Mathis Doulson
31/05/2021
Paris

The Battle Of Neighborhoods

Indian Restaurant in Paris

I. Introduction

1.1. Background

This is the capstone project of IBM Data Science Professional Certificate. The exercise imposes using the Foursquare API in order to cluster different areas of a place in the world according to famous venues in this place. I will use the Foursquare location data to explore neighborhoods in Paris, and come up with a problem for which I can use the Foursquare API.

1.2. Problem

I will explore the neighborhoods in Paris and answer the question: "Where is the appropriate place to open an Indian Restaurant in Paris ?". The business owner wants to ensure a steady and sustainable business. We therefore need to meet the following requirements :

- The store needs to be strategically located inside a very dense area, demographically speaking
- Confirm any assumption by means of modeling and testing the data. Specifically, visually cluster common restaurants in Paris by neighborhood.
    - Locating the new restaurant according to these requirements will ensure the following:
        - lowest cost for delivery
        - shortest travel time to his store for his clients
        - overall lower run costs
        - overall greater customer satisfaction
- Additionally, determine that a good number of people can frequent these restaurants with sustainable frequency inside these neighborhoods.
- The restaurant has to be set in a ‘world restaurants’ area, as it is an Indian restaurant

1.3. Interest

Paris is one of the biggest international cities in the world, and one of the most touristic cities. Opening a restaurant here is an attractive idea. Nevertheless, Indian restaurant isn’t really what tourists are attracted by in the first place when going out in Paris. Are some quarters more suitable for setting an Indian restaurant than others? This is the question we are going to deal with, by exploring different clustering of Paris based on the most famous venues recorded on Foursquare Website.

II. Data

A. Data Sources

Two different kind of data is needed for the comparison.

- **City quarters and respective geographical data:** in order to analyze the cities on a meaningful level, they need to be divided into different areas, in our case in *quartiers (*subdivisions of the 20 *Arrondissements* constituting of Paris. I was able to find the list of the 80 *quartiers* of Paris on Wikipedia[1]. I then web-scrapped the HTML page in order to convert this list into a data frame usable with pandas.
  Using the *Geocoder* python library, I was able to get the geographical coordinates for each quarter.

- **Venue data:** This data, including the Venue name, its category, latitude and longitude, is gathered using the Foursquare API[2].

B. Data Cleaning

1. Obtaining the list of quarters, density and geographical coordinates.

After scraping the table of different *quartiers* of Paris from Wikipedia as well as their according density, I had to get rid of useless columns. I then renamed the different columns and cleaned the display of *arrondissements* as for having only the numbers of the *arrondissement*. From there, I was able to get for each quarter the precise geographical coordinates via the Geocoder API. Here is the final pandas data frame :

	Arrondissement	Quartier	Latitude	Longitude	Densité
0	1	Paris Saint-Germain-l'Auxerrois	48.860211	2.336299	1924
1	1	Paris Halles	48.864614	2.334396	21806
2	1	Paris Palais-Royal	48.864639	2.335815	11661
3	1	Paris Place-Vendôme	48.867463	2.329428	11316
4	2	Paris Gaillon	48.869135	2.332909	7154

2. Obtaining for each quarter, the most famous venues

A rich collection of features was selected from the Foursquare API as follows :

- Venue Name
- Venue Latitude

---

[1] https://en.wikipedia.org/wiki/Quarters_of_Paris

[2] Foursquare City Guide, commonly known as Foursquare, is a local search-and-discovery mobile app developed by Foursquare Labs Inc. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history.

- Venue Longitude
- Venue Category

	Quartier	Quartier Latitude	Quartier Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Paris Saint-Germain-l'Auxerrois	48.860211	2.336299	Cour Carrée du Louvre	48.860360	2.338543	Pedestrian Plaza
1	Paris Saint-Germain-l'Auxerrois	48.860211	2.336299	La Vénus de Milo (Vénus de Milo)	48.859943	2.337234	Exhibit
2	Paris Saint-Germain-l'Auxerrois	48.860211	2.336299	Musée du Louvre	48.860847	2.336440	Art Museum
3	Paris Saint-Germain-l'Auxerrois	48.860211	2.336299	Cour Napoléon	48.861172	2.335088	Plaza
4	Paris Saint-Germain-l'Auxerrois	48.860211	2.336299	Pont des Arts	48.858565	2.337635	Bridge

	Quartier	Accessories Store	African Restaurant	Alsatian Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Venezuelan Restaurant	Video Game Store	Vietnamese Restaurant	Wa Sl
0	Paris Archives	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	
1	Paris Arsenal	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	
2	Paris Arts-et-Métiers	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.041667	
3	Paris Auteuil	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	

<u>Pre-processing the data :</u>

The 253 unique venue categories were converted into categorical (more precisely binary) variables, using one-hot-encoding in order to perform the K-means algorithm. Once each category was transformed into dummy variables, I was able to group rows by neighborhood and compute the mean of the frequency of occurrence of each category :

Out[24]:

	Quartier	Accessories Store	African Restaurant	Alsatian Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Venezuelan Restaurant	Video Game Store	Vietnamese Restaurant	Watch Shop	Win Ba
0	Paris Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	Paris Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	Paris Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	1	0	...	0	0	0	0	
3	Paris Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0		0	0	0	0	
4	Paris Saint-Germain-l'Auxerrois	0	0	0	0	0	0	0	0	0		0	0	0	0	

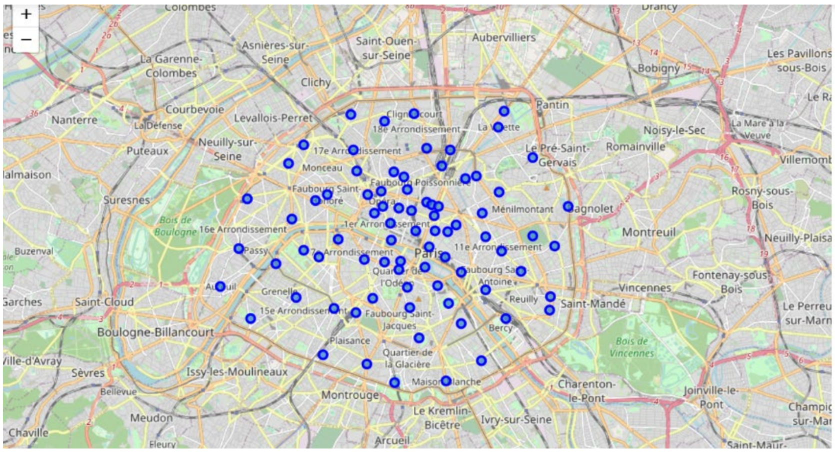
5 rows × 264 columns

III. Methodology

A. Exploratory Data Analysis

1. Geolocalisation of each quarter

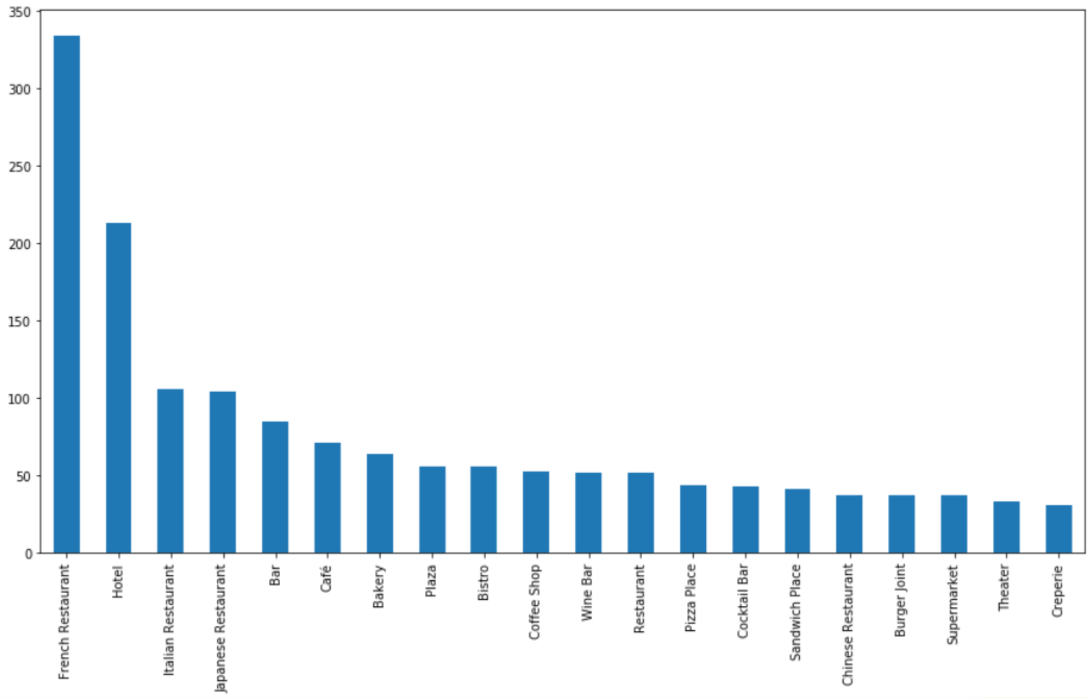
Once I obtained geographical coordinates for each quarter, I was able to display each of them of a map, to make sure all the coordinates were actually situated in Paris. I used Folium from Leaflet API, an open-source JavaScript library for interactive maps.



2. Venues Categories

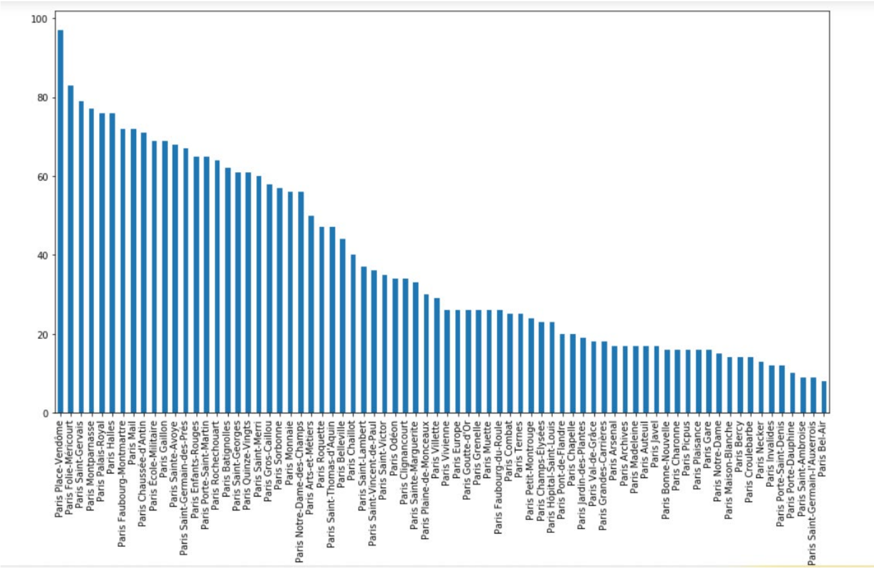
Now that I had associated quarters with all the relevant venues, I was able to determine the number of distinct venues categories. I observed 253 venues categories that have been raised on the Foursquare API for the whole city of Paris. I plotted the 20 most frequent venues categories with their frequency of appearance in Paris :

```
: <matplotlib.axes._subplots.AxesSubplot at 0x1b40a08b5f8>
```



In first place, it can be observed that restoration industry is very prominent in term of frequentation in the city of Paris. French, Italian and Japanese Restaurants can be found in the top 5 of the most frequented venues categories. Cafés, Wine bars and other restaurants are in the top 20.

3. Frequency of venues per quarter



As it can be seen above, the number of different venues per quarter is unequally reported on the foursquare API. That's why whe had to normalize the data, and computing the mean of frequency (cf. *pre-processing* page 3).

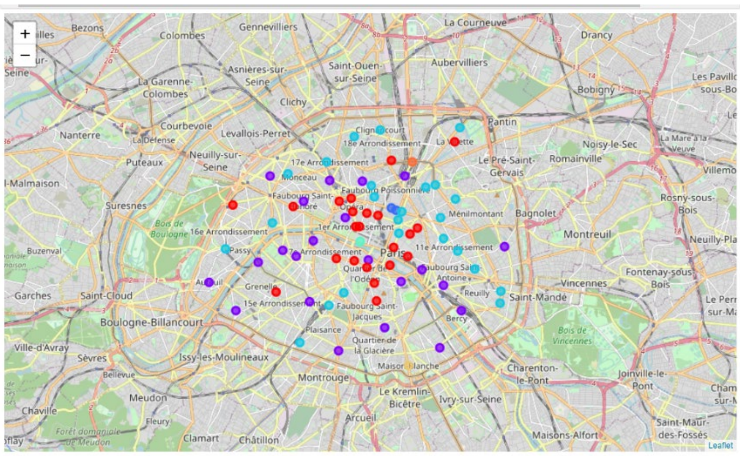
B. Clustering

For this project, k-means is an appropriate clustering algorithm. Because we have a unlabelled dataset, so this is an unsupervisied learning project. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. By clustering the neighborhoods, we can find out the pattern in them, identify the identical neighborhoods and see which is our target. One difficulty of k-means is to determine the hyperparameter k. Based on the *inertia loss* indicator, I chose k=8 for my optimum number of clusters.

IV. Results

A. Geolocalisation of the clusters

After running the K-means algorithm on my pre-processed data frame, I was able to display each cluster on the map, as such :



B. Examine each cluster

In order to determine which cluster would suit the most for my client, I decided to display for each cluster the 8 most common venues for each quarter of the cluster.
For instance, let's have a look at the 2nd cluster :

	Quartier	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
3	Paris Place-Vendôme	Hotel	French Restaurant	Jewelry Store	Hotel Bar	Japanese Restaurant	Italian Restaurant	Cocktail Bar	Boutique
14	Paris Arsenal	French Restaurant	Cocktail Bar	Gourmet Shop	Park	Gym	Supermarket	Vegetarian / Vegan Restaurant	Tapas Restaurant
16	Paris Saint-Victor	French Restaurant	Hotel	Café	Plaza	Wine Bar	Szechuan Restaurant	Syrian Restaurant	Bistro
20	Paris Monnaie	French Restaurant	Hotel	Ice Cream Shop	Plaza	Seafood Restaurant	Japanese Restaurant	Museum	Sandwich Place
25	Paris Invalides	French Restaurant	Plaza	Tea Room	Train Station	Hotel	Cultural Center	Smoke Shop	Embassy / Consulate
26	Paris École-Militaire	Hotel	French Restaurant	Italian Restaurant	Coffee Shop	Café	Japanese Restaurant	Ice Cream Shop	Food & Drink Shop
27	Paris Gros-Caillou	French Restaurant	Italian Restaurant	Hotel	Bakery	Bistro	Café	Convenience Store	Romanian Restaurant
29	Paris Faubourg-du-Roule	French Restaurant	Hotel	Cocktail Bar	Art Gallery	Grocery Store	Resort	Shoe Store	Café
31	Paris Europe	Hotel	Italian Restaurant	French Restaurant	Café	Thai Restaurant	Theater	Korean Restaurant	Music Store

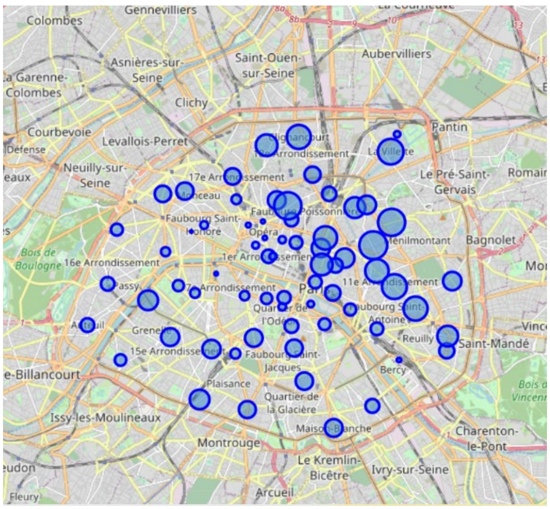
This cluster is mainly composed of French Restaurants and Hotels. Then a large number of Cafés, Bars and international restaurants can be found.

I decided to make a Wordcloud for each cluster in order to visualize more easily the essence of each cluster. Here is an example for clusters 4, 5 and 6.

With a closer look at each cluster, it can be observed that Cluster 1, 2, 4 and 7 match our requirements (high frequency of international restaurants). Taking into account those 4 clusters out of the 8 existing, we are now dealing with 48 quarters.

C. Visualizing density for each relevant quarter

Now that we've found our clusters, it's now possible to merge each quarter with their demographical density scrapped on Wikipedia. I displayed the density of each quartier of the relevant cluster on the map below :



D. Answer the question

From this map, we can observe that quarters of interest in terms of density are essentially situated on the right side of the *Seine*. The business owner may want to choose to set his future restaurant in one of the dense quarters of the *Rive Droite* (right side of the *Seine*). Those quarters taken from relevant clusters are all characterized by high frequency of international restaurants.

V. Discussion

The assignment imposed the use of Foursquare in order to determine the most common venues for each cluster in order to cluster the map by common venues. The sole use of Foursquare is questionable. Indeed, Foursquare users may not be very representative of the whole frequentation in Paris. Biases can exist, for instance people using Foursquare are likely to be young people, a sociologically distinctive cohort whose tastes and choices are very different from adults and older people who are unlikely to report their visits on the social media.

As far as the algorithm is concerned, we obtained quite distinctive cluster which is good although clusters one and two were quite similar though. It is interesting to raise that cluster 7 is solely constituted by the quarter *Maison Blanche* which hosts quasi exclusively *Asian Restaurants*.

VI. Conclusion

In this project, I had to use the location data from Foursquare to solve the problem " Where is the appropriate place to open an Indian Restaurant in Paris ?". I collected the quarters data from the wikipedia page, and formatted it as to be able to get the coordinates for each of them through *geocoder* API. The I was able to invoke Foursquare's API to get the frequented venues for each quarter. After pre-processing the data, I used K-Means algorithm to determine which clusters were the most likely to suit my client. Crossing the relevant clusters with demographical density data, I determined which quarters were the most suitable.