**A Sunday morning in Paris**

**Author: F. Muller**

**Introduction: the business problem**

Paris is a city with an amazing choice of museums to visit, more than a hundred. I visited myself quite a number of them, and I noticed a recurring fact along my visits: museums make people hungry. After a few hours of admiring paintings or sculptures during a cultural Sunday morning, the urge for a snack or a nice lunch can get strong.

Fortunately, Paris also hosts thousands of restaurants, cafés and the like, from cuisines all over the world! The combination “museum + lunch” is therefore very easy to arrange.

Now, say you are a travel agency operating in Paris, and you want to offer short packages combining the visit of a museum and lunch at a restaurant. For the sake of the exercise, I will assume here that the design of such packages will be independent from the museums and will be entirely based on the lunch options available nearby a museum.

You would need to identify which museums are nearby interesting combinations of lunch options in order to offer them as relevant packages to clients. Among many possibilities, museums could for instance be classified between high-end, medium or affordable restoration options.

The goal of this project is thus the following: can we classify the museums in Paris based on their neighboring restaurants options by leveraging data science and machine learning?

### Data

Since the problematic of this little project is not too distant from that of the classification of Toronto neighborhoods, the data required and the general methodology followed will not be very different either.

The final product will be a classification of the museums in Paris into clusters based on their nearby restaurants, using unsupervised machine learning. The source data for this machine learning algorithm will be information about the restaurants nearby each museum, obtained through the Foursquare API. Hopefully enough restaurants exist near most museums, but some museums may have to be skipped if they do not have enough venues nearby.

In order to use the Foursquare API, a database of museums containing their geo-localization must first be created. This database will be the result of a web scraping operation, more precisely from Wikipedia pages. Maps will also be created in order to visualize the distribution of museums in Paris, especially after their clustering by the k-means algorithm.

A list of the museums in Paris is available at the page <https://en.wikipedia.org/wiki/List_of_museums_in_Paris>. This list will be transferred into a pandas DataFrame. For each museum in the list, a link to its specific Wikipedia page is available. The French version of each museum page contains its latitude and longitude, which will allow to create the database required for the Foursquare API usage.

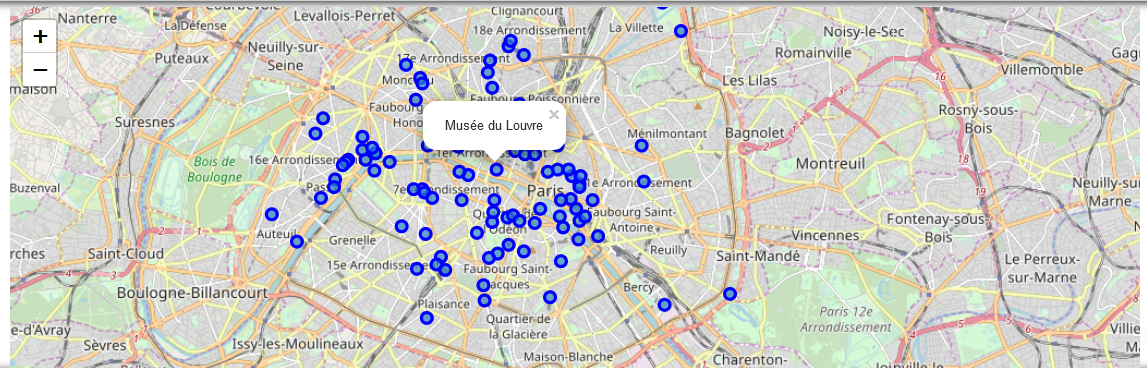


Figure 1: Folium map of the museums in Paris

### Methodology

#### Data cleaning

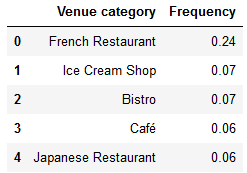
Following the creation of a database containing a list of all museums in Paris with their corresponding latitude and longitude, I used the Foursquare API to obtain a list of 200 venues 300 meters around each museum.

The museums with a very low (<10) number of venues around them were removed from the dataset. Indeed, since the business problem is based on the idea of dual museum/restaurant tours for tourists, these museums would be placed in areas too remote in the city to be interesting.

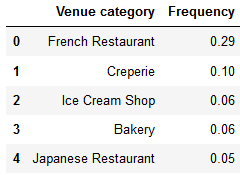
As an exploratory data analysis, I extracted and then examined the list of venue categories obtained. Initially, 125 unique restaurant categories were obtained, which I considered a number quite too large.

Following these data cleaning operations, example tables for the most frequent venues around three museums are shown below.

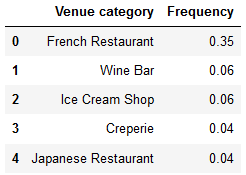
Arab World Institute



Archaeological Crypt of the Paris Notre-Dame



Polish Library in Paris



Three main issues were identified:

1. The "French Restaurant" category (unsurprisingly) was the most frequent venue for all museums.
2. A significant number of categories contained only one venue, or were very similar between one another.
3. A few museums had a very limited number of venues around them.

Both of these issues raised in my opinion a problem of representativity for the data.

I tackled these issues by fusing the single-venue categories into larger, more general categories, and by merging similar categories together.

For instance:

* I removed the generic "Restaurant" and "Diner" types because they did not provide enough information to be valuable here.
* I removed supermarkets, stores, shops, markets and street food from the list, as they are not considered restaurants within this study.
* For restaurants types with less than 3 entries, I reclassified them into more general categories so that they have a more meaningful impact on the classification. Examples: the “Venezuelan”, ‘Peruvian” and “Argentina” restaurant categories were fused into the larger “Latin American Restaurant” category, the “Udon”, “Soba”, “Okonomiyaki” and “Sushi” categories were merged with the “Japanese Restaurant” category.
* I fused categories that were really similar. Examples: I merged the “Brasserie” and “Bistro” categories, and created a new category “Regional French Cuisine” to regroup restaurants from the “Savoyard”, “Provençal”, “Breton”, “Basque”, etc. categories.

These actions lead to a smaller number of venue categories (53 at the end of the data cleaning operations), but each with more venues in average, which would allow them to be significant compared to the "French Restaurant" category in the classification phase.

#### Classification using machine learning

This problem being a classification problem without an a-priori knowledge of the correct results, an unsupervised machine learning was to be used.

I first tested the DBSCAN method, which has the advantage of not requiring a set number of clusters. The results obtained were quite poor, with either most museums classified as outliers or into one large cluster. In my interpretation, the number of variables in this problem (around 50) lead to this poor behavior.

I then applied a k-means method to the problem. The main parameter of this algorithm is the number of cluster desired. Without any knowledge of the possible number of classes, I used the “Elbow method” to gain an idea of the evolution of the cluster dispersion with an increasing number of clusters. Figure 2 was obtained by varying the number of cluster *k* from 2 to 20, and calculating the Sum-of-Squared-Errors.

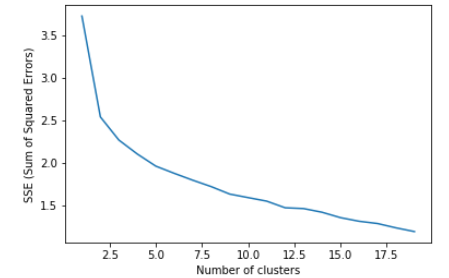


Figure 2: Evolution of the Sum-of-Squared Errors (SSE) with an increasing number of clusters using a k-means algorithm.

The sharpest error decrease was obtained from k = 2 to k = 3. However, I observed more relatable clusters with k = 4 based on my personal knowledge of the city. Therefore, combining this analytical data showing decreasing interest in additional clusters past k = 4 and my personal experience, I chose to select k = 4 for further analysis of the results.

### Results

Following the execution of the k-means algorithm, all museums were sorted into 4 clusters depending on the venues around them. In order to understand how this classification works, we can observe for some museums in each cluster the most frequent venues.

The first cluster contains French Restaurants around 25%, and a lot of Italian (10-20%) and Japanese (around 10%) restaurants. We can label it "French/Italian/Japanese".

|  |  |  |
| --- | --- | --- |
| Gobelins manufactory | Mundolingua | Cernuschi museum |

The second cluster contains mostly French Restaurants (around 50%), and a lot of Cafés and Bistros. We can label it "Traditional French".

|  |  |  |
| --- | --- | --- |
| Galerie nationale  du Jeu de Paume | Rodin museum | Orsay museum |

The third cluster is much more diverse than the others, with more restaurants from Asia or the rest of the World. For some rare museums, other categories than “French Restaurant” are the most frequent. We can label this cluster "Cosmopolitan".

|  |  |  |
| --- | --- | --- |
| Cabinet des Médailles | Musée Pasteur | Museum of Jewish  Art and History |

The fourth cluster contains around 30% of French Restaurants, and Bakeries are coming up more often than in other clusters. We can label this cluster "Baguette zone".

|  |  |  |
| --- | --- | --- |
| Archaeological Crypt of the Paris Notre-Dame | Bibliothèque-Musée de l'Opéra National de Paris | Maison de Victor Hugo |
|  |  |  |

Following this basic identification of the nature of each cluster, we can display them on the map shown in figure 3 using colors.

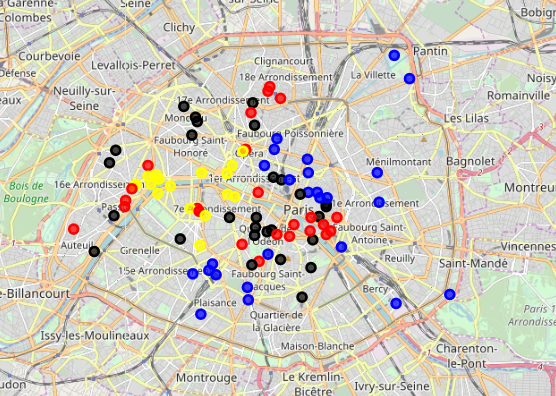


Figure 3: Folium map of the museums in Paris, now grouped into clusters based on the nearby restaurants. Yellow: “Traditional French”, Red: “Baguette zone”, Blue: “Cosmopolitan”, Black: “French/Italian/Japanese”

### Discussion

If we inspect in figure 3 the different clusters, we can recognize some Parisian neighborhoods, which corroborates the classification obtained by the k-means algorithm.

For instance:

1. The "Traditional French" zone in yellow is mostly around the Eiffel Tower, the Louvre and the Orsay Museum. These high-end areas feature more traditional French restaurants.

2. The "Baguette zone" in red corresponds to popular zones such as the student neighborhood of Paris-Dauphine and the Pigalle-Montmartre neighborhood. An increased presence of bakeries and cheap restaurant options in popular areas seems plausible.

3. The "Cosmopolitan" area in blue can especially be found in the Marais which is a gentrified, diverse neighborhood. In this area, restaurants from all over the world are more readily found. This cluster also features various museums along the outer edge of the city, which is plausible since these areas are more diverse than the center of Paris.

4. The "French/Italian/Japanese" zone can mostly be found in the periphery of Paris and in the neighborhoods of Montparnasse and Saint-Germain-des-Prés. These are "upper class" areas, which have included Italian (pizzas) and Japanese (sushi) cuisine into their main preferences.

### Conclusion

The clustering of the museums in Paris based on the restaurants each features in its vicinity using a k-means algorithm has lead us to interesting observations. Each of the four clusters obtained could be related to a socio-economic category of the Parisian population, or to particular neighborhoods.

A possible business plan for a travel agency wanting to offer coherent museum + restaurant combinations could be:

* A “Premium” offer for the traditional area of the cluster “Traditional French”
* A “Business” offer for the upper-class area of the cluster “French/Italian/Japanese”
* A “Economy” offer (maybe aimed at students) for the cluster “Baguette zone”
* A “World” offer aimed at multi-cultural tourists for the cluster “Cosmopolitan”.

This simple study could be expanded by taking into account the type of each museum in the classification, the rating of each venue in the Foursquare API, and the distance between museums in each cluster in order to offer multiple-museum visits.