Applied DS Capstone - Final Project

The Battle of Neighborhood: Night at the Museum (in Paris)

Francesca Collu

November 2020

Introduction

If you only had a couple of days of holidays and you were really into art, which Parisian arrondissement should you head on in order to optimize your time of vacation in the museums capital in the world? In the present report, we will search an answer to this question.

Overview of the Problem

Paris, the capital of France and one of the most visited city in Europe, has been crowned the museum capital of the world as well, accounting for 297 museums. The variety of art expression in Paris is clear in the several shapes of it: not only museums, but also cinemas, theaters, monuments, music venues, outdoor sculptures and so on.

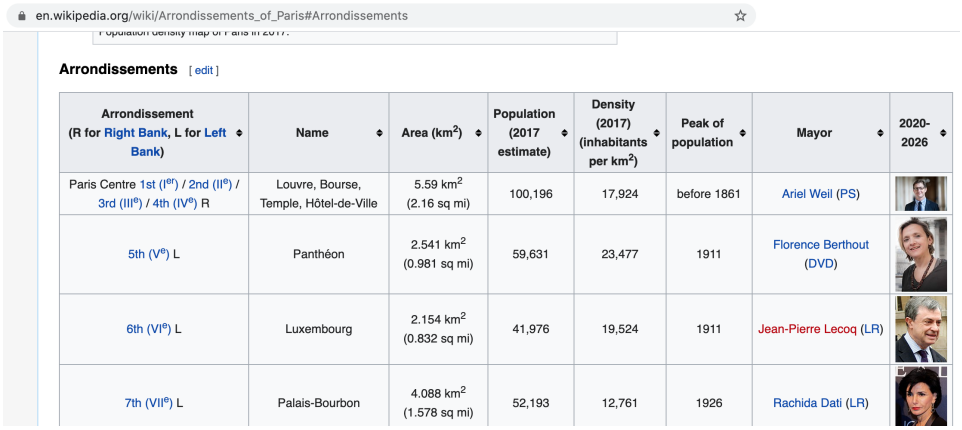
The aim of this report is to analyze the distribution of all the Paris art venues, such as art galleries, libraries, movie theaters, historic sites, theaters, science museums. The target audience of this problem is the art enthusiasts who find themselves in Paris, wondering in which neighborhood they should be staying in for the vacation in order to spend as much time as they can visiting a specific kind of art venue. For the same purpose, another target audience of this problem is also a travel agency, that is planning an itinerary for a journey in Paris for some clients.

Data Sources

In order to accomplish the goal set in the previous section, several data sources have been used and they can be summed up in the following way:

- Districts of Paris Wikipedia page[1]: data was scrapped from this page to create a dataframe with the arrondissements of Paris. In fig. 1 we can see the table where I get the Parisian arrondissements and their names;
- Geocoder: this Python geocoding library was used to identify the geographical coordinates of all the districts of Paris. Thanks to Geocoder, it has been possible to retrieve the dataframe with the arrondissements and their coordinates, as we can see in fig. 2 draw the map of the arrondissements, shown in fig. 3;

[1] `https://en.wikipedia.org/wiki/Arrondissements_of_Paris`



en.wikipedia.org/wiki/Arrondissements_of_Paris#Arrondissements

Arrondissements [edit]

Arrondissement (R for Right Bank, L for Left Bank)	Name	Area (km²)	Population (2017 estimate)	Density (2017) (inhabitants per km²)	Peak of population	Mayor	2020-2026
Paris Centre 1st (I^{er}) / 2nd (II^{e}) / 3rd (III^{e}) / 4th (IV^{e}) R	Louvre, Bourse, Temple, Hôtel-de-Ville	5.59 km² (2.16 sq mi)	100,196	17,924	before 1861	Ariel Weil (PS)	
5th (V^{e}) L	Panthéon	2.541 km² (0.981 sq mi)	59,631	23,477	1911	Florence Berthout (DVD)	
6th (VI^{e}) L	Luxembourg	2.154 km² (0.832 sq mi)	41,976	19,524	1911	Jean-Pierre Lecoq (LR)	
7th (VII^{e}) L	Palais-Bourbon	4.088 km² (1.578 sq mi)	52,193	12,761	1926	Rachida Dati (LR)	

Figure 1: Table from Wikipedia page of the districts of Paris.

- Places API of Foursquare: it has been used to extract the data about the venues of the city, in particular the name, location and category of each venue. In this way it has been possible to retrieve the dataframe shown in fig. 4.

Methodology and Analysis

After scrapping the data from the Wikipedia page of the arrondissements of Paris[2], it has been possible to retrieve a dataframe with the Paris arrondissements and their name; they have been used to get the geographical coordinates of the Parisian districts through the help of *Geocoder*, as we can see in the dataframe shown in fig. 2.

At this point, the *Places API* of Foursquare have been used to locate all the venues in every arrondissement and their exact location. As said before, the goal of this work is to analyze the art venues only, so the results given by Foursquare have been filtered using some key-words as *Art*, *Gallery*, *Museum*, *Comedy*, *Theater*. In this way, 24 categories of art venues have been found out. So, after using the one-hot encoding and taking the mean of the frequency for each art venue, their distribution in the whole city has been studied. This is shown in the plot in fig. 5.

In order to obtain the clusters and find out in which way the arrondissements can be identified and described, we implement the *K-Means* algo-

[2] `https://en.wikipedia.org/wiki/Arrondissements_of_Paris`

	Arrondissement	Name	Latitude	Longitude
0	Paris Centre 1st (Ier) / 2nd (IIe) / 3rd (IIIe...	Louvre, Bourse, Temple, Hôtel-de-Ville	48.857101	2.353064
1	5th (Ve) L	Panthéon	48.846210	2.346110
2	6th (VIe) L	Luxembourg	48.847580	2.340940
3	7th (VIIe) L	Palais-Bourbon	48.860830	2.318590
4	8th (VIIIe) R	Élysée	48.869317	2.316878
5	9th (IXe) R	Opéra	48.882110	2.327990
6	10th (Xe) R	Entrepôt	48.842150	2.375990
7	11th (XIe) R	Popincourt	48.859350	2.376010
8	12th (XIIe) R	Reuilly	48.845002	2.389365
9	13th (XIIIe) L	Gobelins	48.834504	2.353472
10	14th (XIVe) L	Observatoire	48.835960	2.334260
11	15th (XVe) L	Vaugirard	48.839450	2.300620

Figure 2: Dataframe containing the districts of Paris and their coordinates.

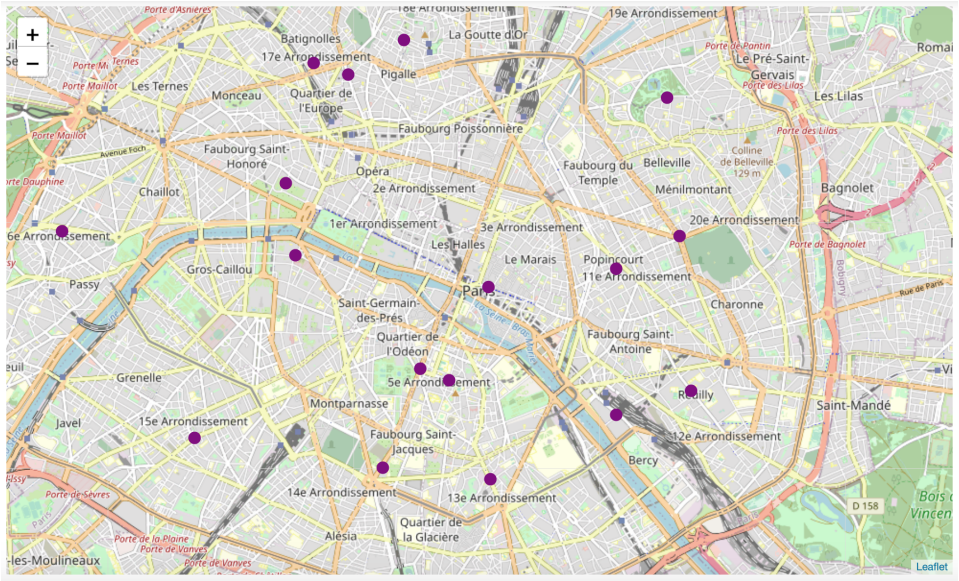


Figure 3: Maps of Paris. The purple circles point out the arrondissements.

	Arrondissement	Arrondissement Latitude	Arrondissement Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Paris Centre 1st (Ier) / 2nd (IIe) / 3rd (IIIe...	48.857101	2.353064	Fleux'	48.858763	2.354161	Furniture / Home Store
1	Paris Centre 1st (Ier) / 2nd (IIe) / 3rd (IIIe...	48.857101	2.353064	Place de l'Hôtel de Ville – Esplanade de la Li...	48.856925	2.351412	Plaza
2	Paris Centre 1st (Ier) / 2nd (IIe) / 3rd (IIIe...	48.857101	2.353064	Maison Aleph	48.857348	2.354873	Pastry Shop
3	Paris Centre 1st (Ier) / 2nd (IIe) / 3rd (IIIe...	48.857101	2.353064	Galerie Azzedine Alaïa	48.857545	2.355217	Art Gallery
4	Paris Centre 1st (Ier) / 2nd (IIe) / 3rd (IIIe...	48.857101	2.353064	Parc Rives de Seine	48.855510	2.351419	Park

Figure 4: Dataframe of the venues in Paris, retrieved with Places API of Foursquare.

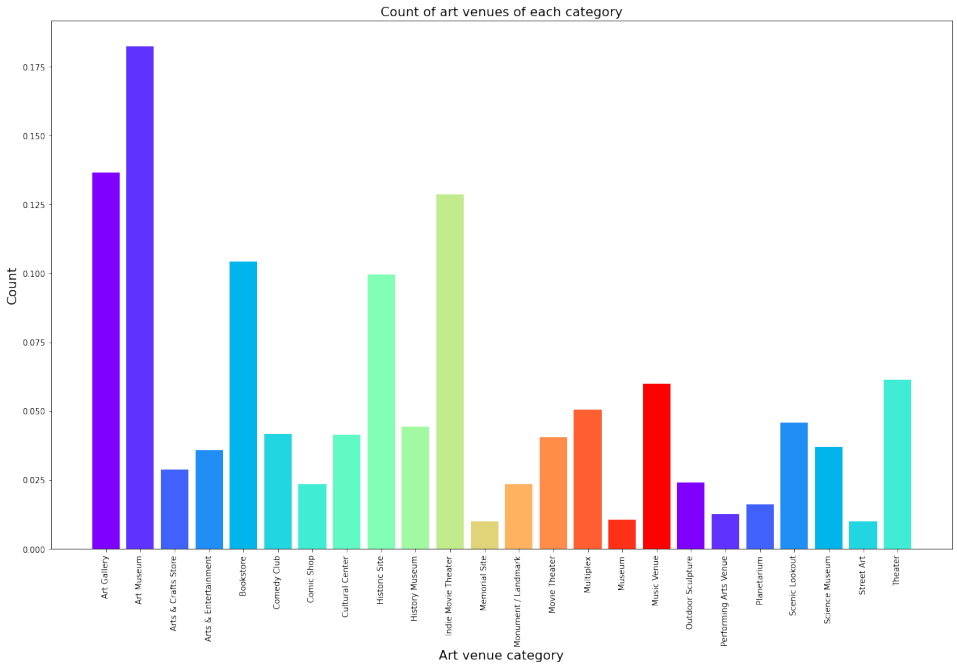


Figure 5: Count of the different categories of art venues in Paris.

rithm. The first question we want to answer is the following: which one is the optimal number of clusters? It will be the one that is not too small to produce a result that does not capture the important aspects of the data, neither too great to overfit the model.

In order to find that optimal number, we have used *KElbowVisualizer* by *Yellowbrick*. In fig. 6 we can see that the optimal number of clusters for this problem is $k = 6$.

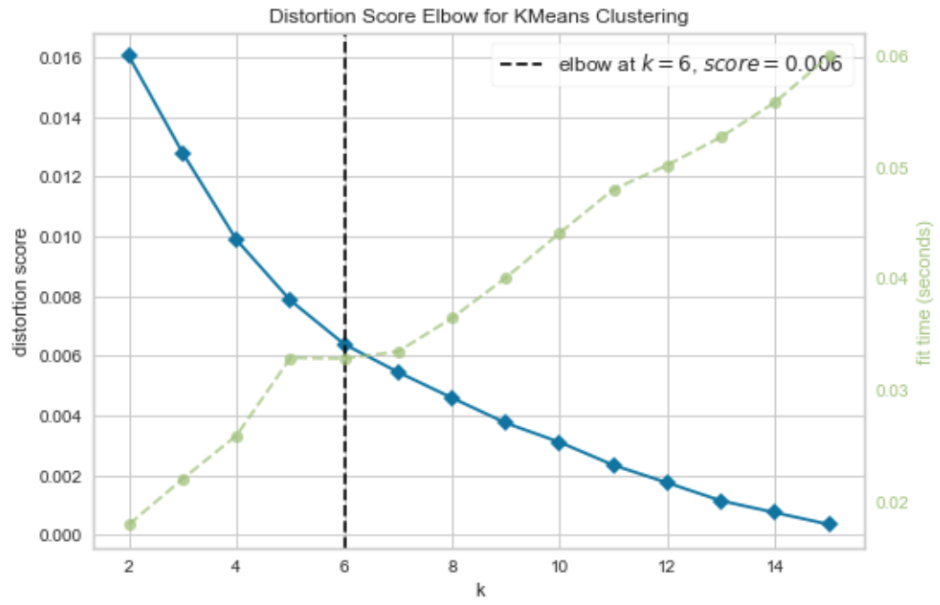


Figure 6: The blue line is the distortion score; the green line is the time needed to make the fit. Both the plots are vs the number of clusters (k). The black dashed vertical line identifies the optimal number of clusters.

We can see the distribution of the clusters on the Paris map in fig. 7.

Results

We have identified six clusters, each of which is characterized by the prevalence of a feature; in order to recognize it, we can observe the plot in fig. 8.

Cluster 1 : it is the cluster that contains the most elevated number of districts, so it involves several venues that can be associated. In this cluster, we find movie theaters of different genres, art museum, music

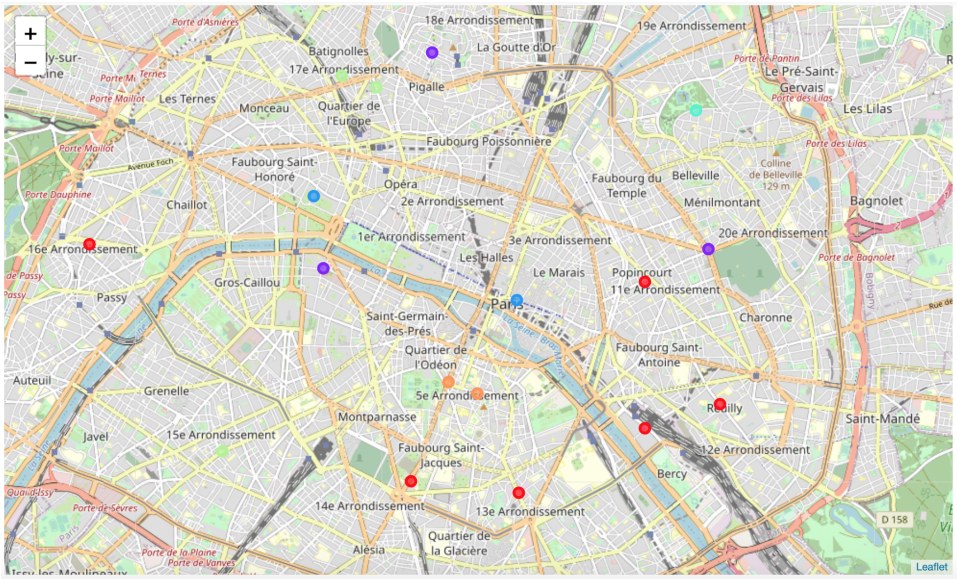


Figure 7: Map of clusters on the Paris map. The red circles represent the Cluster 1, which includes 10th, 11th, 12th, 13th, 14th, 16th arrondissements; the violet circles represent the Cluster 2, which includes 7th, 18th, 20th arrondissements; the blue circles represent Cluster 3, which includes 1st, 2nd, 3th, 4th, 8th arrondissements; the turquoise circles represent Cluster 4, which includes the 19th arrondissement; the light-green circles represent Cluster 5, which includes the 9th and 17th arrondissements; the orange circles represent Cluster 6, which includes the 5th and 6th arrondissements.

venue, historic sites, planetarium, performing arts venues and bookstores. They are all venues that recalls a dynamic tourist, who can easily move from one venue to another to enjoy all of the activity offered by this cluster of districts.

Cluster 2 : in this cluster we can see that art museums, history museums, music venues and cultural centers, but also art galleries, comedy clubs, indie movie theaters, outdoor sculptures, scenic lookouts, street art, theaters are the most common venues. The ideal target for this kind of art is maybe a tourist who like walking around the city and staying outdoors.

Cluster 3 : this cluster is characterized by art galleries, historic sites, theaters, art museums and bookstores, but also outdoor sculptures, memorial sites, arts and crafts stores, cultural centers. This is maybe the perfect cluster for a tourist who love History and Historic art.

Cluster 4 : this cluster is perfectly made up by four categories of art venue: art museum, arts and entertainment, historic site, scenic lookout. This is a cluster the target of which is the tourist who likes art and panoramic views: it could be ideal for taking pictures, so this kind of tourist may be a photography passionate.

Cluster 5 : in this cluster we can observe a relevant presence of bookstores, indie movie theaters, arts and crafts stores, comedy clubs, cultural centers, music venues, art galleries. This can be perfect for a tourist with refined taste on art.

Cluster 6 : this is the cluster maybe more indicated for college students, in general for young people. We can say this because of the presence of indie movie theaters, science museums, theaters, monuments/landmarks, comic shops, history museums, bookstores, comedy clubs.

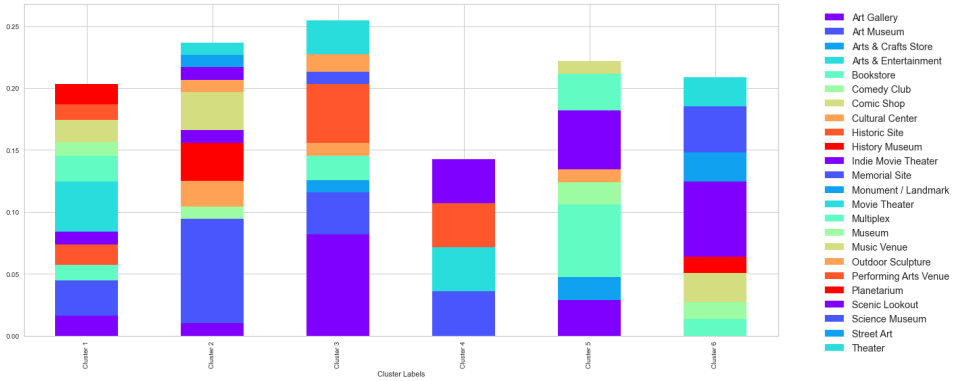


Figure 8: Distribution of art venues in each cluster.

Conclusions

We can sum up the results listed in the previous section in a more readable table, in which we label every cluster with a name able to describe it. We can read it in fig. 9.

	Cluster	Arrondissement	Label
0	1	10th/11th/12th/13th/14th/16th	The dynamic tourist
1	2	7th/18th/20th	The outdoor tourist
2	3	1st/2nd/3th/4th/8th	The history passionate tourist
3	4	19th	The photographer tourist
4	5	9th/15th/17th	The refined tourist
5	6	5th/6th	The young tourist

Figure 9: Pandas dataframe of the six clusters associated with arrondissements and labels.

So, a tourist who acknowledge his affinity in one of the six labels written above, can spend his vacation time visiting the corresponding arrondissement(s).

Of course these work has not the presumption of being thorough; a lot of other aspects are involved in the choice of visiting an arrondissement: the presence of suitable hotels and restaurants or the type of art venues, i.e. in this report it has not been marked a difference between contemporary or romantic art, for example. This is left to future tasks.