Capstone Project

The Battle of Neighborhoods

What is the best place to open a Restaurant in Paris, France?

By Vívian Andrade

1. INTRODUCTION

I have been constantly striving to develop my technical skills to become a Data Science. So, I took the IBM course in order to gain knowledge and pursue the IBM Data Science Professional Certification:

https://www.coursera.org/professional-certificates/ibm-data-science.
During this course, I learned how to use Data Science tools, such as Jupyter Notebook, GitHub and IBM Watson Studio. The main programming language used was Python, which is packed with powerful libraries that can be utilised for Data Science such as Pandas, Numpy, Matplotlib, Seaborn, Folium, Scikit-learn and SciPy.

In the final assignment, called “Capstone Project”, it was required to use various tools and methodologies learned throughout this course to solve a real-life business problem. This business problem had to involve the use of location data derived from Foursquare (https://foursquare.com) using API.

2. BUSINESS PROBLEM

Paris is the capital and located in north-central of France. For centuries, Paris has been one of the important city of the world. According to insee, Paris population in 2021 is 2.1 million and the area is 105.4 sq.km. Since the 17th century, Paris has been one of Europe's major centres of finance, diplomacy, commerce, fashion, gastronomy, science and arts.

The city of Paris is divided into twenty arrondissements municipaux, presented in the Fig. 1, administrative districts, more simply referred to as arrondissements. These are not to be confused with departmental arrondissements, which subdivide the larger French départements.

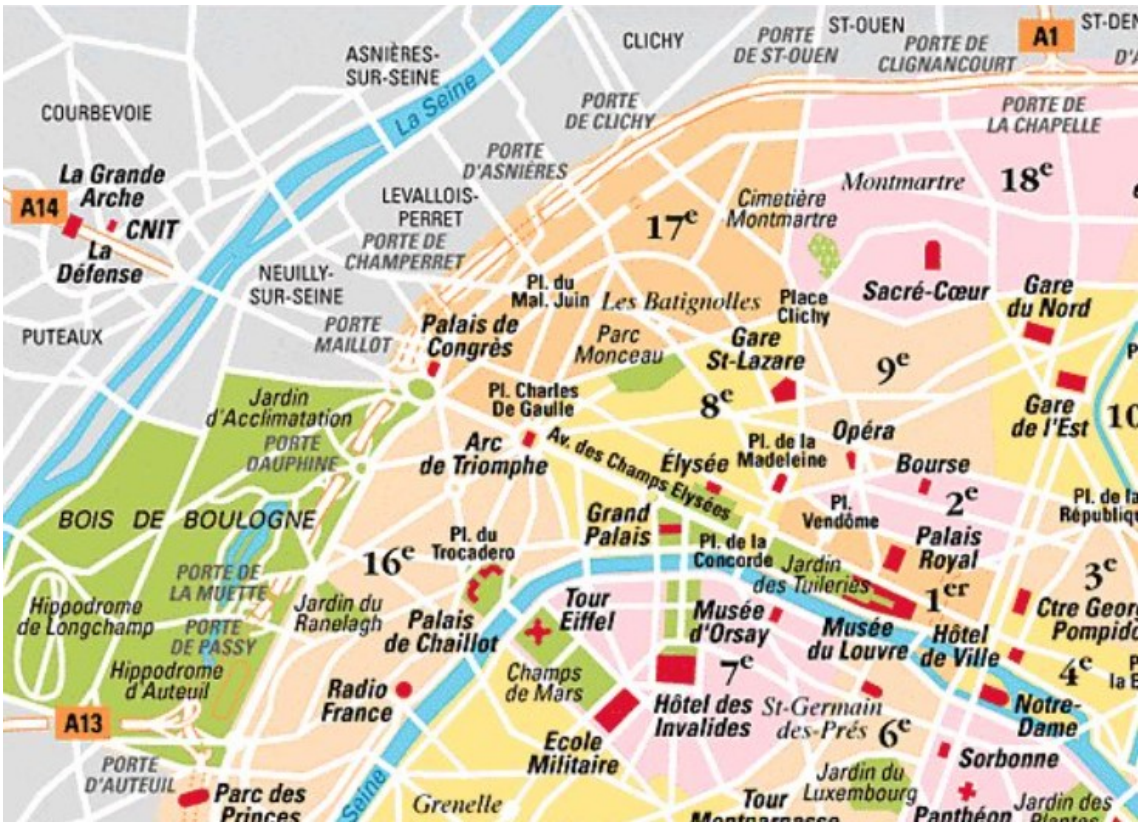

Fig. 1: Arrondissements of Paris

There are currently around 12,000 restaurants in Paris for 2,180 million inhabitants. That is to say, on average, about 1 restaurant per hectare. That's a lot, the competition is tough. Many restaurants are closing, especially with the Covid-19 crisis. For who are think about opening a new restaurant, a market research that indicates the best commercial local to open a new business is very important and can be a big step towards business success.

If someone wants to open a new restaurant in Paris, first, it’s important to conduct a market research. It will help you understand the volume and value of the market, potential customer segments and their buying patterns, the position of your competition, and the overall economic environment, including barriers to entry, and industry regulations.

The main goal of this project is to help these people find the best neighborhood to open a new restaurant. For it, it will be showed the global vision of the

distribution of restaurants and others places that sell foods and beverages in Paris.

3. DATA ACQUISITION

The following list of data will be used to conduct this analysis:

· List of neighborhoods in Central Paris
· Geo-coordinates of the neighborhoods in Paris
· Popular Restaurants by categories, Bakery, Salad Place, Café, Pastry Shop, Pizza Place, Bistro, Coffee Shop, Wine Bar, among other places that sell prepared food and beverages in these neighborhoods in Paris.

The list of neighborhoods will be obtained from the page Open Data / Paris_Data: https://opendata.paris.fr/explore/dataset/quartier_paris/table/?disjunctive.c_ar

The Geo-coordinates will be calculated using the Geopy package within Python.

The Popular Restaurants by categories, Bakery, Salad Place, Café, Pastry Shop, Pizza Place, Bistro, Coffee Shop, Wine Bar, among other places that sell prepared food and beverages list will be gathered from Foursquare using API.

4. METHODOLOGY

4.1 Summary

First, it was used the Paris data to find the geolocation for each neighborhood, then make API calls to Foursquare in order to get the surrounding venue details for each neighborhood. After cleaning this data into a usable format, it was run an unsupervised machine learning algorithm called k-means clustering to group

the neighborhoods based on the restaurant types and others food and beverage business in these neighborhoods.

4.2 Neighborhoods Data

First of all, it was taken some information about the areas of Paris, such as neighborhoods, arrondissement number, population, latitude\longitude.

The data were extracted from Open Data / Paris_Data. The following website was imported as an excel spreadsheet:

https://opendata.paris.fr/explore/dataset/quartier_paris/table/?disjunctive.c_ar

Paris has 20 neighborhoods and the first five dataframe lines are represented below:

	CAR	NAME	NSQAR	CAR.1	CARINSEE	LAR	NSQCO	SUR
0	3	Temple	750000003	3	3	3eme Ardt	750001537	11708
1	19	Buttes-Chaumont	750000019	19	19	19eme Ardt	750001537	67926
2	14	Observatoire	750000014	14	14	14eme Ardt	750001537	56148
3	10	Entrepot	750000010	10	10	10eme Ardt	750001537	28917

Fig. 2: Arrondissements of Paris with some characteristics (only first 5 lines).

After some steps of data cleaning and data preparation, the final result is:

	Arrondissement_Num	Neighborhood	French_N
0	1	Louvre	1e
1	2	Bourse	2eme
2	3	Temple	3eme
3	4	Hotel-de-Ville	4eme
4	5	Pantheon	5eme
5	6	Luxembourg	6eme
6	7	Palais-Bourbon	7eme
7	8	elysee	8eme
8	9	Opera	9eme
9	10	Entrepot	10eme
10	11	Popincourt	11eme
11	12	Reuilly	12eme
12	13	Gobelins	13eme

Fig. 3: Arrondissements/Neighbohoods of Paris.

4.3 Neighborhoods Data

With the coordinates of these 20 main neighbohoods of Paris, the coordinates of Paris were found using Geopy Client Library. After, Folium library was used to visualize geographic details of Paris and marks these areas onto a map to confirm they are accurate. Latitude and longitude values were used to get the visual as below:

Fig. 4: Map of Paris with 20 neighborhoods superimposed on top.

4.4 Foursquare Data

4.4.1 Testing API call for one district

Data from Foursquare were used to view the venues situated within a close radius of each neighbohood. The function was tested for the first neighborhood on the table “Louvre”. By Foursquare API, it was possible to get the top 100 venues that are in Louvre within a radius of 500 meters. The result from the Foursqare API call was a JSON file, which was inspected to create a function in Python to pull out the relevant information and place into a pandas dataframe. Below is the data frame obtained from the JSON file that was returned by Foursquare, after some manipulation and cleaning.

	name	categories	
0	Musée du Louvre	Art Museum	48
1	Palais Royal	Historic Site	48
2	Comédie-Française	Theater	48

Fig. 5: JSON file structured in a pandas dataframe (only first 5 lines).

4.4.2 Extracting information for all districts

The algorithm was scaled to be applied it to all neighborhoods that within a radius of 500 meters of the Louvre. The resulting data frame contains information about venue and venue category. 145 unique venue categories were returned as presented in the following dataframe.

	French_Name	Latitude	Longitude	Venue	Venue Latitude	V
0	1er Ardt	48.862563	2.336443	Musée du Louvre	48.860847	
1	1er Ardt	48.862563	2.336443	Palais Royal	48.863236	
2	1er Ardt	48.862563	2.336443	Comédie-Française	48.863088	
3	1er Ardt	48.862563	2.336443	Cour Napoléon	48.861172	
4	1er Ardt	48.862563	2.336443	La Clef Louvre Paris	48.863977	
...	...	...	...	...	...	
544	20eme Ardt	48.863461	2.401188	Place Édith Piaf	48.866391	

Fig. 6: Venue category and venue for all neighborhood.

Once I was comfortable with the information being returned and I was able to define a function to extract the required details, and apply the same logic to perform the API call for each venue. As I am only interested in restaurants, I extracted only the information related to restaurants, this refined the list to 549 venues of which there were 44 unique categories (restaurant type). I extended the analysis to other different venues related to food and drink represented among all selected venues. These data are useful to understand what is the type of restaurant that has domination in each neighborhood.

Analysis of the top 10 restaurant types in these neighborhoods using the seaborn and matplotlib libraries produced the following graph:

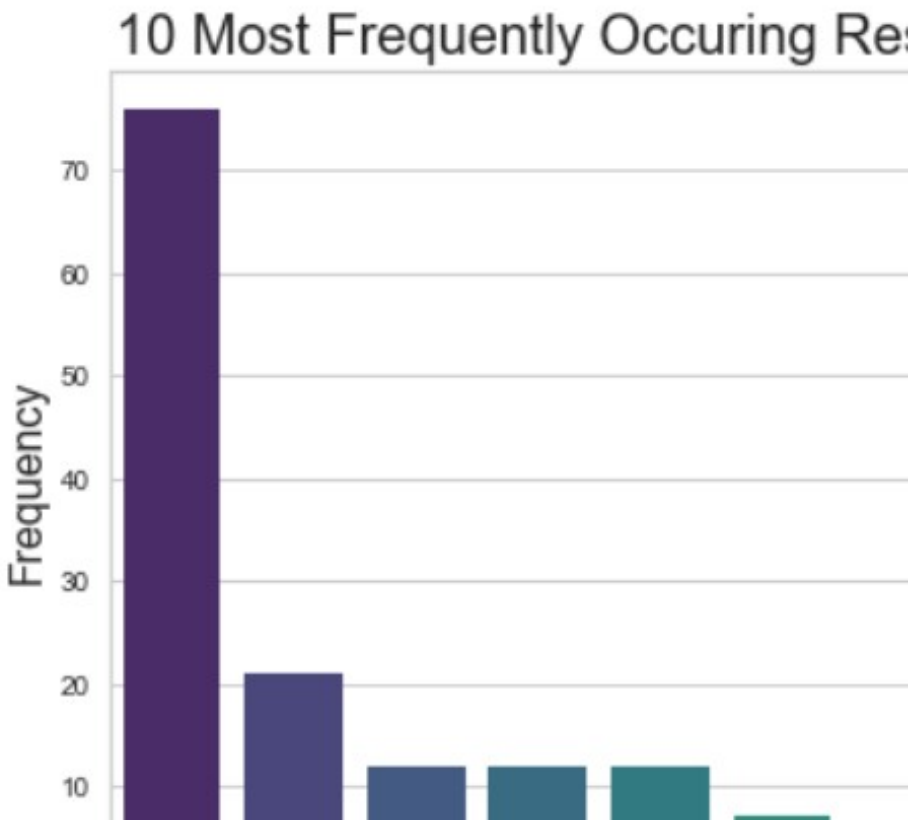

Fig. 7: Top 10 restaurant types in Paris.

In the top 10 restaurant types in Paris, French restaurants are the most popular followed by Italian restaurants. Seaborn and matplotlib were used to compare the number of restaurants in each district.

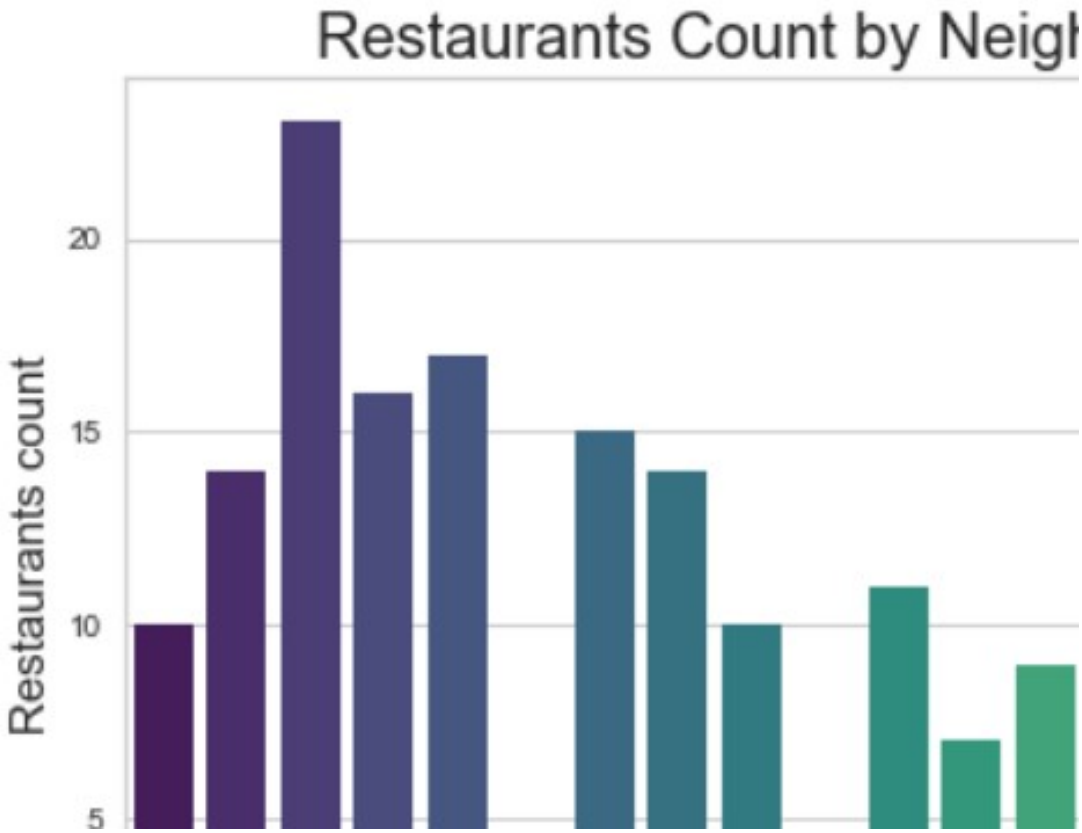

Fig. 8: Number of top-rated restaurants per neighbohood in Paris.

After, other different venues related to food and drink were analysed and represented among all selected venues.

'Drink and Beverages'. Coffee, Wine Bar, Beer Bar, Tea Room, Brasserie and Juice Bar etc.

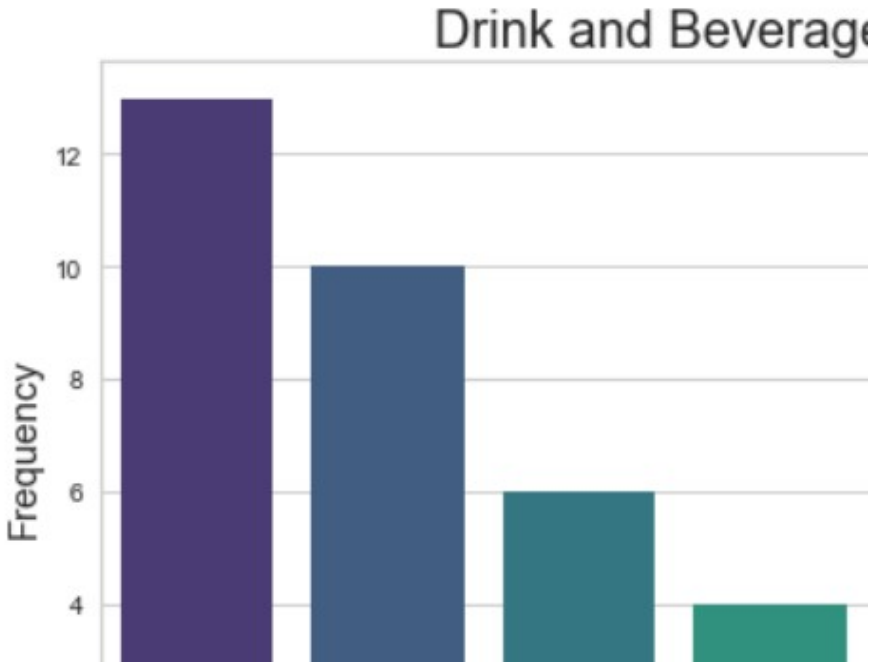

Fig. 9: Number of different venues for food and drink and beverage in Paris.

In the top 10 drinks and beverages types in Paris, Coffee Shop are the most popular followed by Wine Bar.

'Other kind of Foods'. Bakery, Café, Bistro, Pizza Place, Creperie, Ice Cream Shop, Sandwich Place, Pastry Shop and Salad Place.

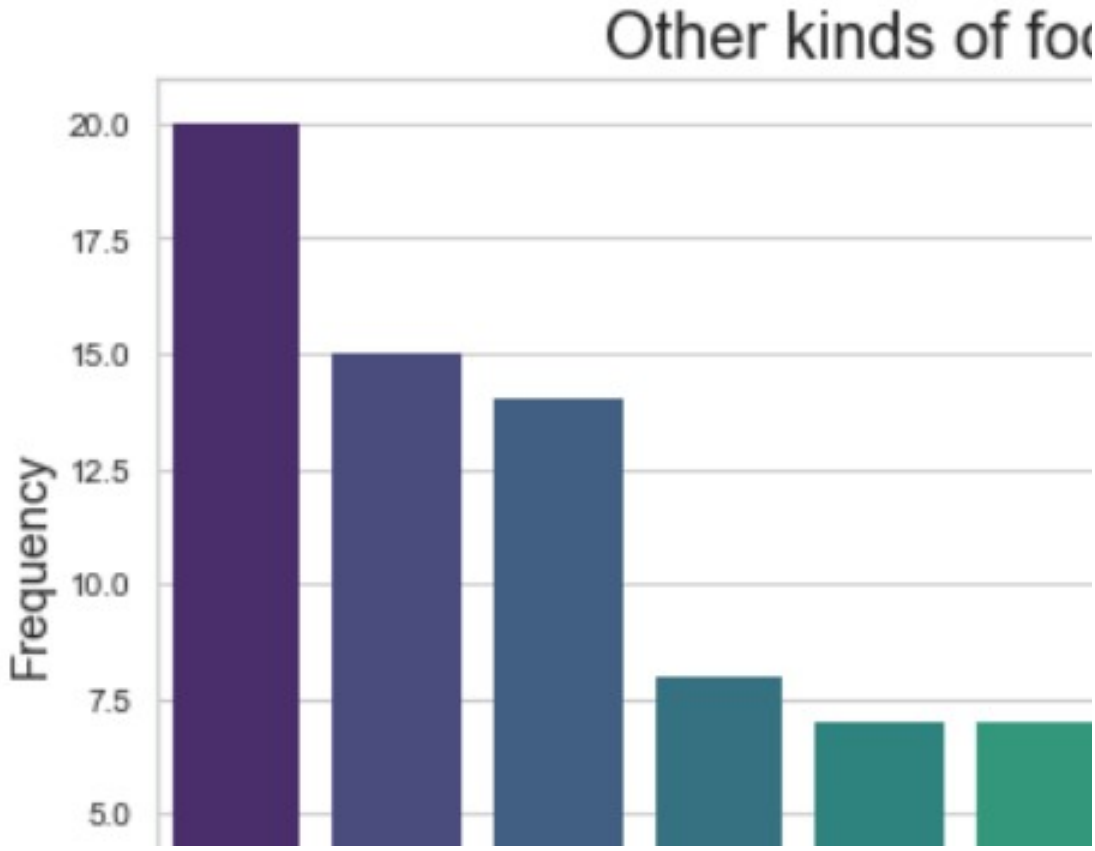

Fig. 10: Number of different Bakery, Café, Bistro, Pizza Place, Creperie, Ice Cream Shop, Sandwich Place, Pastry Shop and Salad Place in Paris.

The Fig. 10 indicates that among the others prepared foods sale points, the most frequent is bakery followed by café.

4.4.3 Analysing each neighborhood

With the extracted information now in a pandas dataframe, one-hot encoding method was performed to convert the categorical values to binary vectors which is required for many machine learning models. Simply put, this transforms the dataframe by placing the unique restaurant types as column headers and the values as either 0 or 1 where 1 is Yes and 0 is No.

	Neighborhood	Afghan Restaurant	African Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Bakery	Bar	Basque Restaurant	...
0	1er Ardt	0	0	0	1	0	0	0	0	0	...
1	1er Ardt	0	0	0	0	0	0	0	0	0	...
2	1er Ardt	0	0	0	0	0	0	0	0	0	...
3	1er Ardt	0	0	0	0	0	0	0	0	0	...

Fig. 11: Dataframe with categorical values converted to binary vectors (only showing top 6 rows)

After the dataframe had been one-hot encoded, the next step was to group the rows by neighborhood in terms of the means of the frequency for each restaurant.

	Neighborhood	Afghan Restaurant	African Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Bakery	Bar	Bas Restau
0	10eme Ardt	0.000000	0.066667	0.0	0.000000	0.0	0.000000	0.033333	0.000000	0.000
1	11eme Ardt	0.033333	0.000000	0.0	0.033333	0.0	0.033333	0.033333	0.033333	0.000
2	12eme Ardt	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.000
3	13eme Ardt	0.000000	0.000000	0.0	0.000000	0.0	0.166667	0.033333	0.000000	0.000

Fig. 12: Dataframe with the mean of the frequency of occurence of each venue category (only showing top 6 rows)

After the dataframe had been one-hot encoded, the next step was to group the rows by neighborhood in terms of the sum of occurrence of each venue category.

	Neighborhood	Afghan Restaurant	African Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Bakery	Bar	Basque Restaurant	...
0	10eme Ardt	0	2	0	0	0	0	1	0	0	...
1	11eme Ardt	1	0	0	1	0	1	1	1	0	...
2	12eme Ardt	0	0	0	0	0	0	0	0	0	...
3	13eme Ardt	0	0	0	0	0	5	1	0	0	...

Fig. 13: Dataframe with the sum of occurence of each venue category (only showing top 6 rows)

The Fig. 14 shows the 10 most common venues in each neighborhood grouped in clusters, which will help to understand briefly the peculiarities of each location.

	Neighborhood	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	
0	10eme Ardt	Coffee Shop	French Restaurant	Bistro	Mediterranean Restaurant	Café	African Restaurant	
1	11eme Ardt	Café	Restaurant	Italian Restaurant	Pastry Shop	Moroccan Restaurant	Ethiopian Restaurant	
2	12eme Ardt	Zoo Exhibit	Zoo	Monument / Landmark	Supermarket	Perfume Shop	Okonomiyaki Restaurant	
3	13eme Ardt	Vietnamese Restaurant	Asian Restaurant	French Restaurant	Thai Restaurant	Chinese Restaurant	Juice Bar	
4	14eme Ardt	French Restaurant	Hotel	Bistro	Food & Drink Shop	Italian Restaurant	Sushi Restaurant	
5	15eme Ardt	Italian Restaurant	Lebanese Restaurant	French Restaurant	Thai Restaurant	Indian Restaurant	Hotel	R
6	16eme Ardt	Lake	Plaza	Bus Station	Art Museum	French Restaurant	Pool	
7	17eme Ardt	Hotel	French Restaurant	Italian Restaurant	Bakery	Restaurant	Turkish Restaurant	Bu
8	18eme Ardt	French Restaurant	Bar	Restaurant	Middle Eastern Restaurant	Deli / Bodega	Pizza Place	R
9	19eme Ardt	French Restaurant	Bar	Seafood Restaurant	Bistro	Beer Bar	Coffee Shop	Bu
10	1er Ardt	Plaza	Café	Coffee Shop	Art Museum	Historic Site	Exhibit	
11	20eme Ardt	Bakery	French Restaurant	Japanese Restaurant	Italian Restaurant	Plaza	Korean Restaurant	
12	2eme Ardt	Cocktail Bar	Coffee Shop	Hotel	Pizza Place	Perfume Shop	Donut Shop	

Fig. 14: Neighborhood grouped in clusters with the top 10 venue categories for each neighborhood.

This table provides very valuable information and can be very useful when making a decision regarding the new business localization.

4.4.4 K-means Clustering

A type of unsupervised learning, K-means clustering, which is used when you have unlabeled data (i.e., data without defined categories or groups) was used to analyze which neighborhood of Paris is good to open a new restaurant. The goal of this algorithm is to find groups in the data, with the number of groups

represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. For it, the first step was to identify *the best "K" using the elbow method.*

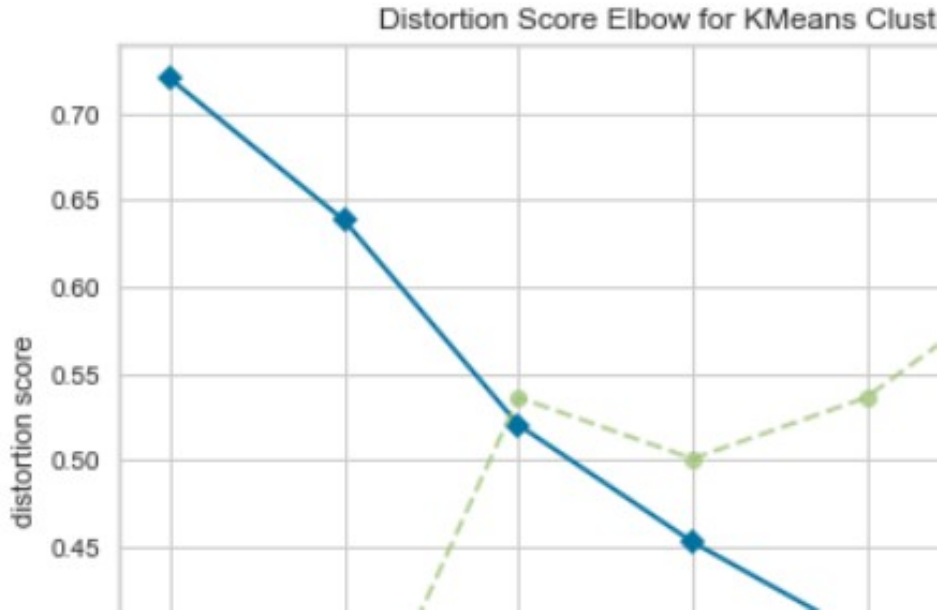

Fig. 15: Distortion score elbow for Kmeans clustering

No 'knee' or 'elbow point' detected This could be due to bad clustering, no actual clusters being formed etc. I settled on the option with 5 clusters.

Finally, we can try to cluster the neighborhood based on the venue categories and use K-Means clustering. The 5 clusters are partitioned based on similar categories venues that belong to neighborhoods. A plot was made to visualize all those results on the Paris map.

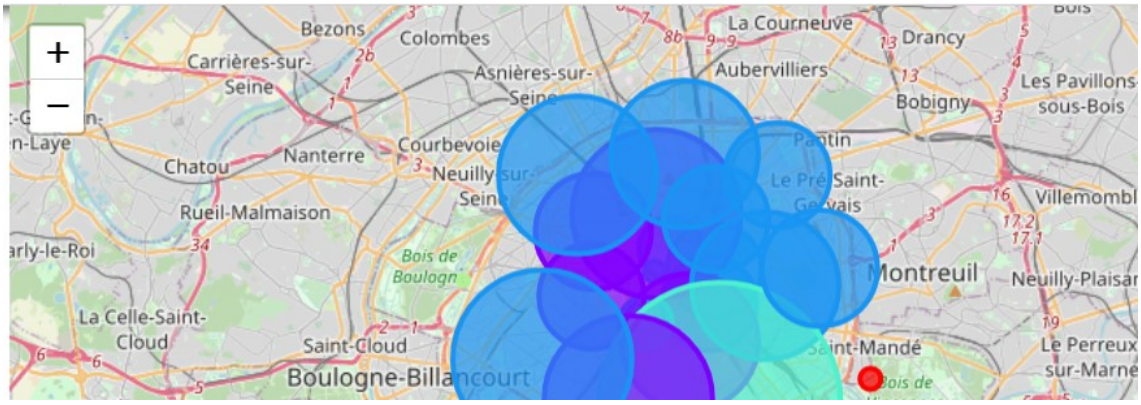

Fig. 16: Visualization the clusters on the Paris map.

The Fig. 16 showed the following clusters:

– Cluster 0 : Arrondissement_Num 12 (Red)
– Cluster 1 : Arrondissement_Num 5, 7, 8, 9 e 14 (Purple)
– Cluster 2 : Arrondissement_Num 1, 2 , 3 , 4, 6, 10, 11, 15, 17, 18, 19 e 20 (Blue)
– Cluster 3 : Arrondissement_Num 13 (Green)
– Cluster 4: Arrondissement_Num 16 (Orange)

Arrondissement_Num	Neighborhood	French_Name	Latitude	Longitude	Cluster	1th Most Common Venue	2th Most Common Venue
1	Louvre	1er Ardt	48.862563	2.336443	2	Plaza	Café
2	Bourse	2eme Ardt	48.868279	2.342803	2	Cocktail Bar	Coffee Shop
3	Temple	3eme Ardt	48.862872	2.360001	2	Hotel	Sandwich Place
4	Hotel-de-Ville	4eme Ardt	48.854341	2.357630	2	Ice Cream Shop	French Restaurant
5	Pantheon	5eme Ardt	48.844443	2.350715	1	French Restaurant	Plaza
6	Luxembourg	6eme Ardt	48.849130	2.332898	2	Pastry Shop	Bakery
7	Palais-Bourbon	7eme Ardt	48.856174	2.312188	1	French Restaurant	Hotel
8	elysee	8eme Ardt	48.872721	2.312554	1	French Restaurant	Hotel
9	Opera	9eme Ardt	48.877164	2.337458	1	French Restaurant	Bakery
10	Entrepot	10eme Ardt	48.876130	2.360728	2	Coffee Shop	French Restaurant
11	Popincourt	11eme Ardt	48.859059	2.380058	2	Café	Restaurant
12	Reuilly	12eme Ardt	48.834974	2.421325	0	Zoo Exhibit	Zoo
13	Gobelins	13eme Ardt	48.828388	2.362272	3	Vietnamese Restaurant	Asian Restaurant

Fig. 17: Neighborhood grouped in clusters.

As showed in the Fig. 17, some districts have characteristic features. Some are dominated by classic restaurants, others, are dominated by non-French restaurants (Italian, Indian and Thai cuisines). The characteristics we obtained as a result of a clear analysis emphasize and confirm the obvious

characteristics of the central areas in which there are many points of interest and in which institutions focus mainly on tourists and the tourism industry.

5. RESULTS

K-means cluster our data into 5 clusters and return those items.

5.1 Cluster 0

	Neighborhood	Cluster	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue

5.2 Cluster 1

	Neighborhood	Cluster	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
4	Pantheon	1	French Restaurant	Plaza	Creperie	Bar	Hookah Bar	Hotel
6	Palais-Bourbon	1	French Restaurant	Hotel	Plaza	History Museum	Garden	Cocktail Bar
7	elysee	1	French	Hotel	Art Gallery	Bakery	Spa	Hotel Bar

5.3 Cluster 2

	Neighborhood	Cluster	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Louvre	2	Plaza	Café	Coffee Shop	Art Museum	Historic Site	Exhibit
1	Bourse	2	Cocktail Bar	Coffee Shop	Hotel	Pizza Place	Perfume Shop	Donut Shop
2	Temple	2	Hotel	Sandwich Place	Dessert Shop	Wine Bar	Burger Joint	Farmers Market
3	Hotel-de-Ville	2	Ice Cream Shop	French Restaurant	Coffee Shop	Wine Bar	Italian Restaurant	Hotel
5	Luxembourg	2	Pastry Shop	Bakery	French Restaurant	Fountain	Dessert Shop	Tennis Court
9	Entrepot	2	Coffee Shop	French Restaurant	Bistro	Mediterranean Restaurant	Café	African Restaurant
10	Popincourt	2	Café	Restaurant	Italian Restaurant	Pastry Shop	Moroccan Restaurant	Ethiopian Restaurant
14	Vaugirard	2	Italian Restaurant	Lebanese Restaurant	French Restaurant	Thai Restaurant	Indian Restaurant	Hotel

5.4 Cluster 3

Neighborhood	Cluster	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue

5.5 Cluster 4

Neighborhood	Cluster	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue

This data analysis shows that each neighborhood has its particularity. Some are dominated by French Restaurant such as, Pantheon, Buttes-Montmartre and Opera, others, as Vaugirard, are dominated by Italian Cuisine.

The region called "Cluster 1" has French restaurants as 1th most commum venue and it is surrounded by good infrastructure including squares, parks, hotels, etc. The characteristics these regions confirm that they are tourist spots and can be interesting for opening restaurants.

The choice of location will depend on the type of restaurant or other kind of food and beverage that the entrepreneur wishes to open. The tables presented important data for decision making. The interested can analyze the frequency of establishments by neighoborhood, as well as all the tourist and commercial attractions in the surroundings. Thus, it is up to the executive to define whether he prefers a region with tourist attractions and a low supply of restaurants, or a region with the same or similar tourist attractions and more establishments (same features of the business you want to open). This last feature indicates that despite the high competition, demand in the region tends to be high as the region already has a defined public profile.

The analysis shows that there are areas where there is a balanced number of restaurants, coffees and other kinds of food and beverages shops. The result emphasizes the actual and general characteristics of the districts in the clusters.

The opening of a restaurant in this cluster 1 and 2 are quite reasonable. The 13eme Ardt (Gobelins) has the highest number of top-rated restaurants as showed in the Fig. 8, therefore, I believe that this region presents a greater risk due to high competition.

The infrastructure of the neighborhoods already meets the needs of people for food and leisure. People are already considering these areas for lunch, dinner, meetings and evening rest. Any venue that opens in these areas will benefit from the status of the place and the habits of the people.

This analysis within this project is quite superficial, it shows the basic methods and opportunities. But, from this study, the stakeholders can refine your search criteria and improve your analysis for specific businesses. I hope this preliminary analysis will be helpful in your decision making. The clustering has revealed characteristic groups of areas on which it is possible to concentrate more specifically. So, for further consideration, I would choose the clusters 1 and 2.

6. CONCLUSION

The study was performed on small set of data, therefore it is possible achieve better results by increasing the neighborhood information. Anyway, Paris is a capital with many different kinds of new restaurant and food and beverage business to offer and we have gone through the process of identifying the business problem, specifying the data required, clean the datasets, performing a machine learning algorithm using k-means clustering and providing some useful tips to our stakeholder. For the Best performance, you can use Foursquare API to major regions of Paris and their neighborhoods. The potential for this kind of analysis in a real-life business problem were discussed in detail. As a final note, all analysis presented in this work carried out with data from Four Square and depends on the adequacy and accuracy your data. As future work, I recommend collecting more external data sources that allow us to

analyze the best location of the business, such as pedestrian flow, per-capita income in the region, number of tourists per day.

.

7 REFERENCES

- Forsquare API https://developer.foursquare.com/
- Coursera https://www.coursera.org/professional-certificates/ibm-data-Science
- https://opendata.paris.fr/explore/dataset/quartier_paris/table/?disjunctive.c_ar
- https://dicasdeparis.net/os-arrondissements-de-paris/