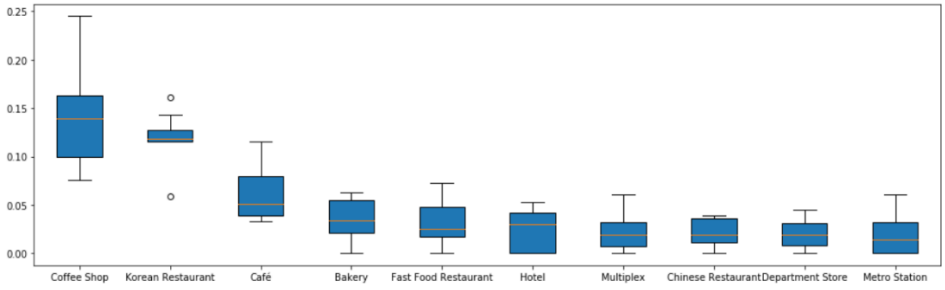
Cluster 4:



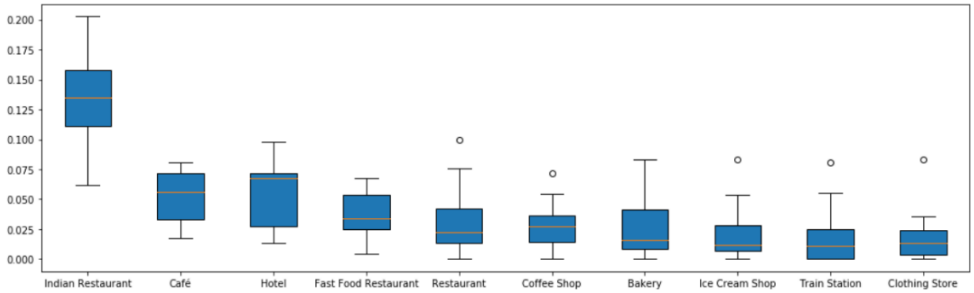
Cluster 5:



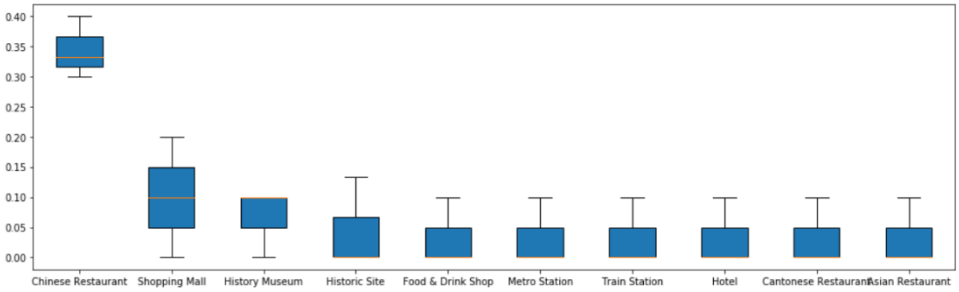
Cluster 6:



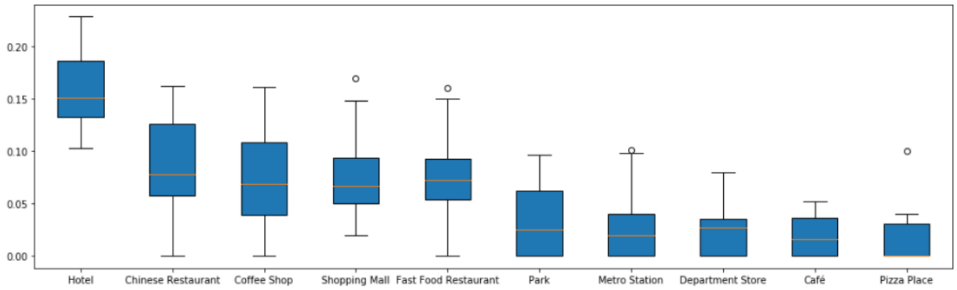
Cluster 7:



Cluster 8:



Cluster 9:



Feedback

We can analyse it on two levels: the regions defined by the clusters and the tops venues in terms of proportions inside a cluster.

First we can see 9 clusters:

1 &2 The first 2 are really blurry, a little bit all over the world, but on the side we can see a lot of outliers in the proportions of venues. This clusters should be the object of a specific analysis because it doesn't give us a lot of informations about the cities in thi clusters.

Then we have some really geographically defined clusters:

3. The third cluster is on the Ecuador and the top venue in this cities are the café, probably because of the sun. In this clusters, there are outliers that are under the average in the café boxplot : It's probably an opportunity to open new coffee places there.

4. The fourth cluster is composed of cities that are close to some water point (rivers, sea, etc), their main venues are hotel and museum, the tourism is probably the thing to develop there.

5. the fifth cluster is centered on Japan. We see that there is a city with a lower proportion of Ramen restaurants, it could be a good idea to develop this in this city, to analyze further with the proportions of the population in this city.

6. The sixth is the South Korea and we see an outlier under the average in term of proportion of Korean Restaurant, maybe an opportunity again.

7. The seventh cluster is about India, we can notice a lot of Indian restaurant, but in third position come the hotels. And a lot of city have a number of hotels above the average, they should take care about this, maybe some cities have too much hotels, to analyze further.

8 & 9 The 8th and 9th clusters are about Chinese cities. But three cities (8th cluster) come out of the group with a low proportion of hotels compare to rest of China. For the others, we can see a big proportion of shopping malls. China has a good consumerist society in fact. It's not for nothing that their economy is gonna overpass the USA's.

As a global feedback, we can see a nice clustering of the biggest cities, but the 2 first clusters should deserve a deeper analysis to get a conclusion on them.