Coursera \ IBM Data Science Professional Certificate Capstone project:

A Battle of Paris Neighbourhoods

Finding a restaurant right place

Sébastien SIME

11-24-2020

Table of Contents:

Executive Summary:

This work is about using Python and the Foursquare location data to draft strategies for a potential restaurant business owner who wants to settle in. The study will be performed only for the city of Paris, the French capital.

Eventually, business owners of other type of business could find insights in the following work as well.

1. Business case: Finding the right place for a restaurant

Searching where a business can bring a service or a product to the end users, is an important question each undertaker or business owner has to face sooner or later. Particularly in the case of a restaurant or a bar we may wonder about the best possible place to open. The answer to that apparently simple question require taking into consideration many parameters regarding what is called the marketing mix (i.e. parameters related to the product \ service, the place, the price and promotion).

Moreover, everyone can access nowadays any kind of information only using internet but, the amount of data and time just looking for the right information and performing some comparisons could be not negligible. Only considering Paris (and not the $\ll$ *île de France* $\gg$ region which is on a more important territory), there are twenty (20) boroughs of four (4) neighbourhoods each which have many possible venues for a restaurant.

Taking the case of an undertaker who wants to open a restaurant in Paris, a data science method can then be used in combination to a defined set of location parameters to advice the owner on possible locations or strategies.

In this report, after choosing some relevant location parameters, I will use Foursquare location data combined with a machine learning algorithm to provide insights on potential location in Paris. I will also use some other datasets to have an idea of the general reputation and accessibility of Paris' neighbourhoods.

1.1 Methodology and Data

1.1.1 Methodology

As mentioned, choosing a restaurant best place, or any business best place, involve lots of parameters which could be related or not. So, choosing require to have in mind the parameters which are more important to the business itself but also to the business owner. For the latter, it is important to have a broad view of Paris before actually picking a place. To perform our analysis, we have chosen to address four majors (4) parameters as analysis axis:

- The general reputation and accessibility of Paris boroughs.
- The global environment in terms of famous venues in Paris neighbourhood.
- Level of restaurants competition in each neighbourhood.

Before going more in-depth in the analysis, we need first to have general information about Paris itself. Having a general idea about Paris, will help to assess Paris subdivisions' reputations according to a French people rank. With this level of information, it will be possible to draw a first broad conclusion about Paris general reputation. At this point, the business owner would have a general idea about Paris and the Paris divisions' reputations.

Knowing the Paris divisions' reputations information, we will explore Paris' neighbourhoods and perform an analysis to categorize the recommended venues. With the different categories or neighbourhoods clusters, it will then be possible to assess the competition level and to advise on potential neighbourhoods' candidates.

1.1.2 Data sources

To tackle the problematic of choosing a potential place for a restaurant in Paris, the following data will be used:

1. Web page providing the rank of the city of Paris compared to others French cities with more than 10000 people. The data will be extracted from REF [1].

2. Web page providing data about general neighbourhood's notations. The data will be extracted from REF [2].

3. Data providing, Latitude, Longitude of Paris' neighbourhoods. This data will be downloaded from REF [3].

4. The recommended existing venues in each neighbourhoods from the Foursquare location data.

The first data source will provide general idea about Paris as a whole compared to others majors' French cities. The second data source will help addressing other more general parameters about the boroughs (Paris divisions) that will complete our understanding of Paris.

The third data source will provide the Paris' neighbourhoods locations (Name, latitude and longitude). The latter are inputs for the Foursquare API calls to explore the neighbourhoods' venues categories.

1.1.3 Data Management

As previously mentioned internet is the major data source. Apart from the Foursquare API, the used datasets will be either directly downloaded from a web page or extracted using web-scraping technics[1]. The data will then be inspected to assess data quality, transformed if necessary and used for analyses (ELT workflow).

Our main objective will be to describe the datasets since we won't have any target features (or features to predict or classify). Remembering that we are looking for the best place for a restaurant in Paris, the datasets will provide some metrics to categorize Paris and built our understanding trough the chosen parameters (reputation, global environment and level of competition). The best analytic approach in this case is a **descriptive analysis with the KMeans clustering algorithm**. So, I performed two clustering analyses:

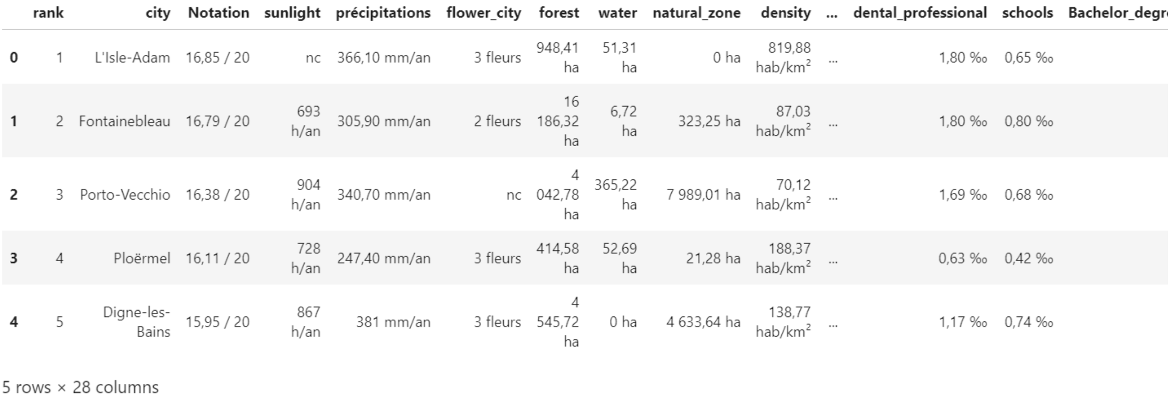
- One with the general Paris' borough notations.
- Another with the Paris neighbourhoods' venues data frames.

Finally, the Paris' boroughs and neighbourhoods' locations data will also be used to place boroughs and venues on a maps using the Python Folium library for categories visualisations.

[1] Principally the BeautifulSoup python library

Below you can see the list of data frames' heads for each data source after transformations at each step:

- ***Step 1: Paris general description with the data frame of most enjoyable French cities with more than 10000 people:***



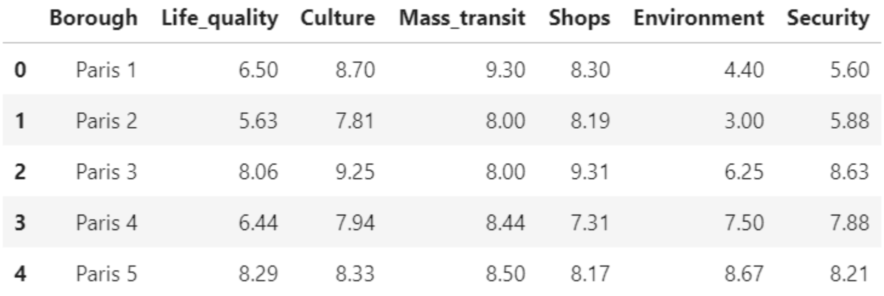
	rank	city	Notation	sunlight	précipitations	flower_city	forest	water	natural_zone	density	...	dental_professional	schools	Bachelor_degr
0	1	L'Isle-Adam	16,85 / 20	nc	366,10 mm/an	3 fleurs	948,41 ha	51,31 ha	0 ha	819,88 hab/km^2	...	1,80 ‰	0,65 ‰	
1	2	Fontainebleau	16,79 / 20	693 h/an	305,90 mm/an	2 fleurs	16 186,32 ha	6,72 ha	323,25 ha	87,03 hab/km^2	...	1,80 ‰	0,80 ‰	
2	3	Porto-Vecchio	16,38 / 20	904 h/an	340,70 mm/an	nc	4 042,78 ha	365,22 ha	7 989,01 ha	70,12 hab/km^2	...	1,69 ‰	0,68 ‰	
3	4	Ploërmel	16,11 / 20	728 h/an	247,40 mm/an	3 fleurs	414,58 ha	52,69 ha	21,28 ha	188,37 hab/km^2	...	0,63 ‰	0,42 ‰	
4	5	Digne-les-Bains	15,95 / 20	867 h/an	381 mm/an	3 fleurs	4 545,72 ha	0 ha	4 633,64 ha	138,77 hab/km^2	...	1,17 ‰	0,74 ‰	

5 rows × 28 columns

Figure 1: Most enjoyable French cities with more than 10000 people data frame extract

This data frame provide general information about 1000 major French cities and their respective notations.

- ***Step 2: Paris boroughs description with the data frame of Paris boroughs notations:***



	Borough	Life_quality	Culture	Mass_transit	Shops	Environment	Security
0	Paris 1	6.50	8.70	9.30	8.30	4.40	5.60
1	Paris 2	5.63	7.81	8.00	8.19	3.00	5.88
2	Paris 3	8.06	9.25	8.00	9.31	6.25	8.63
3	Paris 4	6.44	7.94	8.44	7.31	7.50	7.88
4	Paris 5	8.29	8.33	8.50	8.17	8.67	8.21

Figure 2: Boroughs notation data frame extract

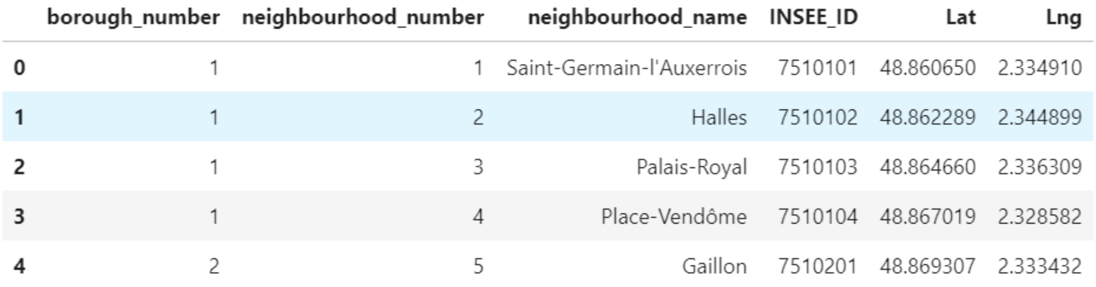
This data frame provide notations (out of 10) of some parameters. The notations have been given by a sample of French people and summarized on the web-page REF [2].

Since the data was well formatted, a clustering analysis with the KMeans algorithm has been used to categorize the dataset.

In order to find the suitable number of clusters to use for the analysis (the input parameter for the kmeans algorithm), we used the "Elbow method[2]" combined with the silhouette score.

[2] Follow REF [4] and REF [5] for more details on the method and score

- ***Step 3: Paris neighbourhoods location data frame:***

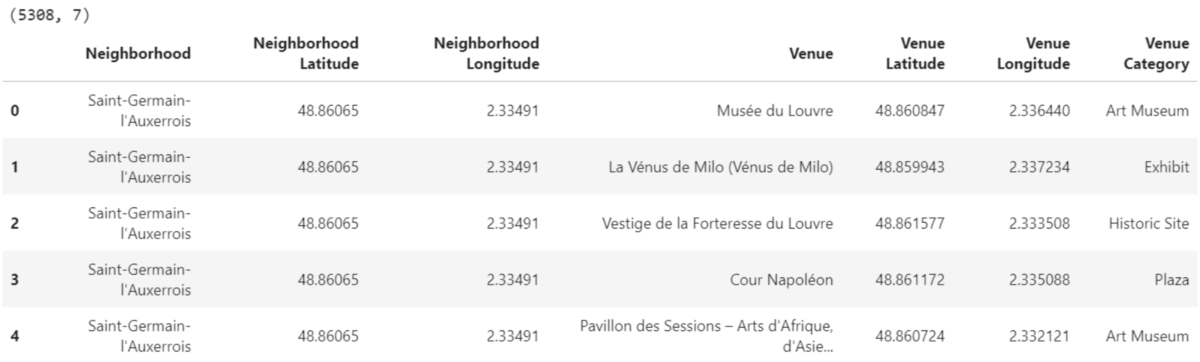


	borough_number	neighbourhood_number	neighbourhood_name	INSEE_ID	Lat	Lng
0	1	1	Saint-Germain-l'Auxerrois	7510101	48.860650	2.334910
1	1	2	Halles	7510102	48.862289	2.344899
2	1	3	Palais-Royal	7510103	48.864660	2.336309
3	1	4	Place-Vendôme	7510104	48.867019	2.328582
4	2	5	Gaillon	7510201	48.869307	2.333432

Figure 3: Neighbourhoods location data frame extract

This data frame provide inputs (using latitude and longitude) for the Foursquare API calls.

- ***Step 4: Paris' neighbourhoods description with the recommended Paris neighbourhoods' venues data frame:***



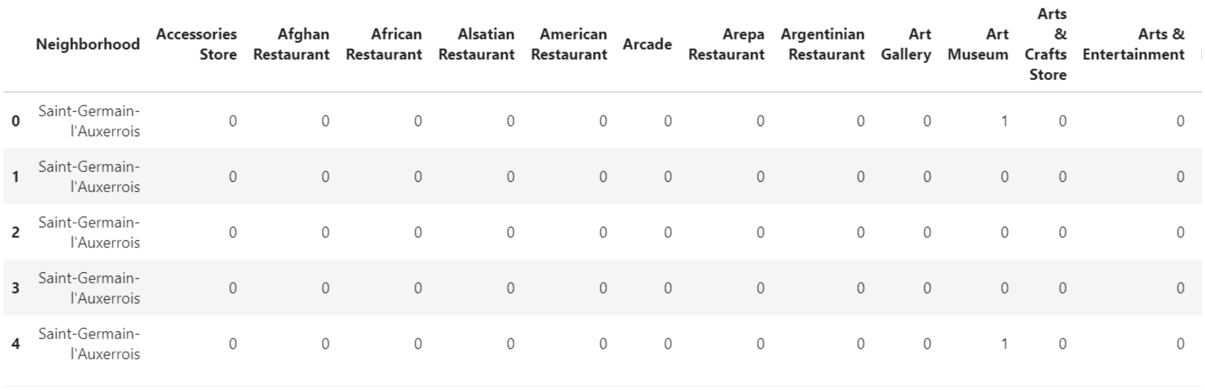
```
(5308, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Saint-Germain-l'Auxerrois	48.86065	2.33491	Musée du Louvre	48.860847	2.336440	Art Museum
1	Saint-Germain-l'Auxerrois	48.86065	2.33491	La Vénus de Milo (Vénus de Milo)	48.859943	2.337234	Exhibit
2	Saint-Germain-l'Auxerrois	48.86065	2.33491	Vestige de la Forteresse du Louvre	48.861577	2.333508	Historic Site
3	Saint-Germain-l'Auxerrois	48.86065	2.33491	Cour Napoléon	48.861172	2.335088	Plaza
4	Saint-Germain-l'Auxerrois	48.86065	2.33491	Pavillon des Sessions – Arts d'Afrique, d'Asie...	48.860724	2.332121	Art Museum

Figure 4: Data frame before hot-encoding

We can see on the latter data frame that we will be dealing with more than 5000 venues.

In order to use the "KMeans" clustering algorithm, I transformed (hot-encoding) the above venue category feature to create the data frame below with venues categories are in rows and the neighbourhoods in columns.



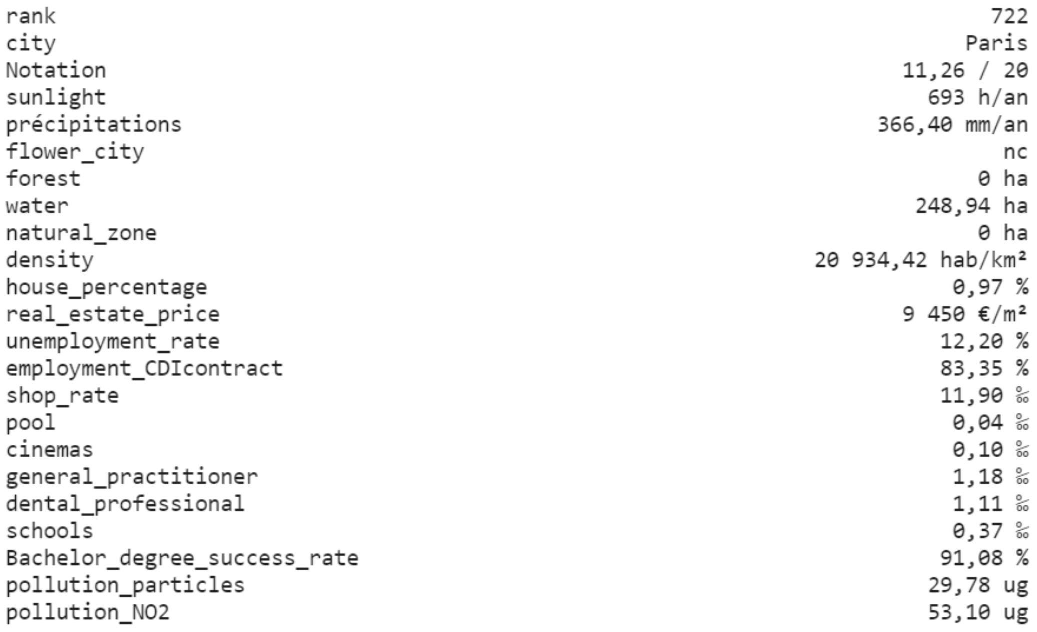
	Neighborhood	Accessories Store	Afghan Restaurant	African Restaurant	Alsatian Restaurant	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment
0	Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0	1	0	0
1	Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0	0	0	0
2	Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0	0	0	0
3	Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0	0	0	0
4	Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0	1	0	0

Figure 5: Data frame after hot-encoding

1.2 Results section

1.2.1 General idea about Paris

Paris is the French capital but by far is not the most enjoyable French city to live in according to the source. The city is ranked 722th behind others French major cities like Marseille or Nice. This statement is totally in contrast with the high population density (about 20934 people per square kilometre) and the fact that there are many companies headquarters (about 83% employment rate) are in Paris.



```
rank                                         722
city                                       Paris
Notation                              11,26 / 20
sunlight                                693 h/an
précipitations                     366,40 mm/an
flower_city                                   nc
forest                                      0 ha
water                                  248,94 ha
natural_zone                                0 ha
density                       20 934,42 hab/km²
house_percentage                          0,97 %
real_estate_price                      9 450 €/m²
unemployment_rate                        12,20 %
employment_CDIcontract                   83,35 %
shop_rate                                11,90 ‰
pool                                      0,04 ‰
cinemas                                   0,10 ‰
general_practitioner                      1,18 ‰
dental_professional                       1,11 ‰
schools                                   0,37 ‰
Bachelor_degree_success_rate             91,08 %
pollution_particles                     29,78 ug
pollution_NO2                           53,10 ug
```

Figure 6: Short numerical Paris description REF [1]

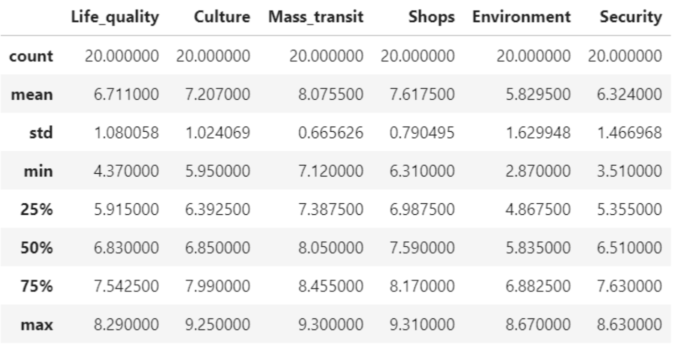
Nevertheless, the city remain in the average in terms of preferences. Paris is also well known for high real estate prices and then the high building density which means not enough green zones.

1.2.2 Paris boroughs description

Paris is divided into 20 boroughs which are not ranked the same according to REF [2]. The provided features notations concern:

- The life quality.
- The culture.
- The mass transit.
- Shops.
- The environnent (pollution, green zones etc.).
- The Security.

All the notations mean values are between 6 and 8 (out of 10). The worst notation is for the security feature and the best for shops. We can also see on the table below that, the mass transit notation mean is about 8, this means that **the transport coverage (and thus the accessibility) is rather good for all the boroughs**.



	Life_quality	Culture	Mass_transit	Shops	Environment	Security
count	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000
mean	6.711000	7.207000	8.075500	7.617500	5.829500	6.324000
std	1.080058	1.024069	0.665626	0.790495	1.629948	1.466968
min	4.370000	5.950000	7.120000	6.310000	2.870000	3.510000
25%	5.915000	6.392500	7.387500	6.987500	4.867500	5.355000
50%	6.830000	6.850000	8.050000	7.590000	5.835000	6.510000
75%	7.542500	7.990000	8.455000	8.170000	6.882500	7.630000
max	8.290000	9.250000	9.300000	9.310000	8.670000	8.630000

Figure 7: Boroughs notations ranges

Another major point is that above boroughs features appear to be highly correlated. The correlation coefficients are all positive, this means for example that if there are security issues, quality of life tends to decline and environmental considerations also tend to be less important.

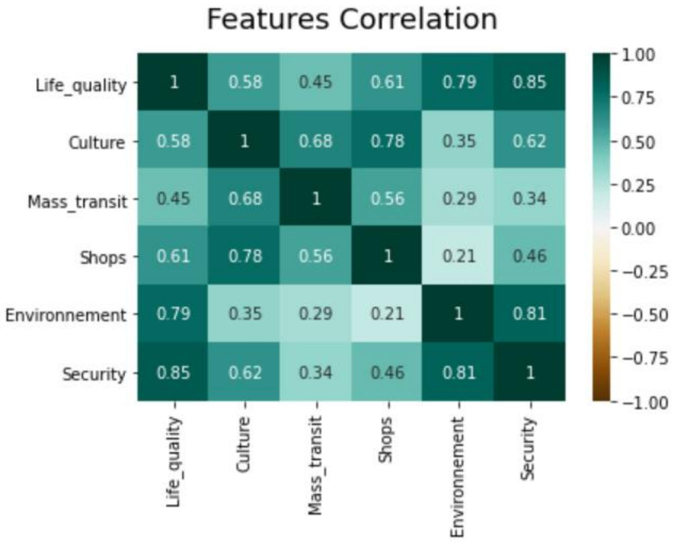


Figure 8: Boroughs features correlation coefficients

Since the data frame is well formatted (all the features are float type), to have a synthetic global description of Paris boroughs, KMeans algorithm has been used. But prior to the clustering analysis, the optimal clusters number has been computed using the Elbow method combined with the silhouette score. This method is simply an indication for a better cluster analysis but not an obligation. You can find more information about the elbow method with REF [4] and REF [5].

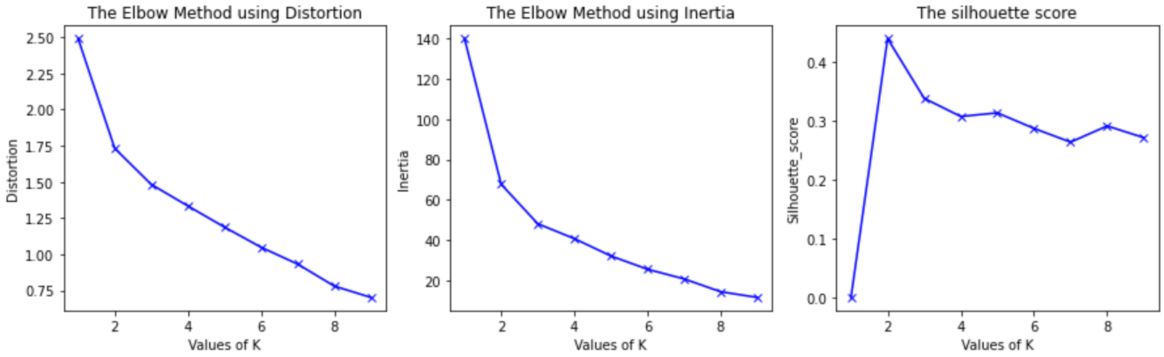


Figure 9: Elbow method graphs

On the distortion and the inertia graphs, we can see that first sharp change in slope are observed for k equal to 2 and this is confirmed by the silhouette score which reaches its maximum for two clusters.

Using 2 as a wanted number of clusters, allowed to group Paris boroughs as shown on the map below:

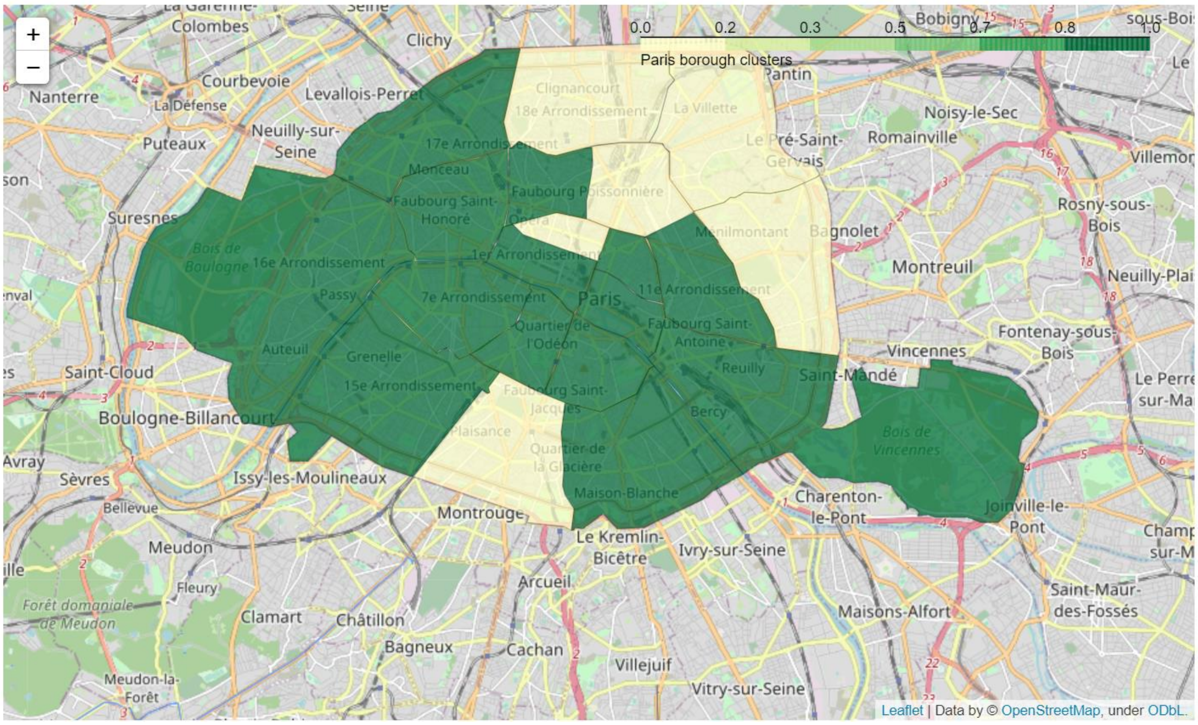


Figure 10: Paris boroughs Clusters (using folium)

The cluster 0 (in light yellow) represent boroughs with less reputation (security + environment) than the cluster 1 (in green). A Z-test of this statement as a hypothesis provided p-values of about 4.3e-5 and 4.4e-4[3] for the security and the environment features.

We can conclude that boroughs 2, 10, 14, 18, 19 and 20 (boroughs in cluster 0) have lower reputations than the other in cluster 1 regarding the security and the environment.

1.2.3 Paris Neighbourhoods description

For the neighbourhoods' venue categories exploration, I used the Foursquare API. As a reminder, the latter takes as inputs data the location data (latitude and longitude) and provide for each couple of coordinates a list of recommended venues. So in order to use the API, the Paris neighbourhoods' location data have been downloaded from REF [3] (a French open source portal).

Focusing on venues categories and using all of the recommended venues, French restaurants is the most represented category before Hotel, with about more than 650 French restaurants in Paris (as you can see on the picture below).

[3] The values of the Jupyter Notebook have been divided by two here since the test hypotheses was one-sided.

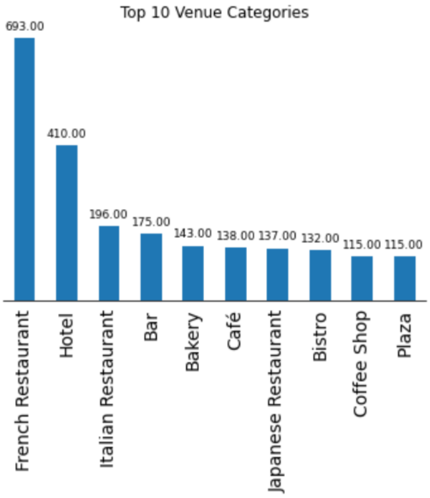
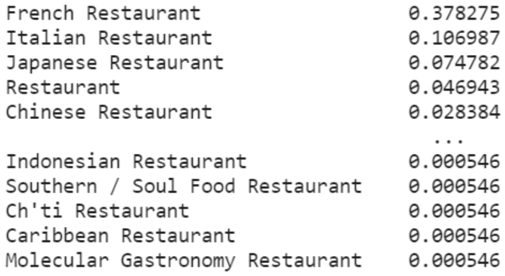


Figure 11: Top 10 venues categories in Paris

I would have expected to see categories like museums or historic places among the top 5 but the gastronomy is globally the first major category. Among all kinds of restaurants or restaurants' specialties, the most represented are listed below:



```
French Restaurant               0.378275
Italian Restaurant              0.106987
Japanese Restaurant             0.074782
Restaurant                      0.046943
Chinese Restaurant              0.028384
                                   ...
Indonesian Restaurant           0.000546
Southern / Soul Food Restaurant 0.000546
Ch'ti Restaurant                0.000546
Caribbean Restaurant            0.000546
Molecular Gastronomy Restaurant 0.000546
```

French specialties are the most represented (after all this is France) but what it is interesting to see is that the others most represented restaurants are Italian and Asian.

Finally as we could have anticipated, "Champs-Elysées" with many others highly touristic neighbourhoods, reach the limit of 100 venues.

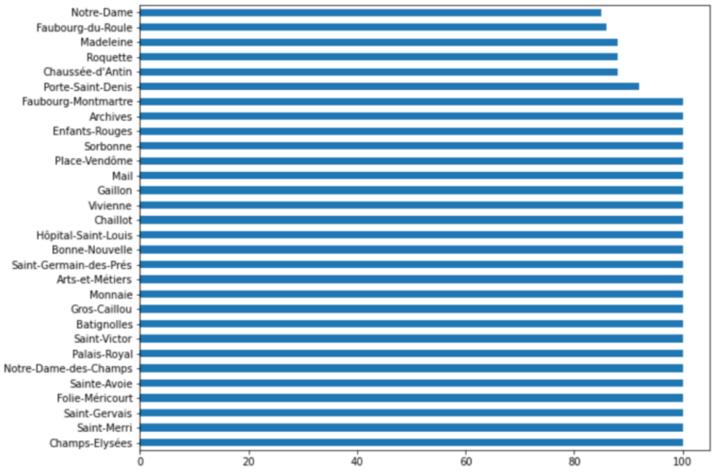
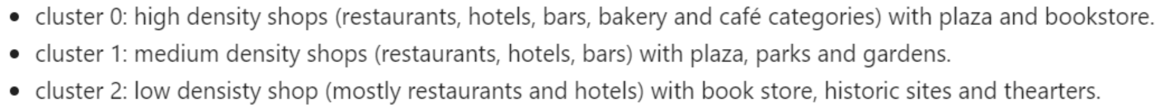


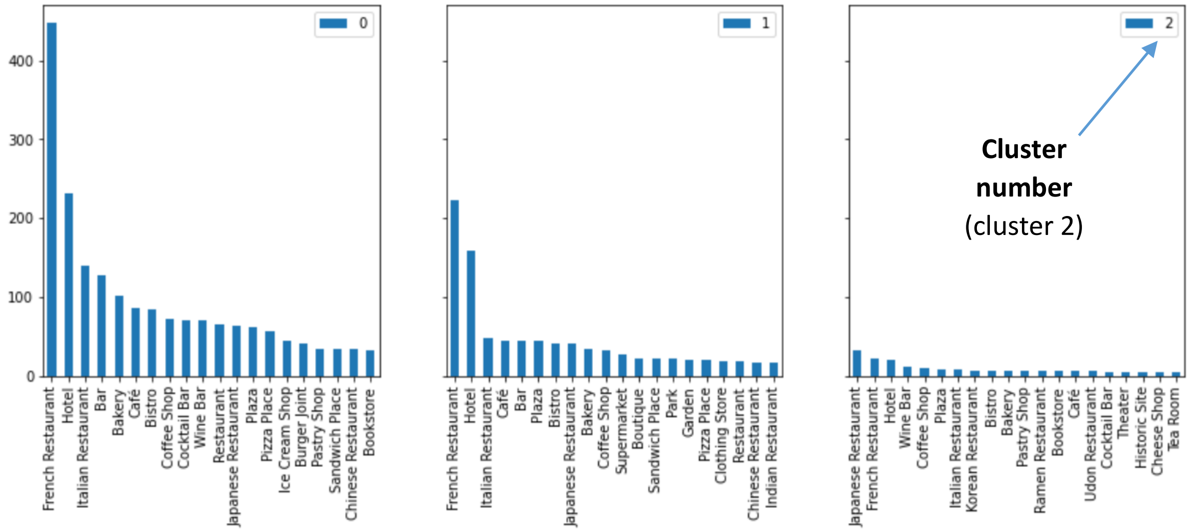
Figure 12: Most represented neighbourhoods

After transformation, the clustering analysis with the KMeans algorithm allowed to divide Paris neighbourhoods into 3 cluster:



- cluster 0: high density shops (restaurants, hotels, bars, bakery and café categories) with plaza and bookstore.
- cluster 1: medium density shops (restaurants, hotels, bars) with plaza, parks and gardens.
- cluster 2: low densisty shop (mostly restaurants and hotels) with book store, historic sites and thearters.

The picture below shows the clusters composition (the first twenty categories):

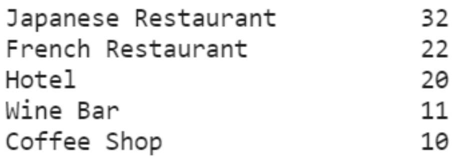


Cluster number (cluster 2)

Figure 13: Top 20 categories of Paris neighbourhoods clusters

French and Italian restaurants are present all over Paris but in **different density**. The cluster 2 has the lowest competition level.

The first 5 categories (and the associated number) of the cluster 2 is listed below:

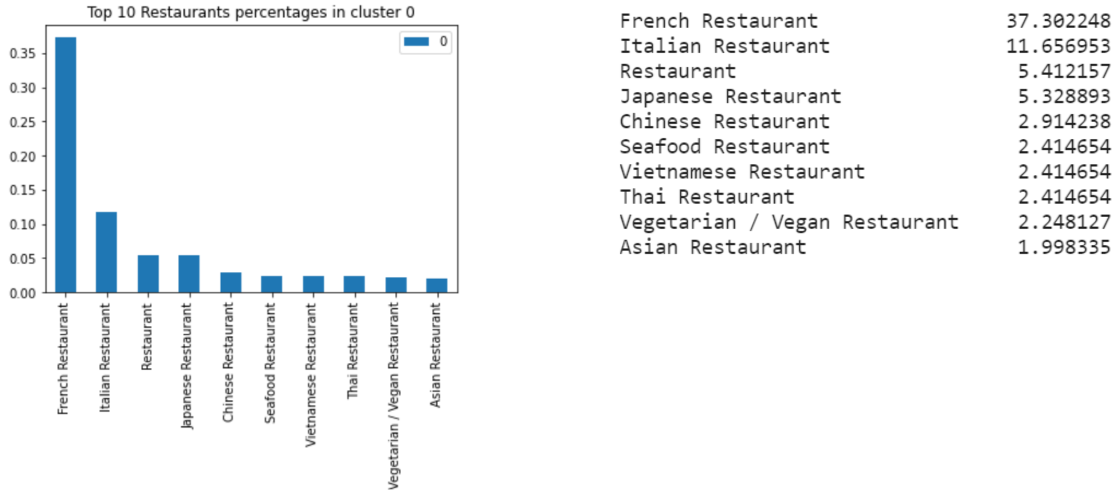


```
Japanese Restaurant         32
French Restaurant           22
Hotel                       20
Wine Bar                    11
Coffee Shop                 10
```

As you can see, the number of Japanese restaurants (32) is above the number of french restaurants in the associated neighbourhoods.

Focusing on restaurants in each cluster:

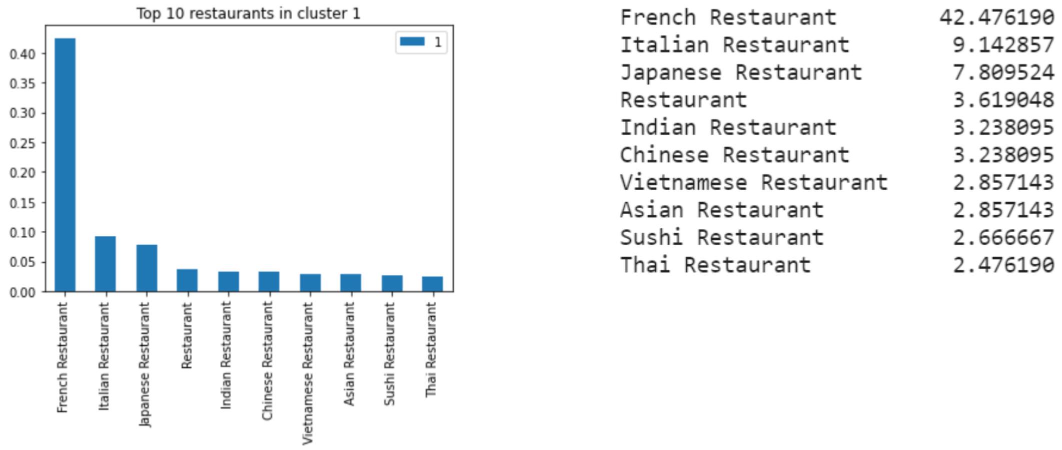
- **In neighbourhood cluster 0** (high density shops cluster): about 37% represent French restaurants, about, about 12% represent Italian restaurants, about 14% represent Asian specialties.



```
French Restaurant                37.302248
Italian Restaurant               11.656953
Restaurant                        5.412157
Japanese Restaurant               5.328893
Chinese Restaurant                2.914238
Seafood Restaurant                2.414654
Vietnamese Restaurant             2.414654
Thai Restaurant                   2.414654
Vegetarian / Vegan Restaurant     2.248127
Asian Restaurant                  1.998335
```

Figure 14: Restaurants in cluster 0

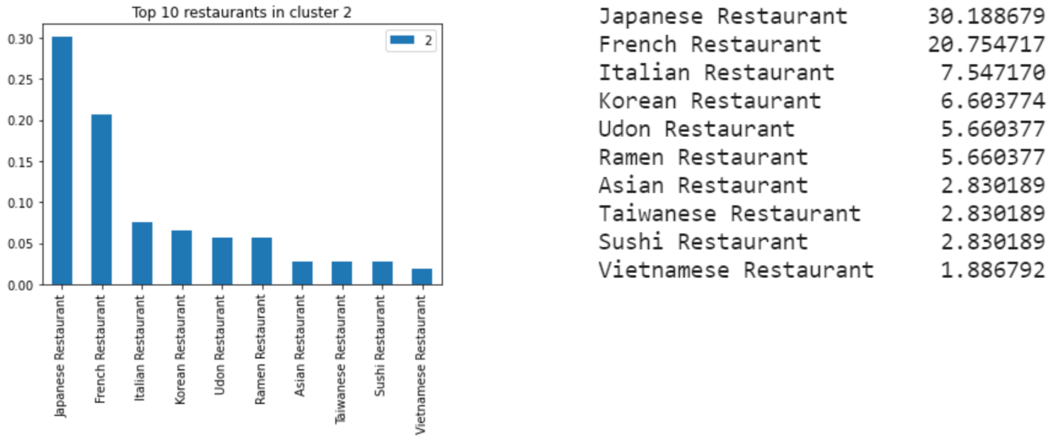
- **In neighbourhood cluster 1**: about 42% of restaurants are French restaurants, about 9% are Italian and about 45%.



```
French Restaurant        42.476190
Italian Restaurant        9.142857
Japanese Restaurant       7.809524
Restaurant                3.619048
Indian Restaurant         3.238095
Chinese Restaurant        3.238095
Vietnamese Restaurant     2.857143
Asian Restaurant          2.857143
Sushi Restaurant          2.666667
Thai Restaurant           2.476190
```

Figure 15: Restaurants in cluster 1

- **In neighbourhood cluster 2**: there is a majority of Asian restaurant despite the lower shop density.



```
Japanese Restaurant      30.188679
French Restaurant        20.754717
Italian Restaurant        7.547170
Korean Restaurant         6.603774
Udon Restaurant           5.660377
Ramen Restaurant          5.660377
Asian Restaurant          2.830189
Taiwanese Restaurant      2.830189
Sushi Restaurant          2.830189
Vietnamese Restaurant     1.886792
```

Figure 16: Restaurants in cluster 2

When the competition level decrease, the Asian restaurant density increase, the French restaurant density decrease with the Italian restaurant density as well.

Finally, on a map below, we can see neighbourhoods clusters represented all over Paris.

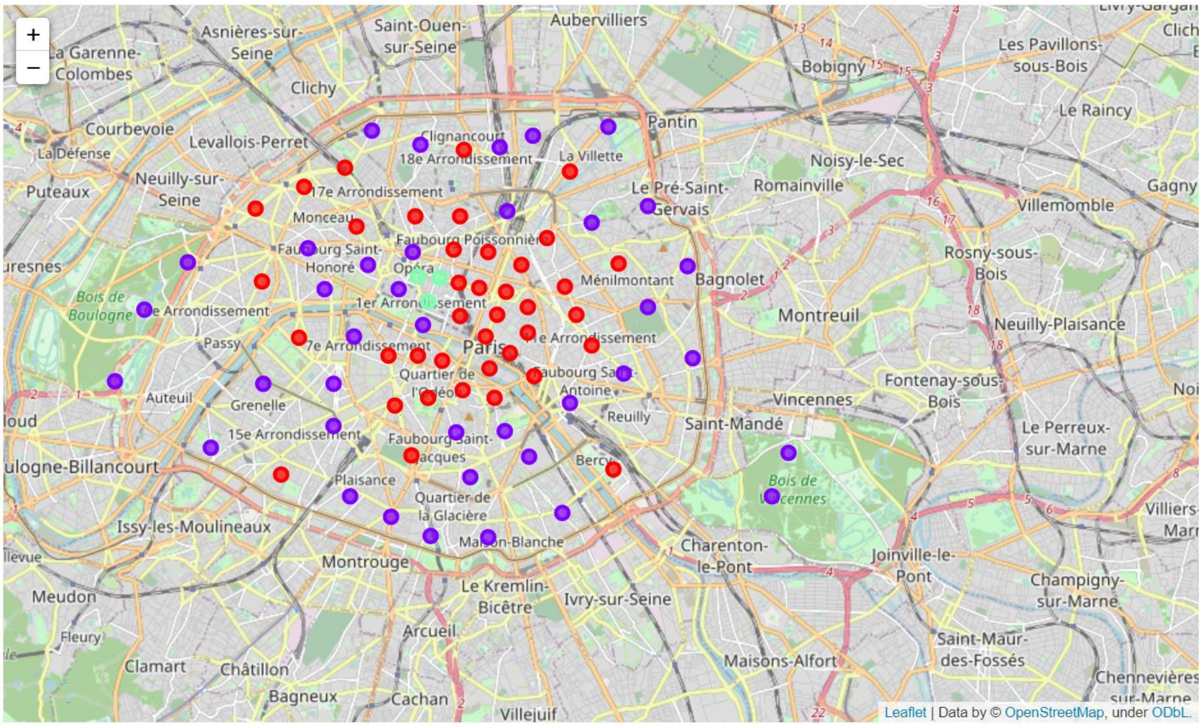


Figure 17: Neighbourhood clustering

On the map above, we observe:

- Neighbourhood cluster 0 (In red): high density for restaurants (French, Italian) and hotels
- Neighbourhood cluster 1 (purple): medium density fort restaurants and hotels
- Neighbourhood cluster 2 (in light green): low density (mostly Asian, French and Italian restaurants)

1.3 Discussion

At the beginning of these study we didn’t' really know a single think about Paris. The study provided information about the general reputation of Paris boroughs and about the level of competition in each neighbourhoods.

On the map below, all of the clusters have been superimposed on the same map (reputation boroughs clusters and neighbourhoods’ density shop clusters).

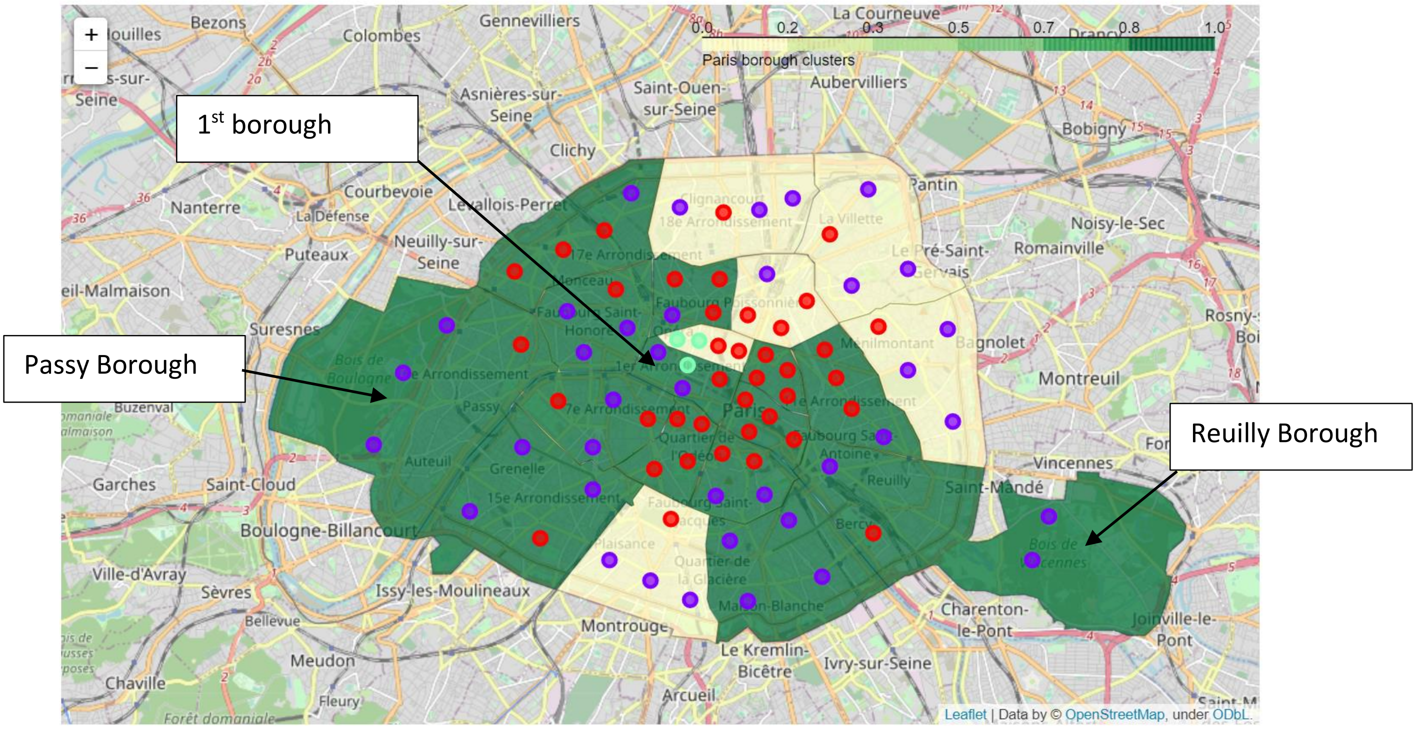


Figure 18: Borough and neighbourhoods clustering

We can see that right at the middle of Paris there is the lowest French restaurants density but the highest Asian restaurants. Famous recommended venues are mainly shops, hotels, cafés, bars and plazas.

Looking at the above map and the clusters compositions we see that we have a shop concentration in the most secure borough clusters. The shop density is rather high in these boroughs. Nevertheless, shops density in boroughs with less reputation appear to be rather medium.

Shops appear to be much concentrated around the Paris centre. The competition at the centre of Paris is then more important and especially for French, Italian and finally Asian specialties.

However, we can see that Passy and Reuilly boroughs (12th and 16th boroughs) have both a medium shop density, a good reputation and more important surface. This could represent an opportunity for a restaurant owner as the neighbourhoods are distant from one another.

We have also seen that the accessibility is not really an issue since the mass transit coverage appear to be rather good all over Paris. But regarding the boroughs, the security and environmental considerations are divisive. This allowed, with a clustering analysis, to divide the 20 boroughs into two groups (good and less good borough reputation clusters).

The clustering analysis on neighbourhoods showed that the level of competition all over Paris is not the same. Regarding the restaurants, dominant categories are French, Italian and Asian. In the global Asian restaurant categories there are a more diversified countries' specialties (Japanese, Chinese, Thai, etc.).

So, according to its preferences and looking at the last Paris map, a business owner could have different strategies:

- A strategy could be to go into higher shop density neighbourhoods with aggressive prices trying to compete or with a not well represented restaurant specialties like may be Caribbean \ African or even Mexican.

- Another way could be to develop a new restaurant concept in lower shop density neighbourhood (like right in the middle of Paris).

- As stated before, the 12th and the 16th Paris boroughs appear to be large enough with a rather medium shop density. So these neighbourhoods could represent good opportunities.

- Finally, we have seen that in each neighbourhoods clusters there are plazas that could be used (if allowed by the town hall) for a food truck approach or something similar.

The latter option could be competitive enough especially if combined for example, with the localization of a web application to guide customers to the food truck.

1.4 Conclusion

Our objective was to find the best place to open a restaurant in Paris. So first we had to choose a set of parameters to define what a best place could be. As parameters we selected the general reputation (as a combination of security, the accessibility, quality of life and others) and the level of competition. Of course these are not all the relevant parameters, even if the latter should be among the first to be considered before going further.

Having the selected parameters in mind, we extracted data from web pages on internet (directly downloading dataset or web-scraping web pages) mainly about Paris itself, Paris boroughs notations and Paris neighbourhoods' locations. After cleaning, we used the collected data to build clustering analyses with the KMeans algorithm as a descriptive approach.

The analyses allowed to divide Paris boroughs according to the reputation and Paris neighbourhoods according to the shop density. The location data combined with the latter Paris subdivisions allowed to represent all of the clusters on a Paris map for visualisations.

Our results was to show how Paris could be divided and to draft some strategies for potential business owners according to the identified clusters. Globally everything depends on the marketing mix the business owner wants to apply: he could be aggressive with prices in high density neighbourhoods, be a niche player targeting not enough represented specialties or try a cutting strategy to differentiate the offer.

However, this study is only the beginning, as there are a lot more parameters to define a best place and then to address the problematic. To continue the work we could gather more information about shops in each neighbourhoods. We could also (with a Foursquare Paid account) collect more information about restaurants (tips, ratings and more) to be more specific about the strategy and obviously know more about potential shops clients and competitors.

An important aspect that we didn't mentioned is the Paris demography. The business owner will need to know more about potential clients to work on its marketing positioning.

But every study needs an ends for a new beginning. So, I thank you for your attention and hope I have given you valuable insights for your idea to come true.

2. References

REF [1] http://www.linternaute.com

REF [2] https://www.parisenigmes.com/guide-arrondissement-paris

REF [3] https://opendata.paris.fr

REF [4] https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

REF [5] https://en.wikipedia.org/wiki/Silhouette_(clustering)

REF [6] (Paris cover page image) https://arzotravels.com/the-most-beautiful-city-at-night-paris/