# CAPSTONE PROJECT

# Battle of Neighbourhood – Report

## 1. Introduction to business problem

### ***Problem Background: -***

Tourism in France has directly contributed 78.9 billion euros to total **Gross Domestic Product (GDP)** in the past years**.** 30% of which comes from international visitors and 70% comes from domestic tourism spending. France was visited by 89 million foreign tourists in 2018, making it most popular tourist destination in the world. In this project, we will be exploring two famous cities of France, *Paris* and *Strasbourg*. Even though France was most populous tourist destination, considering the number of nights spent in the country, it is in sixth place after *United States*, *United Kingdom, China, Spain* and *Italy.*

**Paris**, the Capital City of France is the third most visited city in the world. It has some of the world's largest museums including *Louvre* which is most visited art museum in the world. It hosts some of the world recognizable landmark such as ***Eiffel Tower, The Arc de Triomphe*** and many more.

**Strasbourg** is one of the four main capital of European Union alongside Brussels, *Luxembourg* and *Frankfurt.* It is among the few cities in the world not being a state capital and hosting international organization of the first order. Economically, it is an important centre of manufacturing and engineering. It is the second largest river port in France after Paris. The city is chiefly known for its *sandstone Gothic Cathedral* with its famous astronomical clock.

Evidently, both of these cities are rich in cultural heritages and thus attract millions of international tourists every year. As France stands at sixth position in terms of nights spent by these tourists, even though it is most popular tourist destination in the world, it will be helpful for tourists to have a rough idea about luxurious apartments, hotels and restaurants, pub, café etc. to make their stay more comfortable. This might change the current scenario by improving it's rank from sixth to 2nd or 3rd. If possible, France might stand at first in terms of total number of nights spent by tourists.

### ***Problem description: -***

It is obvious that people who visit these places are somewhere in need of a physical/virtual guide. Through this project, I have explored these two main tourist places to dig some of the useful information about all those luxurious amenities, tourist would be looking for. These basic luxurious resources could be: -

- Hotels
- Restaurants
- Multiplexes
- Opera House
- Mountains
- Museums
- Night Club
- Super Market etc.
  In addition to these, it can be quite helpful for those people who all are international migrants and are looking for perfect place to rent apartments. So, our project could be proven helpful for these immigrants as virtual guide. Our main aim is to provide an outlook of all these available venues within these cities so that people would be less reliable on local guides who often charge these immigrants huge amount in exchange of service.

# 2. Data

## ***Data source - 1***

In this project, we will be exploring **Paris** and **Strasbourg.**

The dataset has been collected from Kaggle. It can be downloaded from this link. The dataset prepared by INSEE. It is the official French institute gathering data of many types around the France.

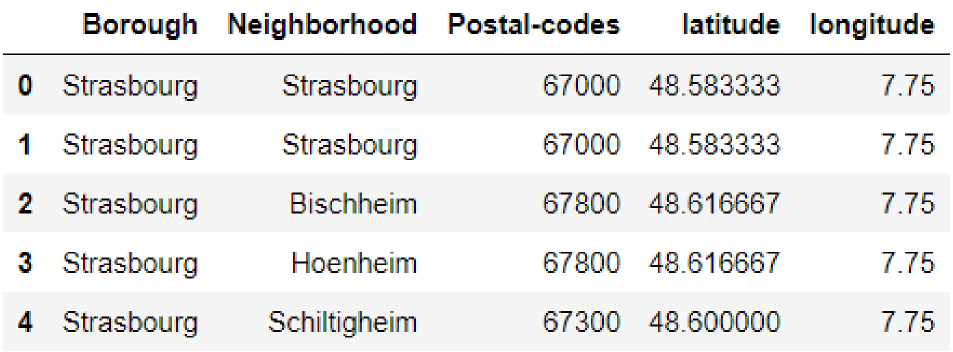
There were four files in the dataset, but as per the requirements, I have only used ***name_geographic_information.csv*** dataset. Given Dataset contains following features: -

- EU_circo: name of the European Union Circonscription
- Code_region: code of the region attached to the town
- nom_région: name of the region attached to the town
- chef.lieu_région: name the administrative center around the town
- numéro_département : code of the department attached to the town
- nom_département : name of the department attached to the town
- préfecture : name of the local administrative division around the town
- numéro_circonscription : number of the circumpscription
- nom_commune : name of the town
- codes_postaux : post-codes relative to the town
- code_insee : unique code for the town
- latitude : GPS latitude
- longitude : GPS longitude
- éloignement : I couldn't manage to figure out what was the meaning of this number

out of above features, only few were helpful. So, I performed Data wrangling to extract useful features, so that appropriate Machine-Learning algorithm could be used to extract useful information with more accuracy. Those features which were used as primary features for our models are listed below: -

- prefecture – renamed as Borough
- nom_commune – renamed as Neighborhood
- codes_postaux – renamed as Postal-codes
- latitude
- longitude

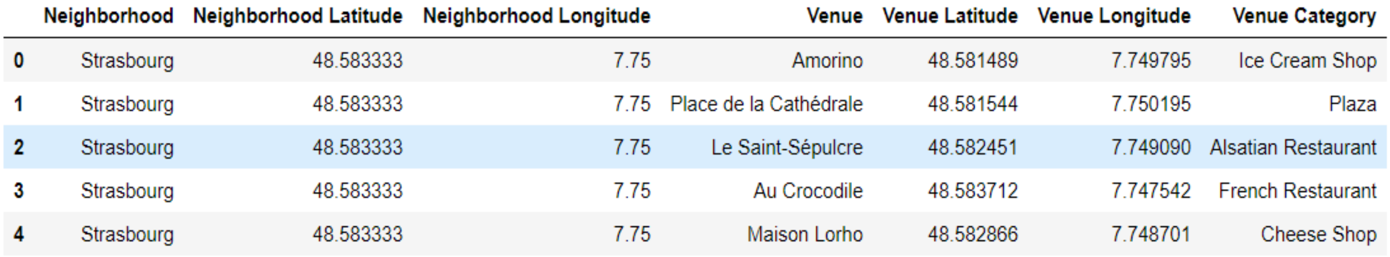
Given dataset contains many prefectures out of which only ***Paris*** and ***Strasbourg*** have been taken into consideration as only we are interested in exploring only these two cities. Since the dataset contains missing values, and we are only interested in exploring the venues, dropping the missing rows would be appropriate choice. After data cleaning, dataset looks like –



| | Borough | Neighborhood | Postal-codes | latitude | longitude |
|---|---|---|---|---|---|
| **0** | Strasbourg | Strasbourg | 67000 | 48.583333 | 7.75 |
| **1** | Strasbourg | Strasbourg | 67000 | 48.583333 | 7.75 |
| **2** | Strasbourg | Bischheim | 67800 | 48.616667 | 7.75 |
| **3** | Strasbourg | Hoenheim | 67800 | 48.616667 | 7.75 |
| **4** | Strasbourg | Schiltigheim | 67300 | 48.600000 | 7.75 |

### ***Data source – 2***

we will be using Foursquare API to leverage neighbourhood venues by providing geographical coordinates along with user credentials. Once, the neighbourhood venues are explored, data frame looks like –



| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| **0** | Strasbourg | 48.583333 | 7.75 | Amorino | 48.581489 | 7.749795 | Ice Cream Shop |
| **1** | Strasbourg | 48.583333 | 7.75 | Place de la Cathédrale | 48.581544 | 7.750195 | Plaza |
| **2** | Strasbourg | 48.583333 | 7.75 | Le Saint-Sépulcre | 48.582451 | 7.749090 | Alsatian Restaurant |
| **3** | Strasbourg | 48.583333 | 7.75 | Au Crocodile | 48.583712 | 7.747542 | French Restaurant |
| **4** | Strasbourg | 48.583333 | 7.75 | Maison Lorho | 48.582866 | 7.748701 | Cheese Shop |

## 3. Methodology

### ***Business Understanding*: -**

Our main aim is to segregate the suitable places for tourists to stay where they could afford all the facilities mainly affordable hotels, restaurants, and amusement parks. We also aim at providing virtual guides to international migrants regrading suitable places to buy apartments.

### ***Analytic approach: -***

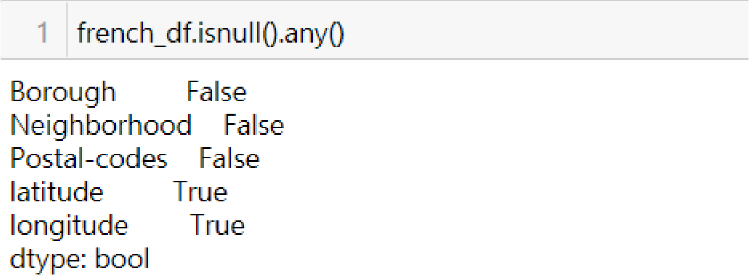
Original data frame consists of 36840 rows and 24 columns. After cleaning the data, the clean data frame consists of 519 rows and 5 columns. There is significant decrease in the number of rows because we would be considering those rows only which contains the information about **Paris** and **Strasbourg.** We will be using K-Means clustering machine learning model to cluster neighbourhoods of these cities based on certain criteria.

### ***Exploratory Data Analysis (EDA):* -**

### Geographical data exploitation: -

Original dataset contains many rows out of which only some are useful. As our main aim is to cluster neighbourhood, we will extract useful information from the given dataset at first.

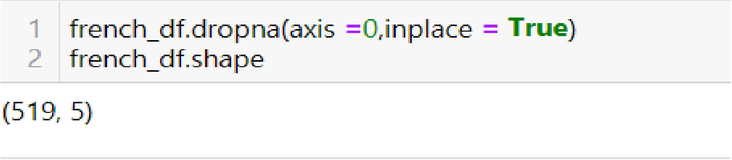
- Original dataset contains missing values which can be seen below.



```
french_df.isnull().any()
```

```
Borough         False
Neighborhood    False
Postal-codes    False
latitude         True
longitude        True
dtype: bool
```

As we can see that latitude and longitude columns of ***french_df*** dataset contains missing values, we will ignore all those rows which contains missing values by using below code –

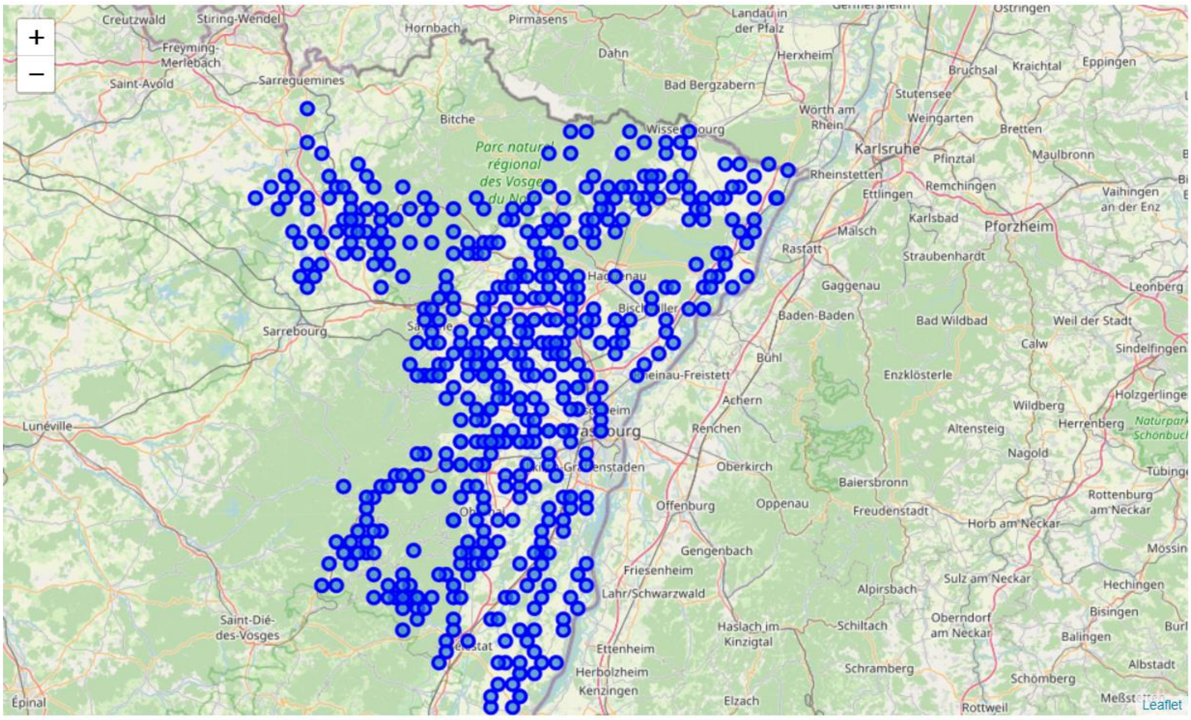


```
french_df.dropna(axis =0,inplace = True)
french_df.shape
```

```
(519, 5)
```

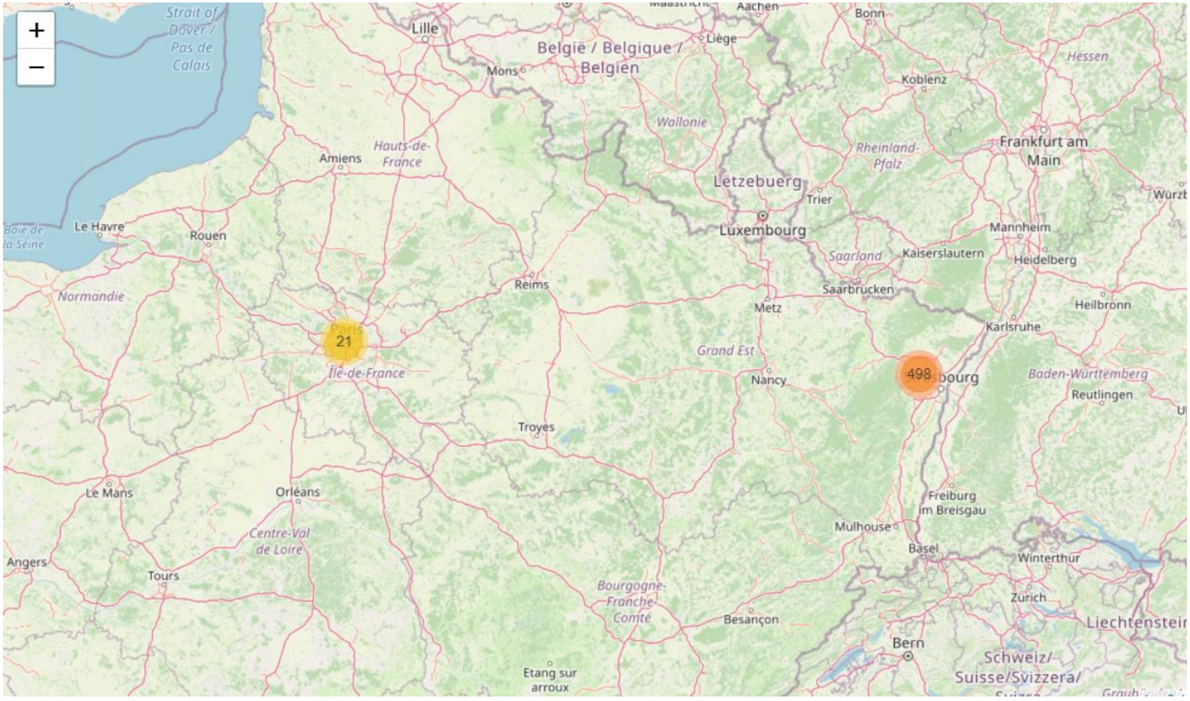
As of now, only 519 rows remain in the ***french_df*** dataset.

- We used Geopy along with Folium library to plot the neighbourhoods on the map.



From the above plot, we can conclude that these cities are densely populated.

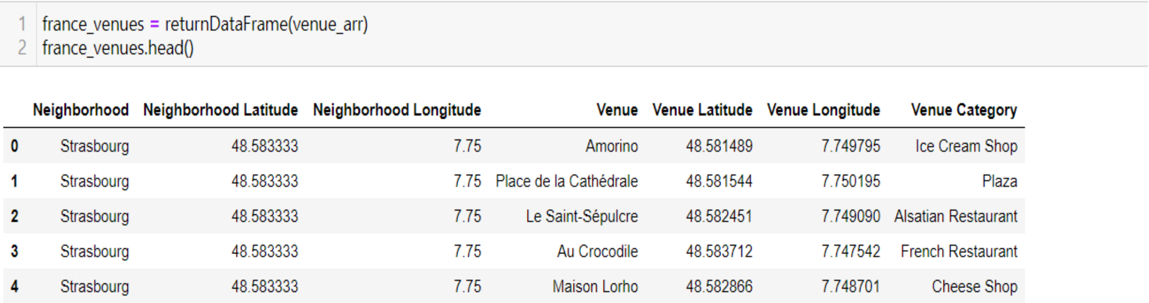
- Let’s cluster these neighbourhood into two cluster namely*,* ***Paris*** and ***Strasbourg.*** We will visualize these cluster using folium library and Geopy.



From the above plot, we can see that there are 21 localities in the Paris’s neighbourhood while there are 498 localities in Strasbourg’s neighbourhood.

## **Quantitative analysis of Neighbourhood venues** –

- Foursquare API was used to collect all the neighbourhood venues of these two cities. Once the neighbourhood data are collected, new data frame would look like below.

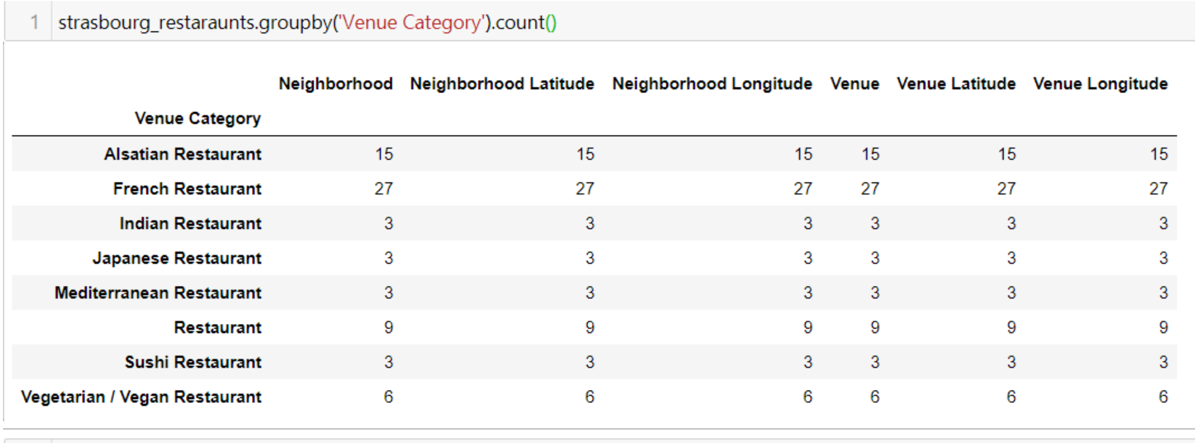


```
france_venues = returnDataFrame(venue_arr)
france_venues.head()
```

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Strasbourg | 48.583333 | 7.75 | Amorino | 48.581489 | 7.749795 | Ice Cream Shop |
| 1 | Strasbourg | 48.583333 | 7.75 | Place de la Cathédrale | 48.581544 | 7.750195 | Plaza |
| 2 | Strasbourg | 48.583333 | 7.75 | Le Saint-Sépulcre | 48.582451 | 7.749090 | Alsatian Restaurant |
| 3 | Strasbourg | 48.583333 | 7.75 | Au Crocodile | 48.583712 | 7.747542 | French Restaurant |
| 4 | Strasbourg | 48.583333 | 7.75 | Maison Lorho | 48.582866 | 7.748701 | Cheese Shop |

We get a total of 2793 rows along with 7 columns.

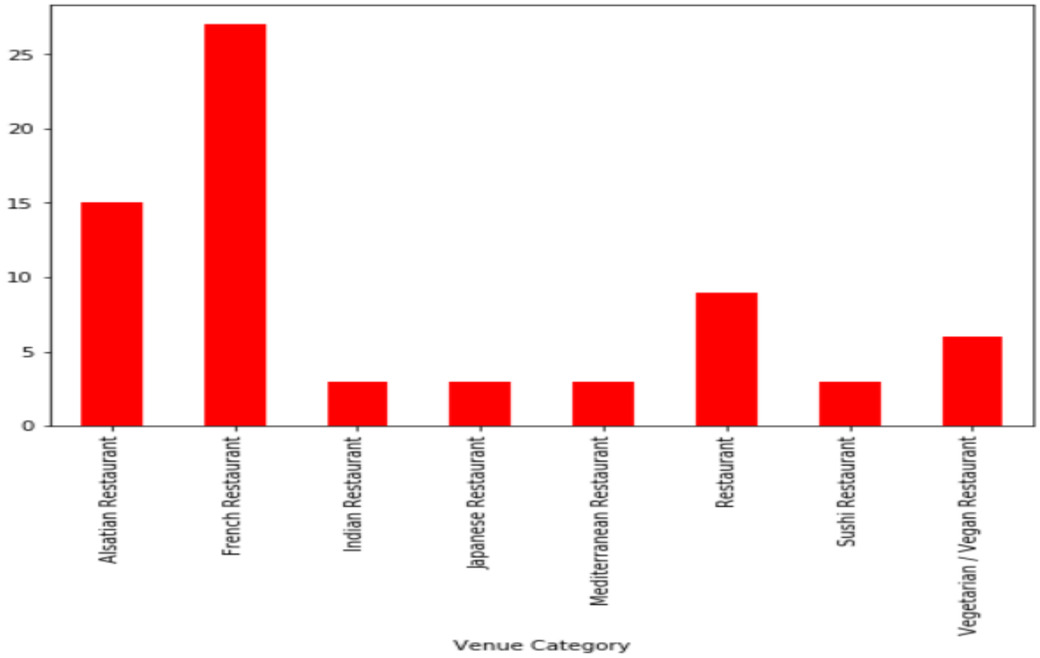
- we explored these cities one by one. We found that **Strasbourg** is having total of 69 restaurants out of which there are 27 French restaurants only which is 39% of total.



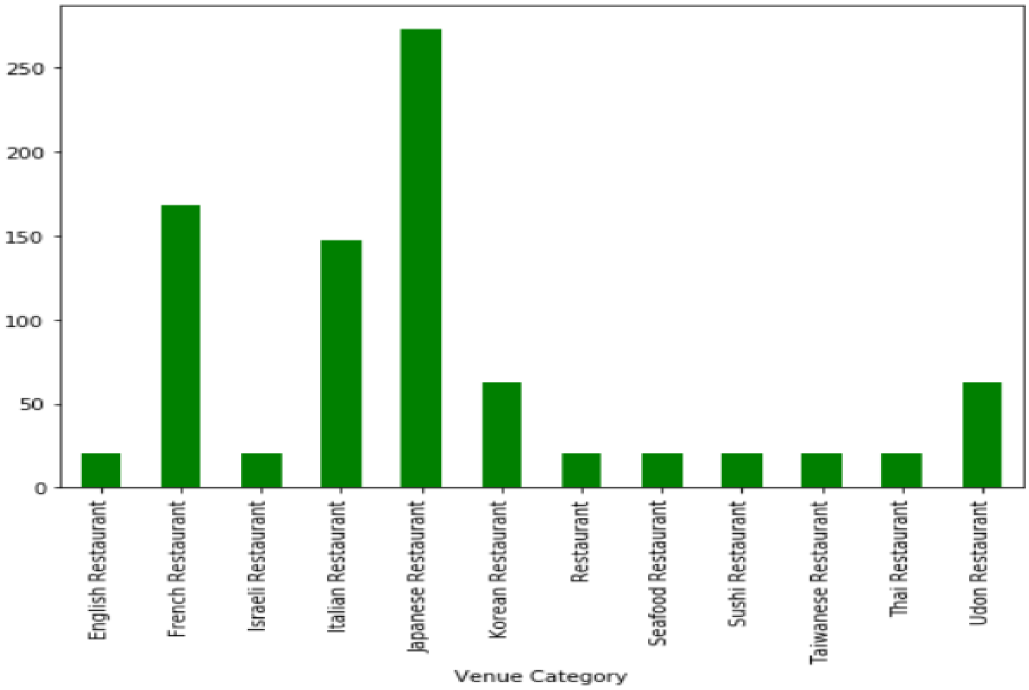
```
strasbourg_restaraunts.groupby('Venue Category').count()
```

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude |
|---|---|---|---|---|---|---|
| Venue Category | | | | | | |
| Alsatian Restaurant | 15 | 15 | 15 | 15 | 15 | 15 |
| French Restaurant | 27 | 27 | 27 | 27 | 27 | 27 |
| Indian Restaurant | 3 | 3 | 3 | 3 | 3 | 3 |
| Japanese Restaurant | 3 | 3 | 3 | 3 | 3 | 3 |
| Mediterranean Restaurant | 3 | 3 | 3 | 3 | 3 | 3 |
| Restaurant | 9 | 9 | 9 | 9 | 9 | 9 |
| Sushi Restaurant | 3 | 3 | 3 | 3 | 3 | 3 |
| Vegetarian / Vegan Restaurant | 6 | 6 | 6 | 6 | 6 | 6 |

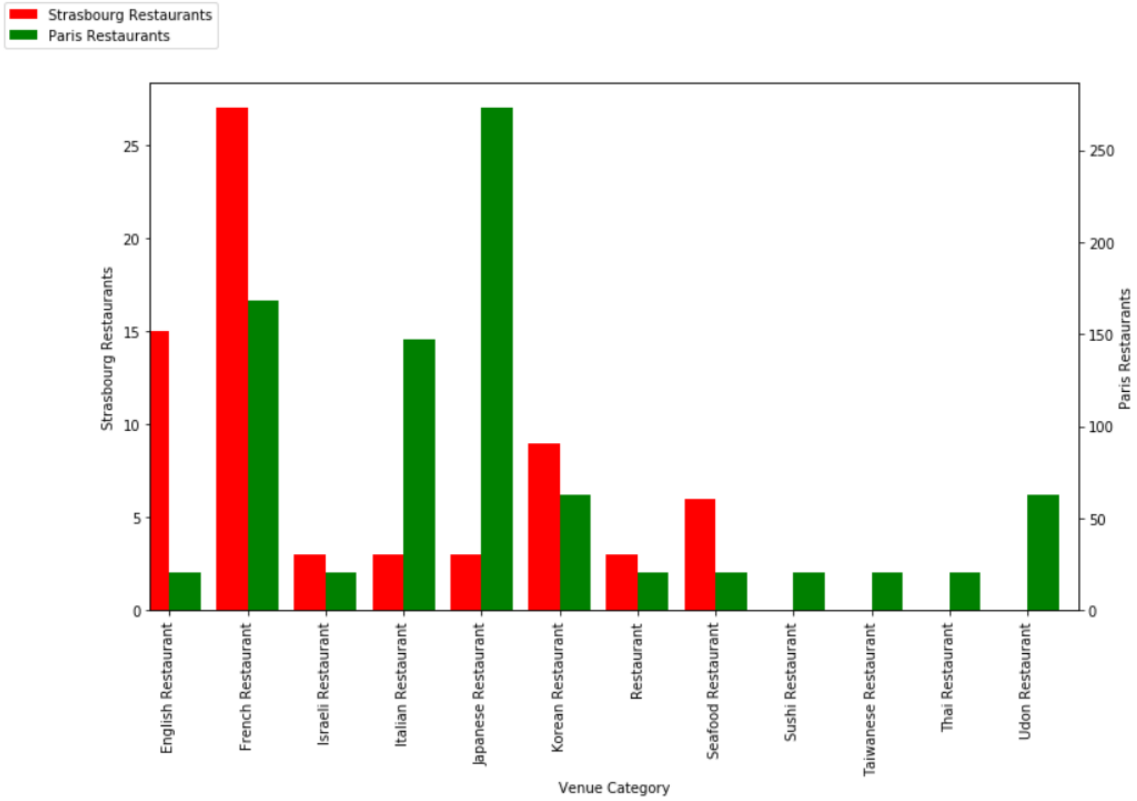
We can observe this by looking at graph below.



- Once we explored **Paris,** we found out that there are total of 861 restaurants in its neighbourhood, out of which there are 273 Japanese restaurants only which is even more than total of **Strasbourg.**
  we can verify it by looking at graph below.



- We compared these two cities’ restaurants, we found that there are a smaller number of restaurants in **Strasbourg** than in **Paris.** There could be many reasons for this huge margin. One such reason is that **Strasbourg** is European capital region where there are a greater number of administrative blocks than restaurants.
  We can verify it by using joint plot of above bar graphs.



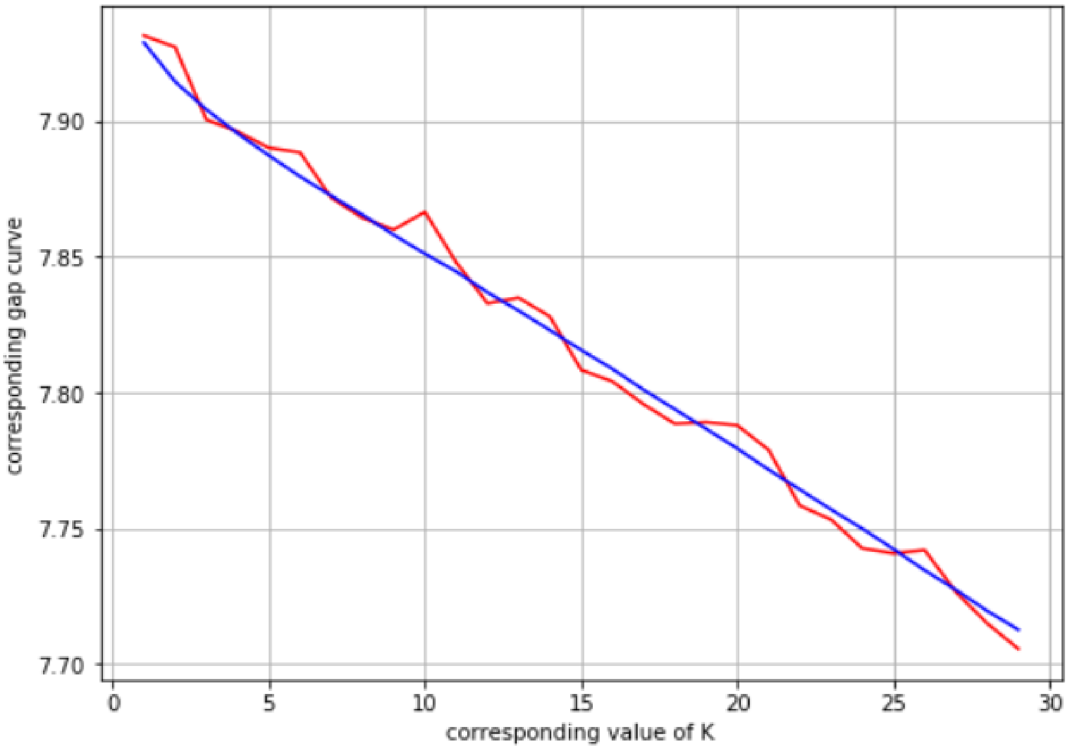
# 4.Result

On the above dataset, we performed K-Means cluster analysis to cluster these neighbourhood on the basis of mean of frequency of these venues. We will study all the neighbours of these two cities jointly.

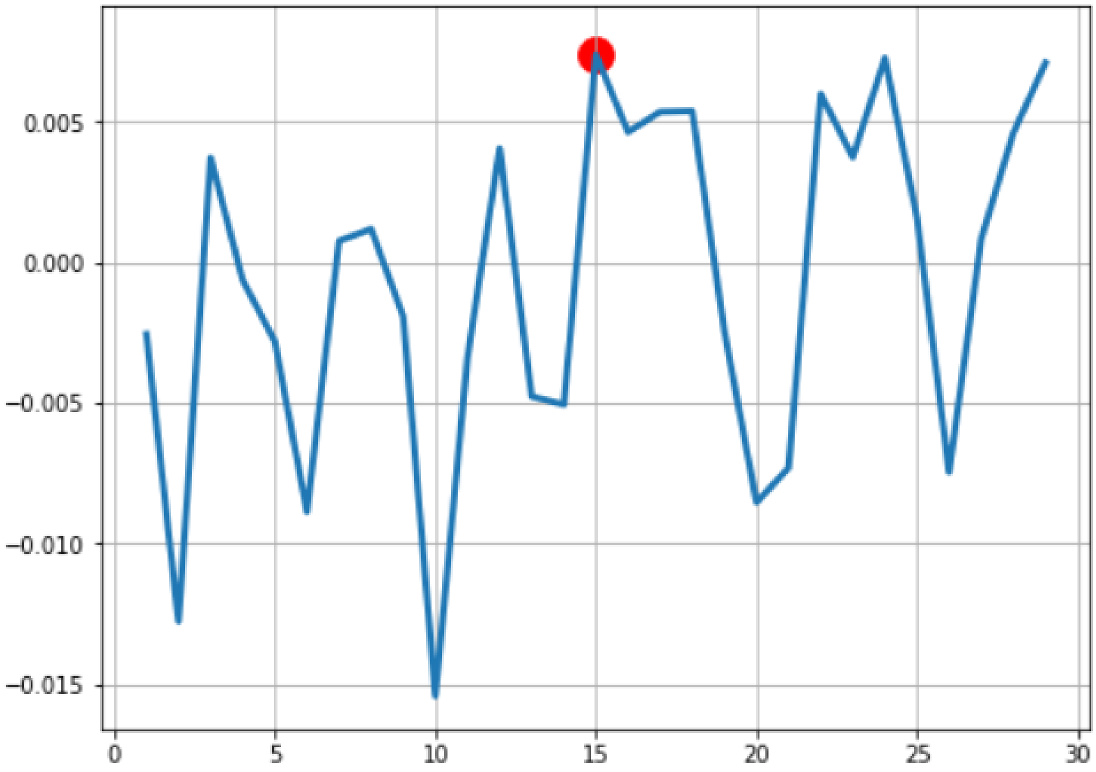
**K- Means cluster –**

we will cluster these neighbourhood into **k** number of clusters. We derived the optimal value of **k** using gap statistics. The optimal value of **k** came out to be 15. We will cluster these neighbourhood into 18 cluster using K-Means clustering machine learning algorithm.

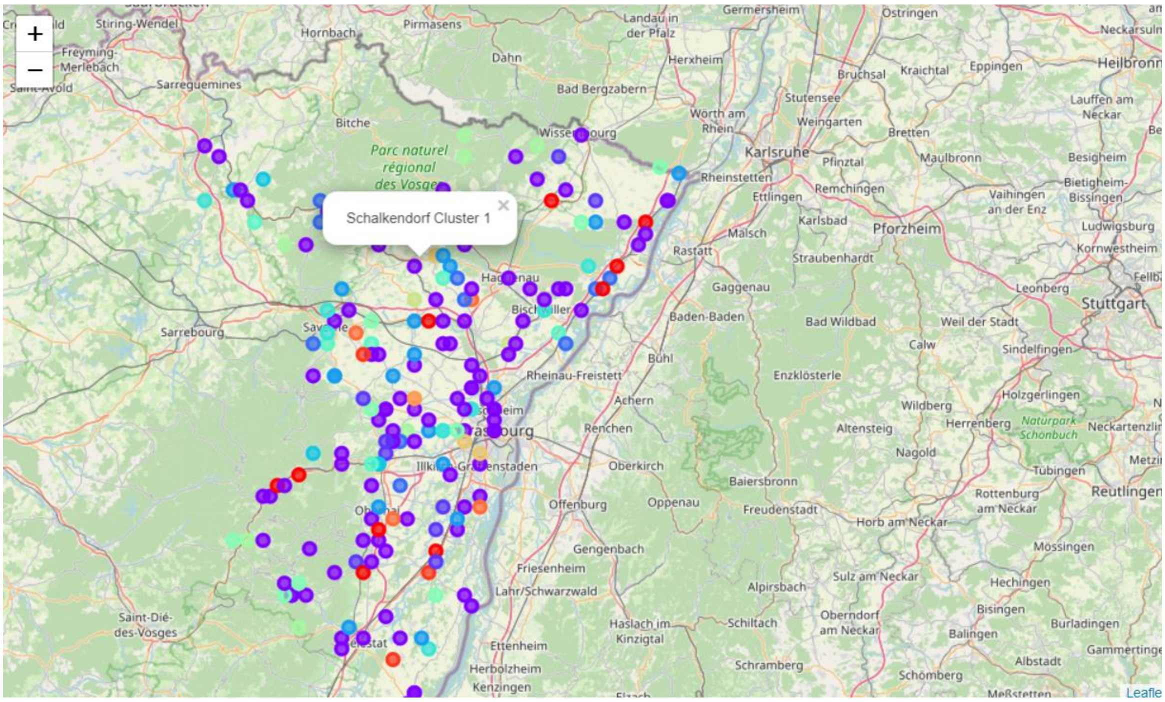
- Optimal value of **k** can be verified below.
  <u>*Gap curve*</u> –



<u>*Optimal value of K-*</u>
After

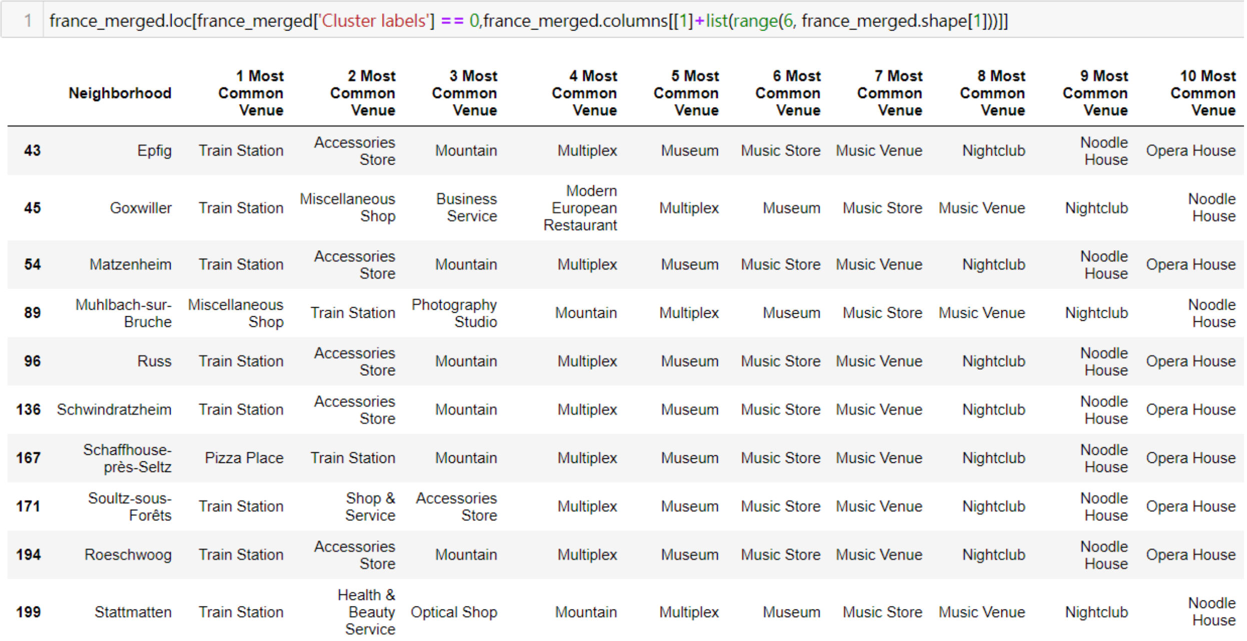


After clustering, map looks like below –



By looking at above folium map, we can conclude that Blue dots are more compact and are greater in numbers. By looking at the labels, we can say that cluster 1 is having a greater number of venues that can be help migrants decide where to rent apartment.

- ❖ Let’ examine **Cluster-1** in details.

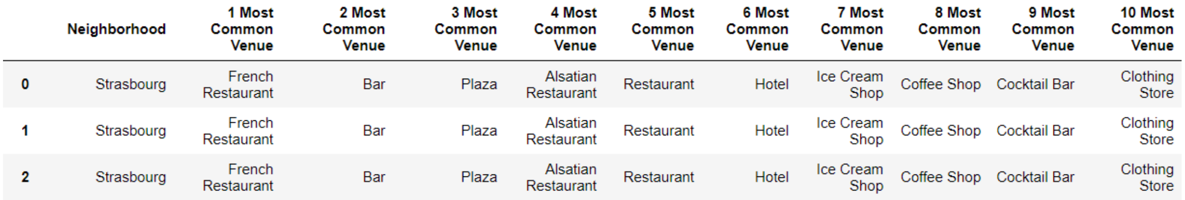


```
france_merged.loc[france_merged['Cluster labels'] == 0,france_merged.columns[[1]+list(range(6, france_merged.shape[1]))]]
```

| | Neighborhood | 1 Most Common Venue | 2 Most Common Venue | 3 Most Common Venue | 4 Most Common Venue | 5 Most Common Venue | 6 Most Common Venue | 7 Most Common Venue | 8 Most Common Venue | 9 Most Common Venue | 10 Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **43** | Epfig | Train Station | Accessories Store | Mountain | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House | Opera House |
| **45** | Goxwiller | Train Station | Miscellaneous Shop | Business Service | Modern European Restaurant | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House |
| **54** | Matzenheim | Train Station | Accessories Store | Mountain | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House | Opera House |
| **89** | Muhlbach-sur-Bruche | Miscellaneous Shop | Train Station | Photography Studio | Mountain | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House |
| **96** | Russ | Train Station | Accessories Store | Mountain | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House | Opera House |
| **136** | Schwindratzheim | Train Station | Accessories Store | Mountain | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House | Opera House |
| **167** | Schaffhouse-près-Seltz | Pizza Place | Train Station | Mountain | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House | Opera House |
| **171** | Soultz-sous-Forêts | Train Station | Shop & Service | Accessories Store | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House | Opera House |
| **194** | Roeschwoog | Train Station | Accessories Store | Mountain | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House | Opera House |
| **199** | Stattmatten | Train Station | Health & Beauty Service | Optical Shop | Mountain | Multiplex | Museum | Music Store | Music Venue | Nightclub | Noodle House |

In the above dataset, we can see that most common venue in cluster-1 is Train station, followed by Mountain, Museum, music-store etc. so by looking at these venues, we can suggest tourists to visit these places without any discomfort.

- ❖ If someone is fond of French food, then neighbours in cluster-2 as well as in cluster-5 would be best places to visit.
  We can verify this below.



| | Neighborhood | 1 Most Common Venue | 2 Most Common Venue | 3 Most Common Venue | 4 Most Common Venue | 5 Most Common Venue | 6 Most Common Venue | 7 Most Common Venue | 8 Most Common Venue | 9 Most Common Venue | 10 Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Strasbourg | French Restaurant | Bar | Plaza | Alsatian Restaurant | Restaurant | Hotel | Ice Cream Shop | Coffee Shop | Cocktail Bar | Clothing Store |
| 1 | Strasbourg | French Restaurant | Bar | Plaza | Alsatian Restaurant | Restaurant | Hotel | Ice Cream Shop | Coffee Shop | Cocktail Bar | Clothing Store |
| 2 | Strasbourg | French Restaurant | Bar | Plaza | Alsatian Restaurant | Restaurant | Hotel | Ice Cream Shop | Coffee Shop | Cocktail Bar | Clothing Store |

Venues in cluster-2 are not limited to this only. If someone is foody and want to explore different kinds of cuisines, cluster-2 could be the best choice.

## 5. Conclusion

This analysis has been performed on the legal dataset. These two cities are not the only places to visit in France and to rent apartments. There are many such destinations. But our main idea was to highlight the available venues within these two places as they are the centre of tourism. There are many venues within these cities like book-store, bar, music store, supermarket, Clothing store etc. These venues can be added benefits to the tourists’ interests. One might look on these two cities in different perspective. Choice is completely independent of interests. Some tourists might be interested in doing outdoor recreations, some might be looking for suitable places to open a restaurant, shopping malls, book-store, clothing-store, Pastry shop etc. Our model can be helpful to these people and thus they can decide according to their choice of interests.