INTRODUCTION AND BUSINESS PROBLEM

INTRODUCTION

The capstone assignment project aims to study the business opportunities in Paris (France) intra-muros (without the suburbs) can offer to any contractors, especially in the domain of foods and restaurants but also for any other good deals like looking for a place to stay with amenities and a good quality of life.

Paris, the capital city of France, is a large metropolis with more than 2.2 million inhabitants with a rich history and a cosmopolitan and multicultural population. Initiating a business in such a context presents enormous risks of failure without further and deep study.

Paris intra-muros is divided into 20 districts from the first to the 20th. The district is an administrative division, headed by an elected mayor and created since 1859.

Nowadays some districts are like a big town in the town. Thus, in the 19th districts, the population of approximately 180,000 inhabitants is the equivalent of a town like Reims. But in the 1st district (Louvre) there are only about 17,000 inhabitants and in the 8th (Elysées) approx. that 40,000 inhabitants. District areas differ from each other from the smallest (the 2nd - 99 ha) to the largest (the 15th - 848 ha).

The districts form a spiral, a snail shell, starting from the center outwards and turning in the direction of the needles of an old clock. The 1st district is in the center, and the 20th district in the east.

For each district there are some clichés: the 16th arrondissement is reputed as the "rich" district, the 5th is for the universities, the 11th (where is the concert hall "Bataclan" that the terrorists attacked on November 13, 2015 making 90 deaths) has become a trendy district where you meet at the end of the week for a drink. The 13th is Chinatown while the 18th, the 19th and the 20th are popular neighborhoods.

BUSINESS PROBLEM

In the scope of this capstone, we would like to segment the 20 districts of Paris to group those who presents some similarities and characteristics in terms of venues categories around each district but also considering the number of inhabitants in the analysis. The objective is to have a tool to guide any users for decision making to respond to the following questions:

- If I want to open a new restaurant in Paris, depending on the type of my restaurant, in which district would I create it according the existing restaurant in the area ?

- If I want to rent a house, in which district can I look first to fullfill my needs in terms of local amenities and quality of life?

MORE ON THE 20 PARIS DISTRICTS

1st district : the Louvre former palace (royal power and his court) became museum.

2nd district : Bourgeois neighborhood, beautiful architecture

3rd district : Museums, old and historical streets, pretty markets

4th district : The Paris City Hall, Le Marais.

5th district : La Sorbonne (University since 12nd century)

6th district : Saint-Germain-des-Prés et du Jardin du Luxembourg - Most expensive district

7th district : Tour Eiffel - embassies and ministries - National Assembly- fine items and haute-couture or ready-to-wear boutiques

8th district : is luxury and fashion. Arc de Triomphe, the Champs-Elysées, the Place de la Concorde, the Church of the Madeleine and of course the Elysee Palace seat of the Presidency of the Republic.

9th district : La Pigalle - Night live

10th district : 2 train stations

11th district : Bars, restaurants, rues commerçantes

12th district : Parks (floral, Vincennes, Bercy)

13th district : Chinatown, festive and arty district on the heights of Paris

14th district : Garden

15th distrcit : Residential, parks

16th district : Residential, Museums, Bourgeois, Art Nouveau et néo-classicisme bourgeois

17th district : Bourgeois, populaire and animated.

18th district : The basilica of the Sacred Heart of Montmartre overlooking Paris

19th district : Parks, museums, popular

20th ditstrict : popular and festive of old and continue to mix cultures

---

DATA INVENTORY – DESCRIPTION - SOURCES

DATA INVENTORY AND DESCRIPTION

To achieve our goal thru this capstone assignment, we should have the following data:

- The geospatial coordinates of Paris (France)
- The order and the name of the 20 Paris Districts.
- The coordinates of a location in each district: we can take here the well-known coordinates of the Hall of the City in each district.
- The number of inhabitants in each district.
- The area of each district

In addition, we need to get the existing venues in each district. For each venue, we will need to know the category of the venue and their coordinates for our analysis. We will limit the number of venues to 150 and we will cover on 1,500 meters around of the location point.

After loading, cleaning and transforming the data, we will collect all the venues categories in each district and try to use a machine learning algorithm to group districts in some clusters presenting the same characteristics. From this result, we will try to respond to the business problem asked for that assignment.

DATA SOURCES

About the data related to the 20 Paris districts, Wikipedia provides on the name, area, number of inhabitants, density for each district [1]. Specific information on each district can be collected also from this page [2].

To get the data about the venues, categories and their coordinates, we will use the Foursquare API [3] which is the best place to collect the venues locations in any town over the world. We will use the standard free subscription to achieve this exercise.

[1] https://fr.wikipedia.org/wiki/Arrondissements_de_Paris#R%C3%A9partition_de_la_population

[2] https://www.unjourdeplusaparis.com/paris-essentiel/paris-par-arrondissements

[3] https://enterprise.foursquare.com/products/places

	Arr	Nom	Latitude	Longitude	Superficie	Population	Densite
0	1er	Louvre	48.866879	2.340376	183	16545	9041
1	2e	Bourse	48.866879	2.340376	99	20796	21006
2	3e	Temple	48.864025	2.361470	117	35049	29956
3	4e	Hotel de Ville	48.856804	2.351056	160	27146	16966
4	5e	Pantheon	48.846249	2.344604	254	59333	23359

Methodology section

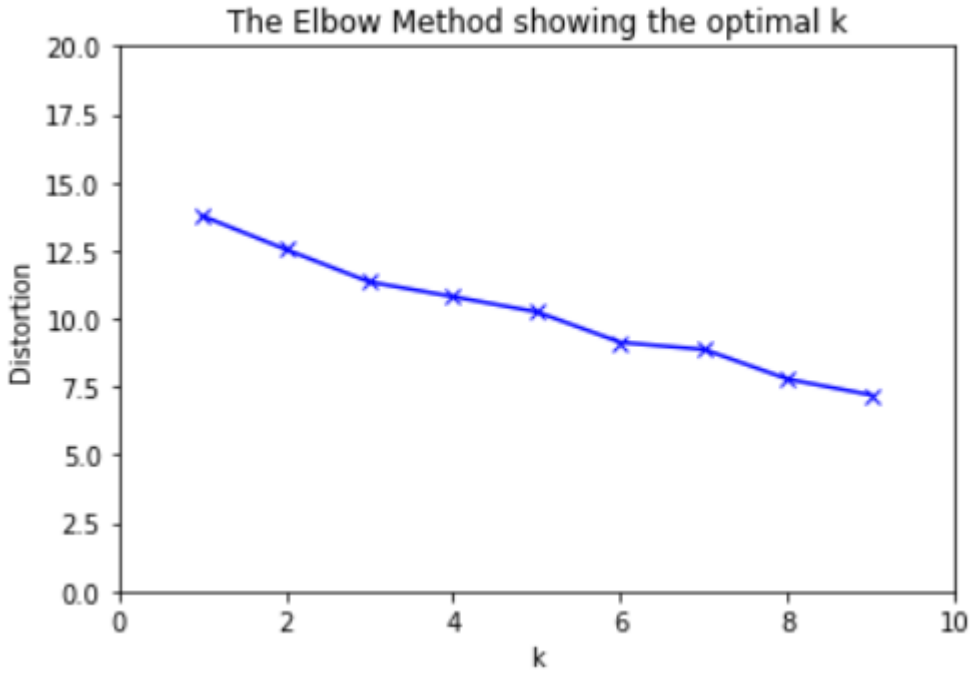
The first step of our methodology is to state clearly the business problem (first paragraph), which is I remind to identify the best district in Paris to implement my new restaurant but also if I want to rent a house, where can I go to fulfil my needs according the local amenities for a good quality life.

Our approach is to collect all places and venues in each district. They will be classified into categories like restaurants, hotel, museum, park, etc … We will limit to the first 150 venues around 1,500 meters from the specified location. The later, for the sake of simplicity, has been positioned at the Hall of the City of each district. The next step is to propose a model to let us extract the main characteristics of each district which can be validated cross-checked with other information (like “More on the 20 Paris districts” in the previous paragraph). Finally, a descriptive analysis of the clustering results let us to respond to the initial questions of the business problem.

Like presented earlier, data needed to achieve the assignment have been collected from different sources. The venues and categories are grabbed from Foursquare using the available standard API converting the addresses of the places to their respective location (latitude and longitude). Google Geocoder tool also has been used to locate Paris it self and the coordinates of the Halls of the city in each district.

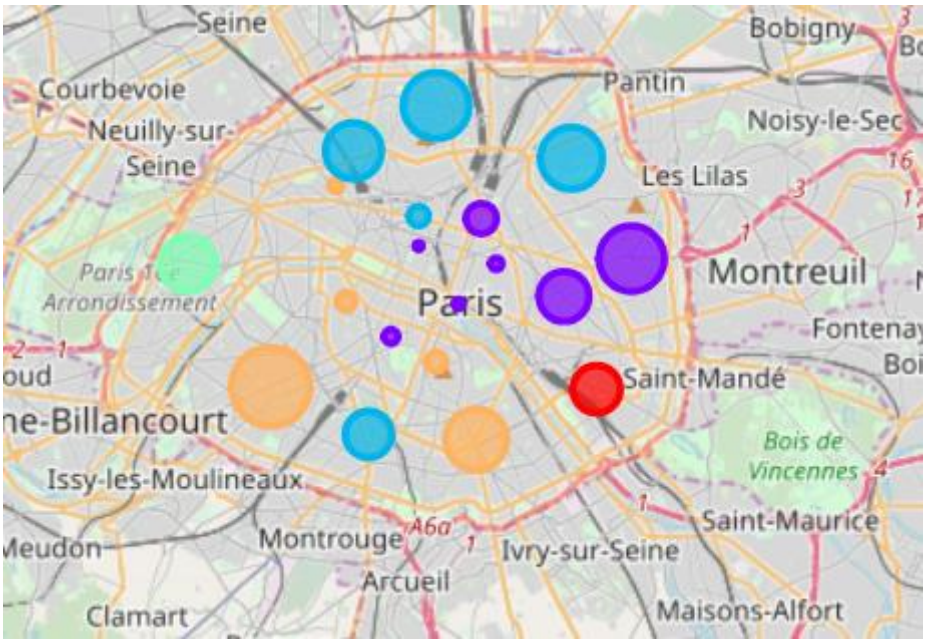
Below is the main steps of the adopted methodology (see details in the notebook):

- Load data into pandas data frames from csv files.
- Use Foursquare API for places geo-localisation.
- Use Google Geocoder to get the coordinated of some places in Paris by districts.
- Use Folium libraries to visualize the places in a map.
- Search the first 150 venues by district at 1500 meters around with their geospatial coordinates¶
- Get the categories of each venue (1,513 for the 20 districts with the previous restriction).
- Group venues by categories (205 unique categories).
- Select the 10 most common venue for each district.
- Use the machine learning algorithm K-Means to segment Paris districts to 5 clusters by categories and the population.
- Use the elbow method for K-Means with best value of K.
- Visualize the results in a map with specific color for each cluster.
- Characterize each cluster from the most common venue and the number of inhabitants.
- Revert back to the initial business problem and discuss.



Results section

The final result of our analysis is to present the 5 groups (clusters) of districts having similar characteristics in terms of existing venues and places. Each district with their respective number of inhabitants is drawn in the following map:



In this map, the color determine the cluster to which the district belongs and the size of the circle shows the number of inhabitants in the district.

The following worksheet resulting from our analysis can help us to respond to our initial questions.

CLUSTER #	Number Of Districts	DISTRICTS	CHARACTERISTICS (Segmentation)
1	**1**	12th – Reuilly	Residential district with many parks/gardens and commodities (Hotel, Restaurants, …)
2	**8**	1st - Louvre 2nd - Bourse 3rd – Temple 4th – Hotel de Ville 6th – Luxembourg 10th - Entrepot 11th - Popincourt 20th - Mesnilmontant	Particularly provided in terms of food (Restaurants, Café, Bar, Bakery, Bistro with the lowest population in Paris.
3	**5**	9th – Opera 14th - Observatoire 17th - Batignoles-Monceau 18th – Buttes-Montmartre 19th – Buttes-Chaumont	Most popular districts in Paris with a lot of hotels and bars (wine) but not so much restaurants.
4	**1**	16th – Passy	Bourgeois population with high density with commodities like café and bakery.
5	**5**	5th – Pathéon 7th – Palais Bourbon 8th - Elysée 13th - Gobelins 15th - Vaugirard	Multicultural and very popular districts with all commodities (Hotel, Bar, Restaurant, café, …). Include universities and touristic places.
	20		

Discussion section

First of all, the results show that for all the districts, "French Restaurant" come as the first or second most important venue unless for the cluster 1 where it comes at the 4th position. This is normal because of the reputed french "gastronomie". Thus, we'll eliminate in the classification this type of venue because it is not discriminatory. Some venues like museums, parks or garden are also seen in most of the districts but we considered them in the analysis.

The results show homogenous clusters in terms of categories of venues but also in terms of number of inhabitants. Indeed, unless the first and the 4th cluster (where there is only one particular district for each, the other clusters make a group of 5 to 8 districts each.

As stated in the results section, the previous worksheet can indicate any user to identify the best cluster fitting with the requirements. A further analysis can be done to select the best district in the cluster. As an example, if we want to open a new restaurant targeting the student community, we should select one of districts in the cluster 5. Obviously, some more analysis with other data should be conducted to fine tune the choice (e.g. considering the cost of rent in each district).

Conclusion section

This report demonstrates the strength and the efficiency of data analysis coupled with the use of machine learning algorithm to solve a concrete business problem as stated previously. Once the later is clearly identified, the collect of data from different sources is a crucial step in the process. The approach in the choosen methodology consists to collect the venues around a location in each administrative division in the area off study. These venues are classified by categories. A machine

learning algorithm named K-Means, reputed to be very efficient has been used to segment and group these division according their similarity in terms of venues around.

The result has been visualized in a map and a worksheet describing each cluster of districts has been provided. These tools can help the users to respond to the initial questions about the business problem.