

**IBM Data Science Capstone Project**

**The battle for restaurants in the Paris neighborhoods**

Philippe BOTTIER | Applied Data Science Capstone by IBM | May 19, 2021

# **Introduction**

## Description and discussion of the context

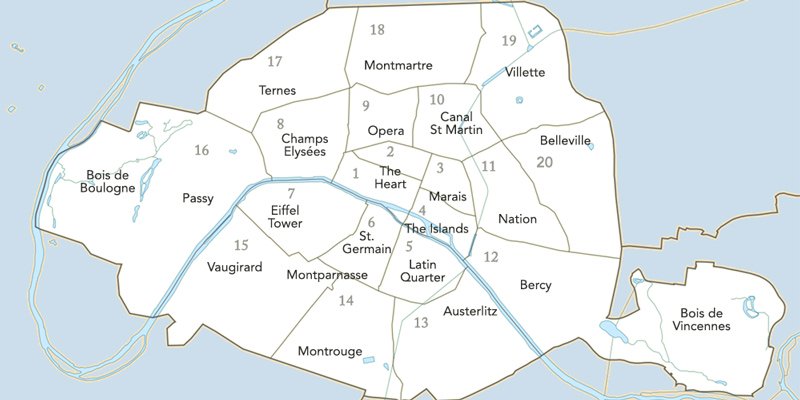
Paris is the [capital](https://en.wikipedia.org/wiki/Capital_city) and [most populous city](https://en.wikipedia.org/wiki/List_of_communes_in_France_with_over_20,000_inhabitants" \o "List of communes in France with over 20,000 inhabitants) of [France](https://en.wikipedia.org/wiki/France), with an estimated [population](https://en.wikipedia.org/wiki/Population) of 2,175,601 residents as of 2018, in an area of more than 105 square kilometres (41 square miles). Since the 17th century, Paris has been one of Europe's major centres of [finance](https://en.wikipedia.org/wiki/Finance), [diplomacy](https://en.wikipedia.org/wiki/Diplomacy" \o "Diplomacy), [commerce](https://en.wikipedia.org/wiki/Commerce), [fashion](https://en.wikipedia.org/wiki/Fashion" \o "Fashion), [science](https://en.wikipedia.org/wiki/Science), [arts](https://en.wikipedia.org/wiki/Arts) and [gastronomy](https://en.wikipedia.org/wiki/Gastronomy)

Paris received 38 million visitors in 2019, measured by hotel stays, with the largest numbers of foreign visitors coming from the United States, the United Kingdom, Germany, and China. It was ranked as the second most visited travel destination in the world in 2019, after [Bangkok](https://en.wikipedia.org/wiki/Bangkok) and just ahead of London.

The city of [Paris](https://en.wikipedia.org/wiki/Paris) is divided into twenty administrative districts called arrondissements. The twenty arrondissements are arranged in the form of a [clockwise](https://en.wikipedia.org/wiki/Clockwise" \o "Clockwise) [spiral](https://en.wikipedia.org/wiki/Spiral) (often likened to a [snail shell](https://en.wikipedia.org/wiki/Gastropod_shell" \l "Morphology" \o "Gastropod shell)), starting from the middle of the city, with the first on the [Right Bank](https://en.wikipedia.org/wiki/Rive_Droite) (north bank) of the [Seine](https://en.wikipedia.org/wiki/Seine).

In French, notably on street signs, the number is often given in Roman numerals. For example, the [Eiffel Tower](https://en.wikipedia.org/wiki/Eiffel_Tower) belongs to the [VIIe arrondissement](https://en.wikipedia.org/wiki/7th_arrondissement_of_Paris) while [Gare de l'Est](https://en.wikipedia.org/wiki/Gare_de_l%27Est_(Paris_M%C3%A9tro)) is in the [Xe arrondissement.](https://en.wikipedia.org/wiki/10th_arrondissement_of_Paris)

Each Parisian arrondissement has four neighborhoods, so a total of 80, which constitute the highest level of public administration in Paris.

[](https://www.parisinsidersguide.com/paris-arrondissements.html)

As a resident in a suburb near the capital, I chose the city of Paris to lead my project.

## Business problem

The problem to be solved is totally imaginary.

My company, PhB Data Consulting, specialized in data analysis, was contacted by the Tourist Office of the city of Paris to provide them with an analysis that would allow them to advise tourists on the types of restaurants they could find during their visits to the Paris neighborhoods.

This analysis could be carried out using an unsupervised learning model that would reproduce on the map of Paris the groupings of neighborhoods according to the types of restaurants that are most represented there.

## **Data Gathering**

To carry out this project, I needed the following data :

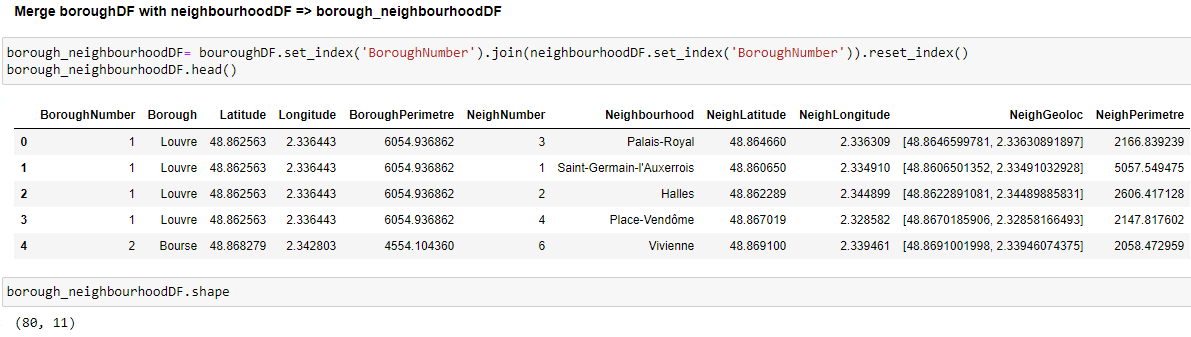
The list of Parisian districts (arrondissements in French) as well as the list of Parisian neighborhoods (quartiers). These lists were imported from the Open Data site of the city of Paris :

* <https://opendata.paris.fr/explore/dataset/arrondissements/export/?disjunctive.c_ar&disjunctive.c_arinsee&disjunctive.l_ar>
* <https://opendata.paris.fr/explore/dataset/quartier_paris/export/>

The Geo-coordinates of the districts in Paris, obtained with the help of the geocoder tool in the notebook.

The Top venues data of neighborhoods, obtained from Foursquare through an API

After cleaning, preparing and merging the data, here is the dataframe that was used for the rest of the project. In this dataframe, we therefore have the 20 arrondissements of Paris with for each of them, their name, their geographical coordinates as well as the four neighbourhoods which depend of the arrondissement.



# **Methodology**

To conduct this study, I list below the different steps of the methodology that I followed :

* Step 1 : Data Acquisition with the JSON library
* Step 2 : Preparing, Cleaning and Merging Data with the PANDA library
* Step 3 : Venues Acquisiton with the Foursquare API
* Step 4 : One Hot Encoding with the PANDA library
* Step 5 : Clustering K-means with sklearn.cluster library
* Step 6 : Cluster Analysis
* Step 7 : Creation of a clusters map with the Folium library

## Step 1 - Data Acquisition with the JSON library

As I said before, I have retrieved the data from the Open Data site of the city of Paris (<https://opendata.paris.fr/page/home/>) :

* list of Paris arrondissements : <https://opendata.paris.fr/explore/dataset/arrondissements/export/>
* list of Paris neighbourhoods :

<https://opendata.paris.fr/explore/dataset/quartier_paris/export/>

These files are in GeoJSON format. To be able to use the data, I have saved the files in the same directory as my Python notebook.

To import the files, I used the json library :

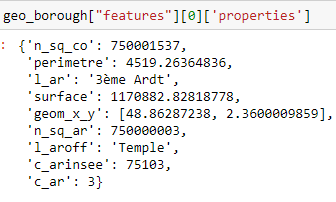
|  |
| --- |
| import json  geo\_borough **=** json.load(open("arrondissements.geojson")) *# Paris arrondissements*  geo\_neighbourhood **=** json.load(open("quartier\_paris.geojson")) *# Paris neighborhoods* |

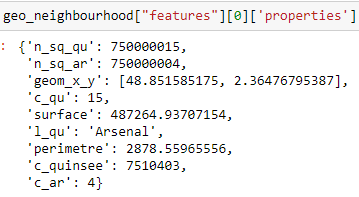
The GeoJSON format is used to represent data of a geographic type. In this object, transformed into a dictionary under python, there are two elements: the type and the information (named features).

In each object of the features list, we also have different objects types :

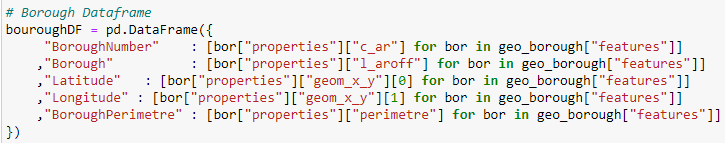


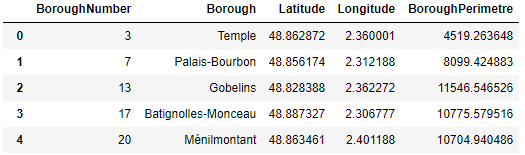
In the properties field, there are various useful information, including the longitude and latitude coordinates of the center of the borough (or of the neighbourhood).

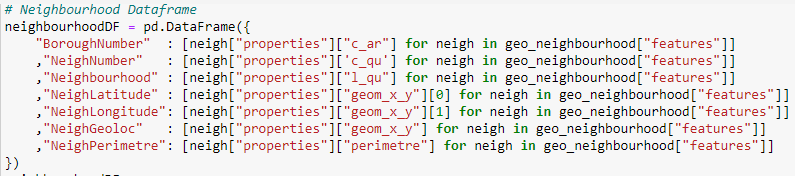


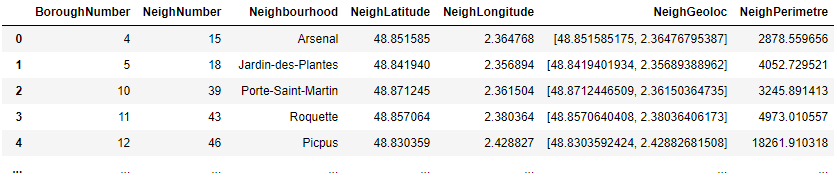


For the rest of the project, with the Panda library imported previously, I have created 2 DataFrames with the useful information.



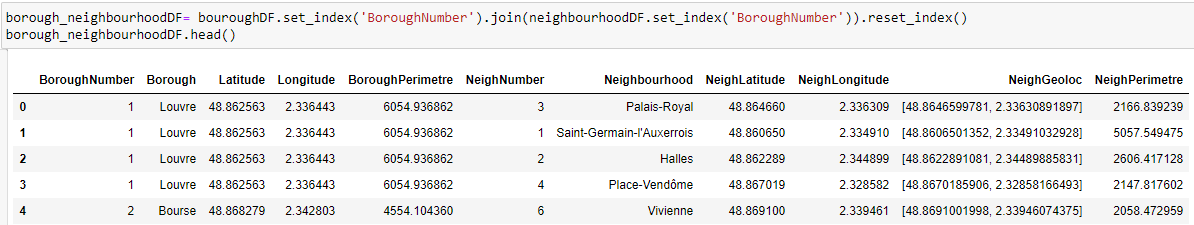


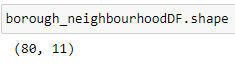




## Step 2 - Preparing, Cleaning and Merging Data with the PANDA library

After cleaning, merging, combining the values ​​of the data frame, I obtain the following data frame which will allow me to continue the study :





In this dataframe we have 80 Paris Neighbourhoods.

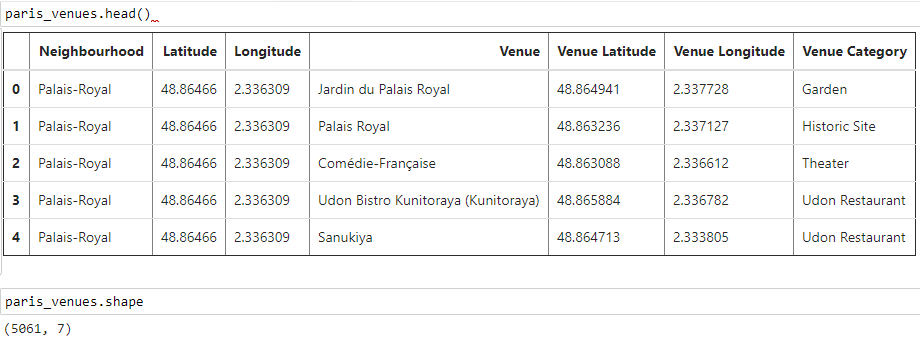
I used this dataframe for the rest of the project. I didn't need to clean the data.

## Step 3 - Venues Acquisiton with the Foursquare API

I have used Foursquare API to get venues suggestions for each neighbourhood in Paris. Data from Foursquare API is received in json format.

The parameters passed during the call to the Foursquare API made it possible to retrieve 100 suggestions of venues for each neighbourhood within a radius of 500 meters from the coordinate point of the neighbourhood.

The created and arranged data frame looks like this :

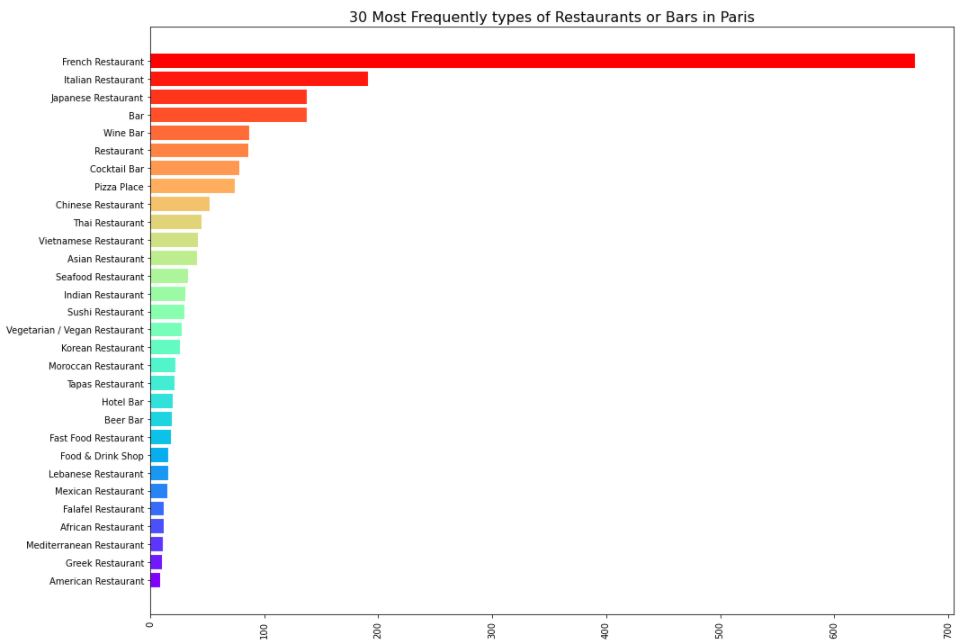


Since our client, the Tourist Office of the City of Paris, wants an analysis to advise tourists on the types of restaurants they might find during their visits to Parisian neighborhoods, we have decided to filter the places that deal with only food and drink.

The new data frame created looks like this :



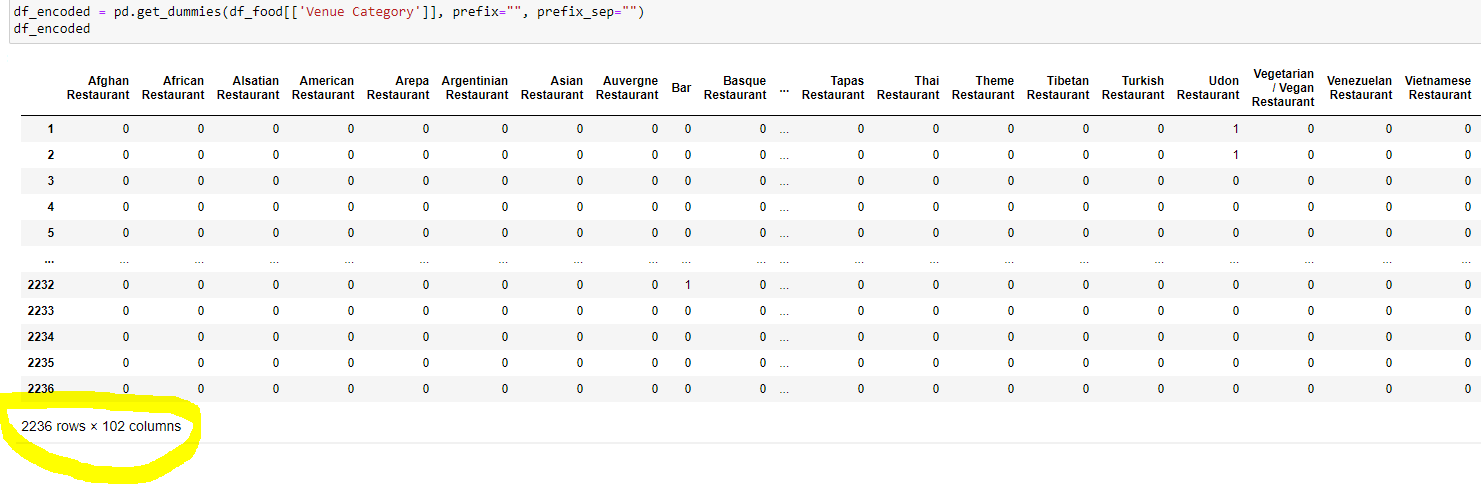
By exploring the data, we realize that French restaurants are widely represented in Paris. Indeed :



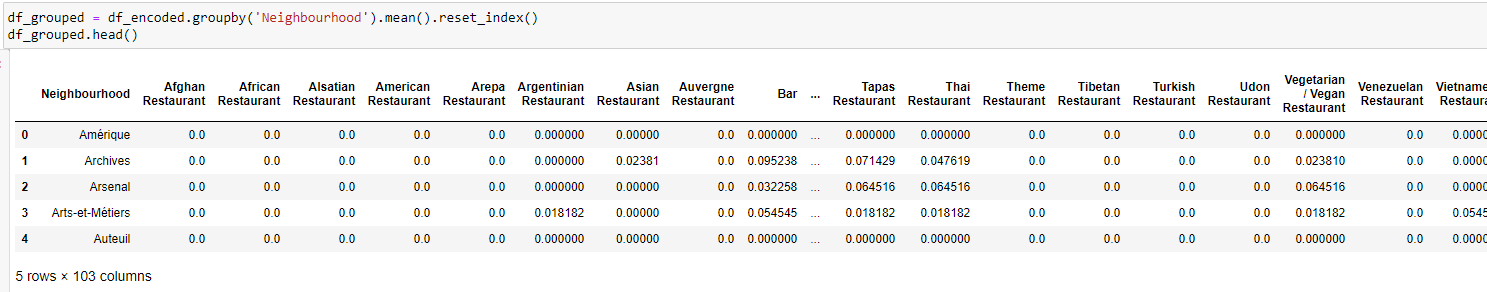
## Step 4 - Performing one-hot encoding to analyze neighbourhoods with the PANDA library

One-Hot Encoding is a technique which ensures that machine learning algorithms can process data. Namely, it converts categorical variables into the binary Boolean ones.

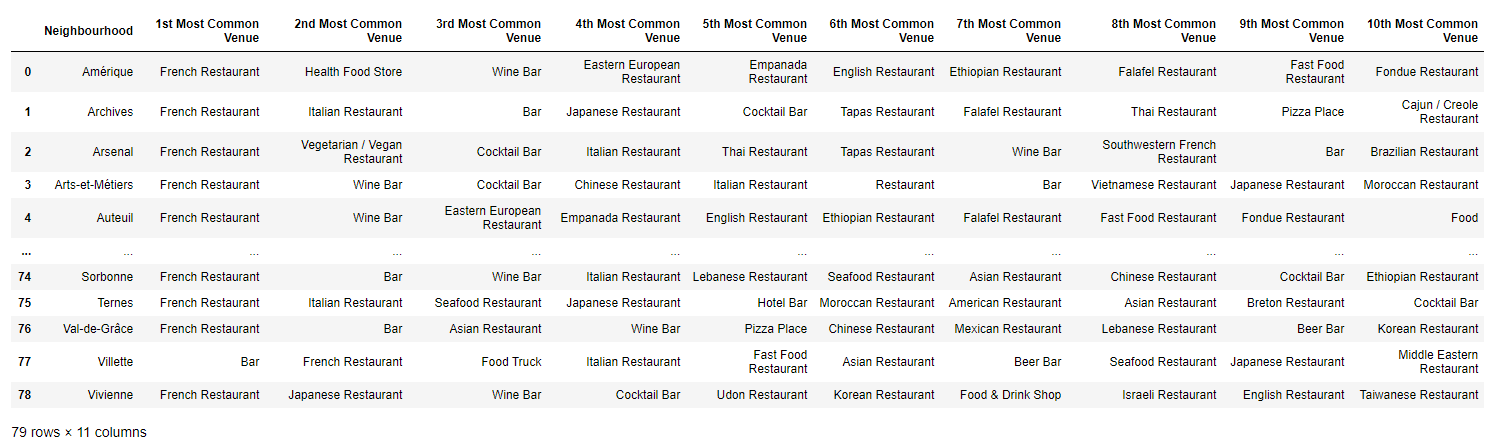
Below, an extration of the data frame created after applying the One-Hot Encoding function :



Then, I have used this data frame to create a new one grouping the categories of restaurants in each neighborhood of the city. Below, an extraction of this new data frame:



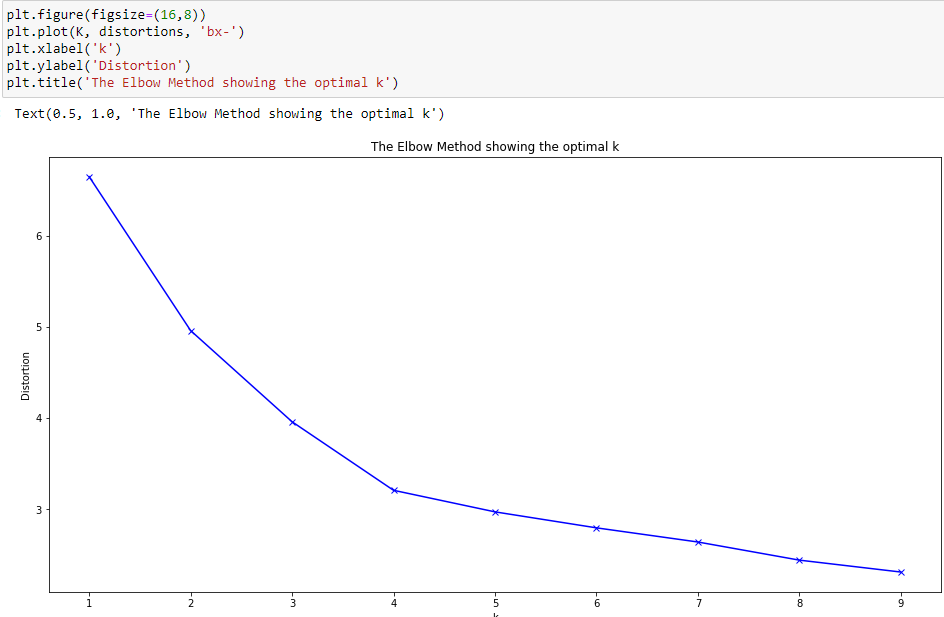
Finally, I have used the previously created data frame to create a new one with the 10 most common types of restaurants for each neighborhood in the city.



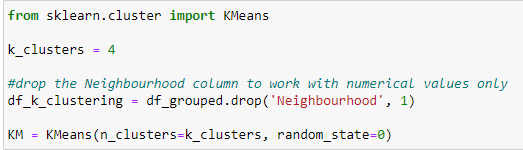
This last data frame allowed me to run an unsupervised machine learning algorithm, more precisely, a k-means clustering algorithm from the scikit-learn package.

## Step 05 - Clustering K-means with sklearn cluster library - To do

But before I could run the k-means clustering algorithm, I have used the ellbow method to set the value of the optimum k :



The graph of the elbow method, shows that the optimal value of k is 4. So I chose k as being 4, to run the k-means clustering algorithm.



After adding the label of the cluster to the data frame and after having merged it with the data frame containing the borough information, I get this data frame which will allow me to analyze the clusters that have been created :

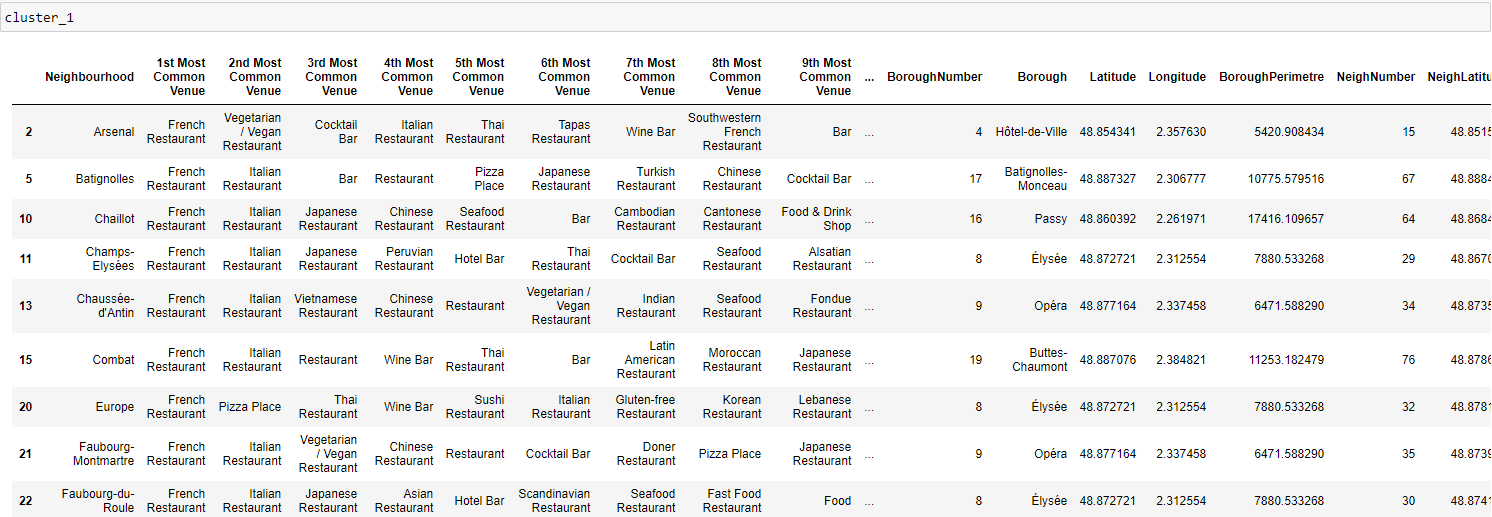


## Step 06 - Cluster Analysis

So four groups of concentrations of types of restaurants were created by the K-Means model for the city of Paris. I have named these groups according to the frequency of the types of restaurants that appear the most among the first 3 most common venues.

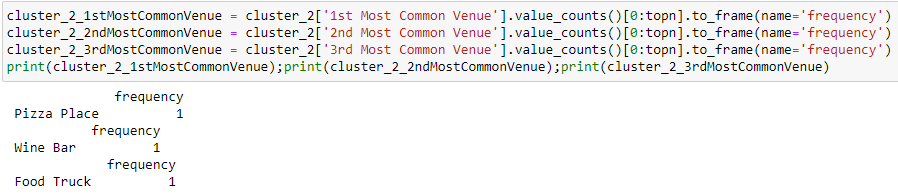
### Cluster 1 (0) - French and Italian Restaurants

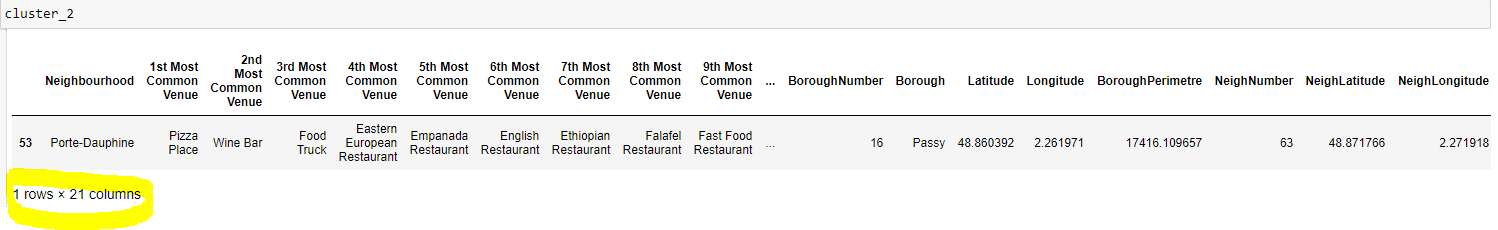




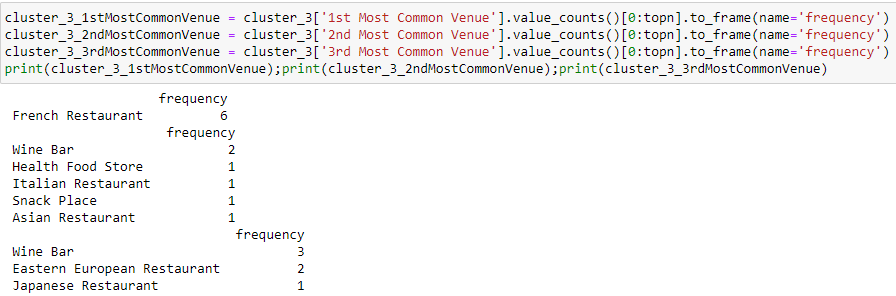


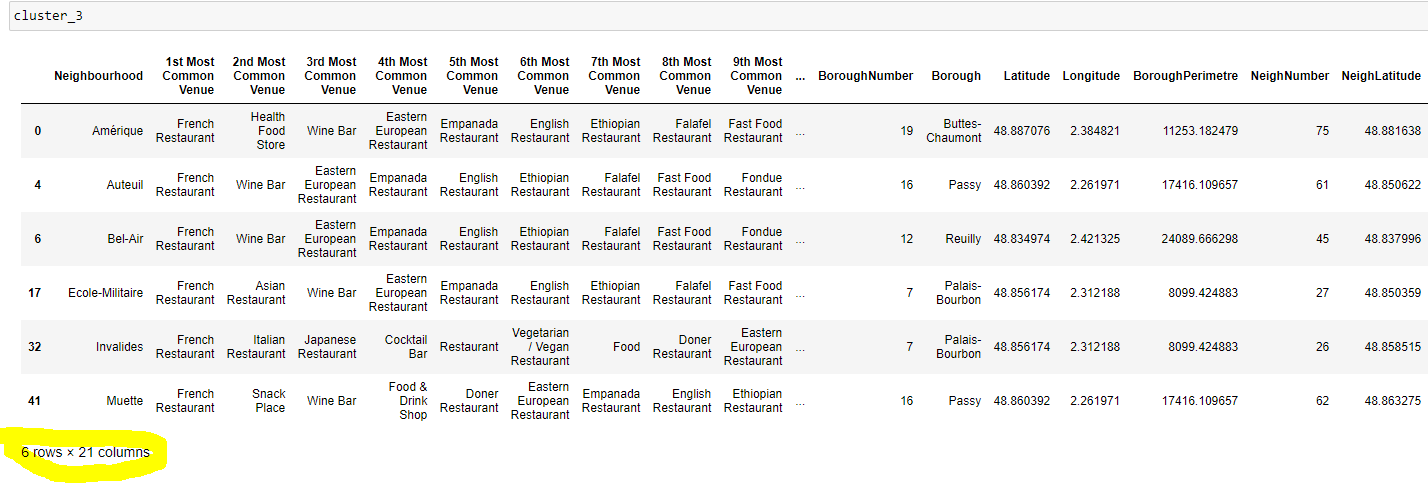
### Cluster 2 (1) - Pizza Place and International Cuisine



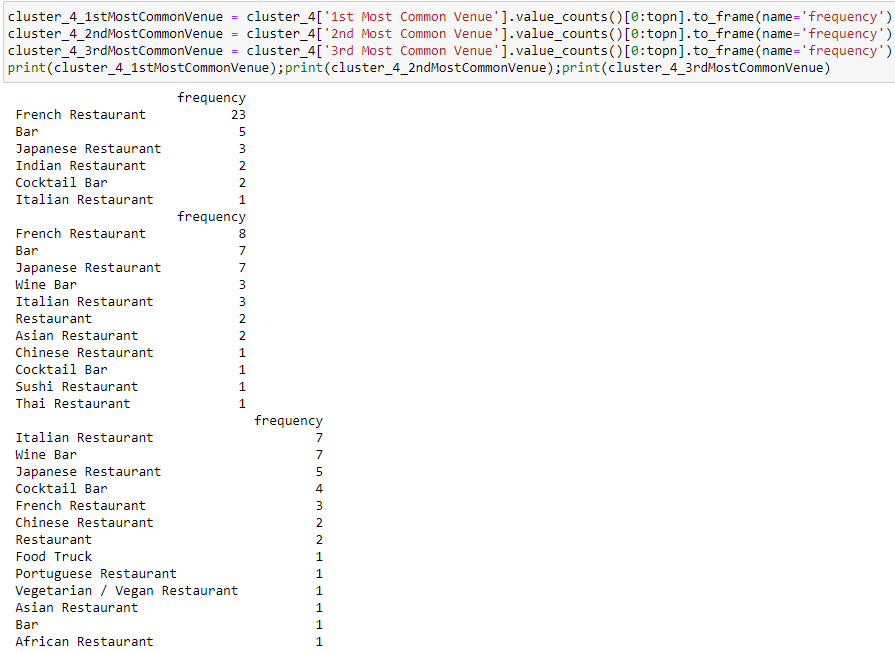


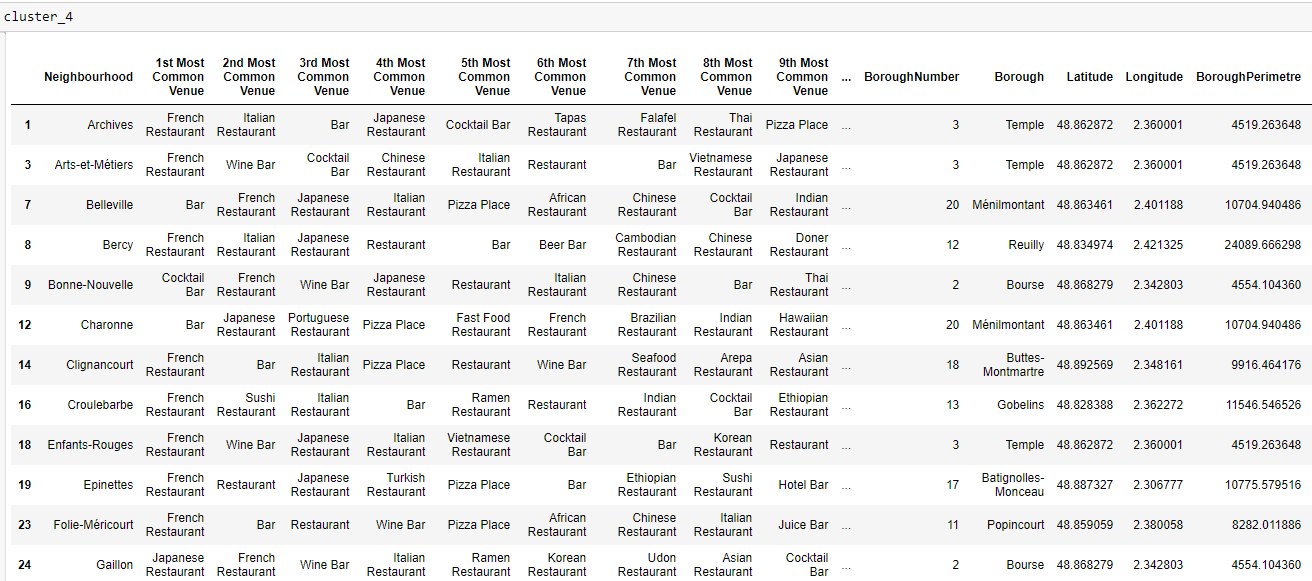
### Cluster 3 (2) - French Restaurants and Wine Bars





### Cluster 4 (3) - French Restaurants

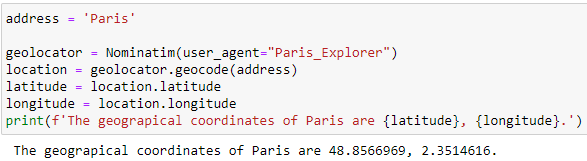




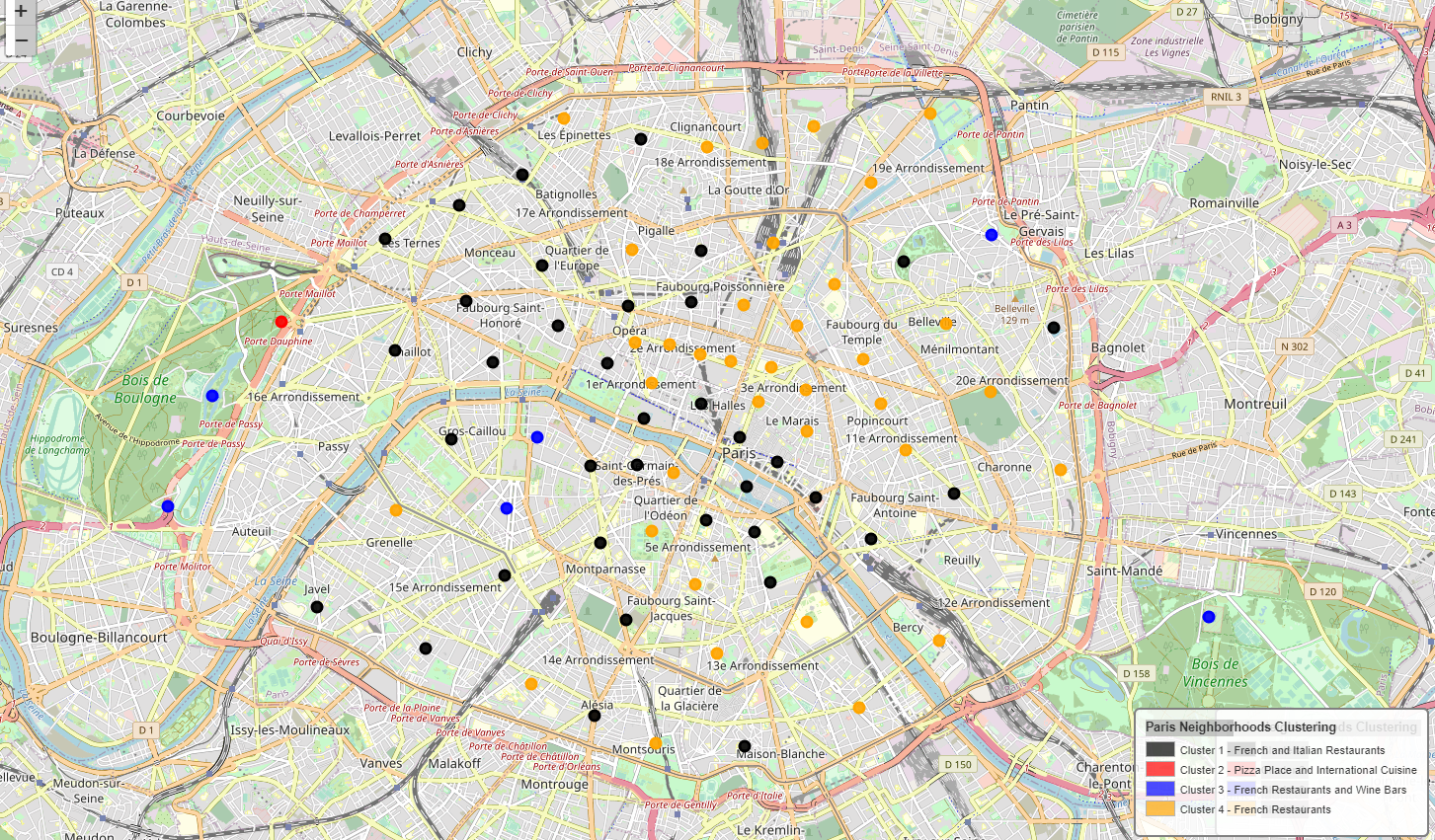


## Step 07 - Creation of a clusters map with the folium library

The geographical coordinates of Paris were extracted using GeoPy libray in Python.



The geographic coordinates will be used to draw the map of Paris with arrondissements colored according to the types of restaurants that are present.



# **Discussion**

As I indicated previously, French restaurants are widely represented in Paris. Indeed, there are nearly 680.

In 3 out of 4 clusters reproduced by the K-Means model, there is a large number of French restaurants in the first most frequent place of each of the districts.

Only group 2 (Pizza Place and International Cuisine) does not have a French restaurant. This cluster is made up solely of the Porte Dauphine district located in the 16th arrondissement of Paris.

We also note that only the Picpus district, located in the 12th arrondissement of Paris, does not have a restaurant.

# **Conclusion**

The objective of the project presented at the beginning of this report has been achieved. Indeed, my company has delivered as planned to the Tourist Office of the city of Paris a tool which now allows them to advise tourists on the types of restaurants they could find during their visits to the Parisian districts.

# **Acknowledgement, References & Links**

## Acknowledgement

This report refers to the lab projects of the IBM Data Science Professional Certificate course on Coursera.

To carry out this project, I recovered a large part of the Python code used in the different labs of Course 9.

I was also inspired by the many Notebooks published by students who had thought about the subject before me.

## References

Source : https://en.wikipedia.org/wiki/Arrondissements\_of\_Paris

**Source :** [**https://www.parisinsidersguide.com/paris-neighborhoods.html**](https://www.parisinsidersguide.com/paris-neighborhoods.html)

Source : <https://fr.wikipedia.org/wiki/Liste_des_quartiers_administratifs_de_Paris>

## Links

The notebook with the code for this project, as well as the report, can be found in my [github repository](https://github.com/philippe-bottier/Coursera_Capstone/blob/4c9d8eadc640a7830f03707b17bde46cdedf8a04/Capstone%20Project%20-%20Battle%20Neighborhoods%20Paris%20V3.ipynb).