Capstone Project

Exploring Paris neighborhoods

Choosing location for a hotel

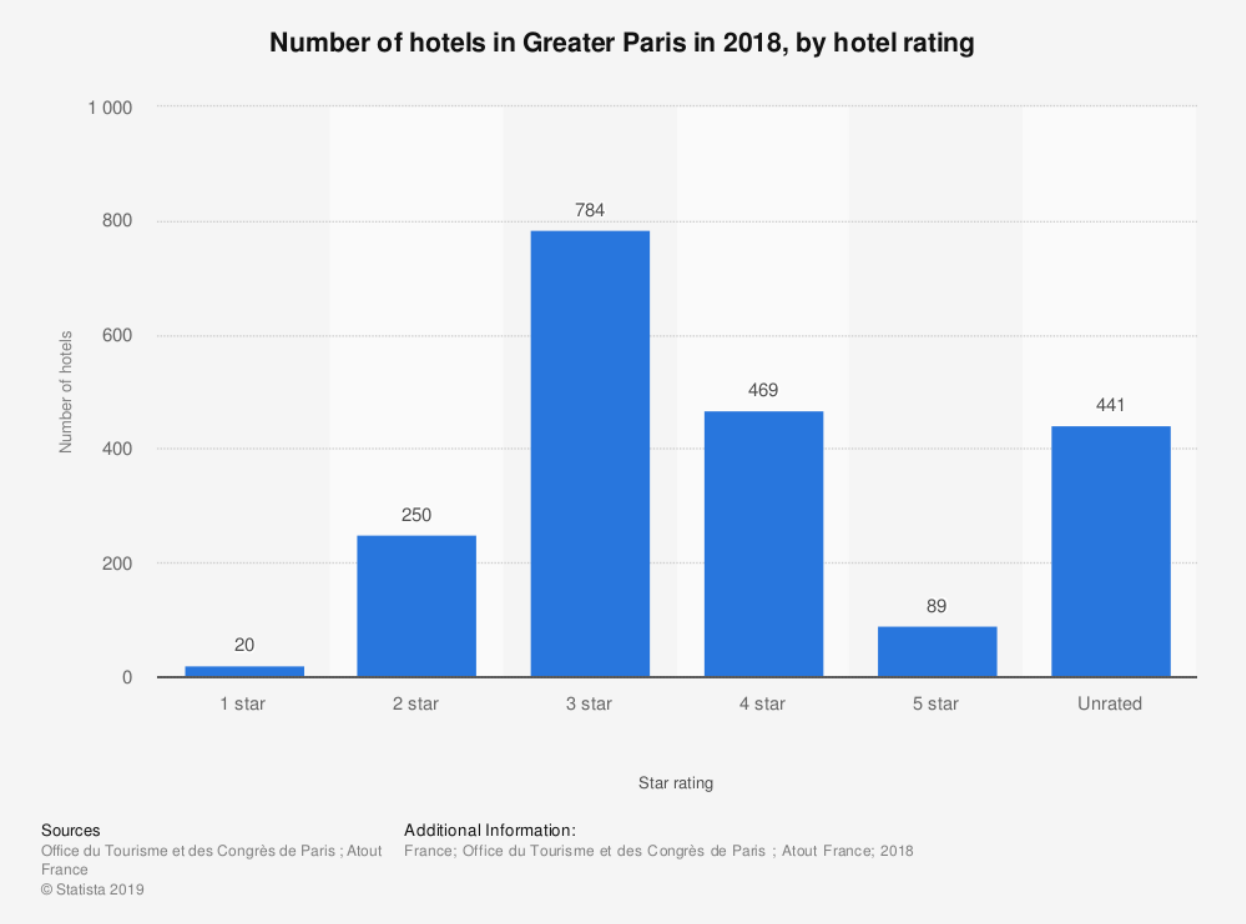
Stefanskaya Anna

June, 2020

1. Introduction

1.1 Background

Paris is the biggest city in France with population over 2 million people. It is very popular among tourists all over the world. Paris received 17.5 million visitors in 2018, 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 sixth most visited travel destination in the world in 2018, after Hong Kong, Bangkok, London, Macao and Singapore [1]. There is a great variety of accommodations to stay in for tourists, such as hotels, hostels, Airbnb and so on. There were more than 2,000 hotels in Paris in 2018 [2].



1.2 Problem

Usually when choosing the best hotel for our next vacation we consider many factors such as hotel rating, price, accommodations, location, meals and of course nearby places. Depending on our preferences and the place of staying we want some special venues to be in a walk distance from our hotel. For example, if we go to Paris, we may want to try French food and to explore local bars, or sometimes it is more important to stay in a walk distance from museums and exhibitions, or we may prefer our hotel to be near large shopping centers. However, online hotel search services usually don't have such search criteria where we can choose nearby places that we prefer.

1.3 Interest

The algorithm of neighborhoods clusterisation based on nearby places may be used personally while choosing the best place for staying during vacation or business trip. In business it may be used by online hotel search services to improve their search criteria for their users.

2. Data acquisition and cleaning

2.1 Data sources

First of all to explore boroughs and neighborhoods of Paris we need corresponding location data. We may find France boroughs' information here [3] and Paris neighborhoods' information here [4]. For getting nearby venues for each neighborhood we will use Foursquare API.

2.2 Data cleaning and preparation

France data is a text file that contains a lot of information about France regions. First of all we should choose the features that we are interested in and convert it into a dataframe. We will get the table with shape (1300, 10), where 1300 – is a number of boroughs in France. As we need only Paris boroughs we sort our dataframe by choosing only Paris city. After deleting unnecessary rows we get the table with shape (20, 10) – which is correct, as Paris has 20 boroughs.

Neighborhoods' data is in correct shape initially and has an information about all 80 neighborhoods in Paris. So we need only choose the features we need from text file and convert them into a dataframe. Then we merge two tables based on boroughs' postal codes, which are the same in both dataframes. Finally, we drop columns we don't need and rename other columns for better understanding of the data. In a resulting dataset we have the following information: neighborhood's name, surface and perimeter, latitude and longitude, number and name of borough, postal code and borough's population. The head of the final dataset:

	neighborhood	surface	perimetre	lon	lat	borough	postal_code	borough_name	population
0	Notre-Dame-des-Champs	8.613070e+05	4559.989773	2.327357	48.846428	06	75006	PARIS-6E-ARRONDISSEMENT	43.1
1	Petit-Montrouge	1.345774e+06	5490.636672	2.326437	48.826653	14	75014	PARIS-14E-ARRONDISSEMENT	137.2
2	Pont-de-Flandre	2.376238e+06	6397.871676	2.384777	48.895556	19	75019	PARIS-19E-ARRONDISSEMENT	184.8
3	Muette	5.477898e+06	11962.438594	2.259936	48.863275	16	75016	PARIS-16E-ARRONDISSEMENT	169.4
4	Chaillot	1.424035e+06	5207.046446	2.291679	48.868434	16	75016	PARIS-16E-ARRONDISSEMENT	169.4

Venues data we will get using explore function of Foursquare API. As a result of request 6529 venues were downloaded for 80 neighborhoods, which means that on average 81 venues were found for each neighborhood, given that API has a limit of 100 venues per request. We consider that it's a great number of data for the purposes of this study given the limitations. After transforming the results of the request into the dataframe, we get the following table (first 5 rows shown):

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Notre-Dame-des-Champs	48.846428	2.327357	Legend Hotel	48.845316	2.325507	Hotel
1	Notre-Dame-des-Champs	48.846428	2.327357	Gilles Verot	48.847118	2.326819	Deli / Bodega
2	Notre-Dame-des-Champs	48.846428	2.327357	Sadaharu Aoki | 青木定治	48.848013	2.330366	Dessert Shop
3	Notre-Dame-des-Champs	48.846428	2.327357	Marché de Raspail	48.848807	2.327526	Market
4	Notre-Dame-des-Champs	48.846428	2.327357	Bagels & Brownies	48.846537	2.327329	Bagel Shop

3. Methodology

3.1 Business understanding

The main goal of this project is to explore neighborhoods in Paris and to find the best place for a hotel according to our preferences of the nearby places required.

3.2 Analytic Approach

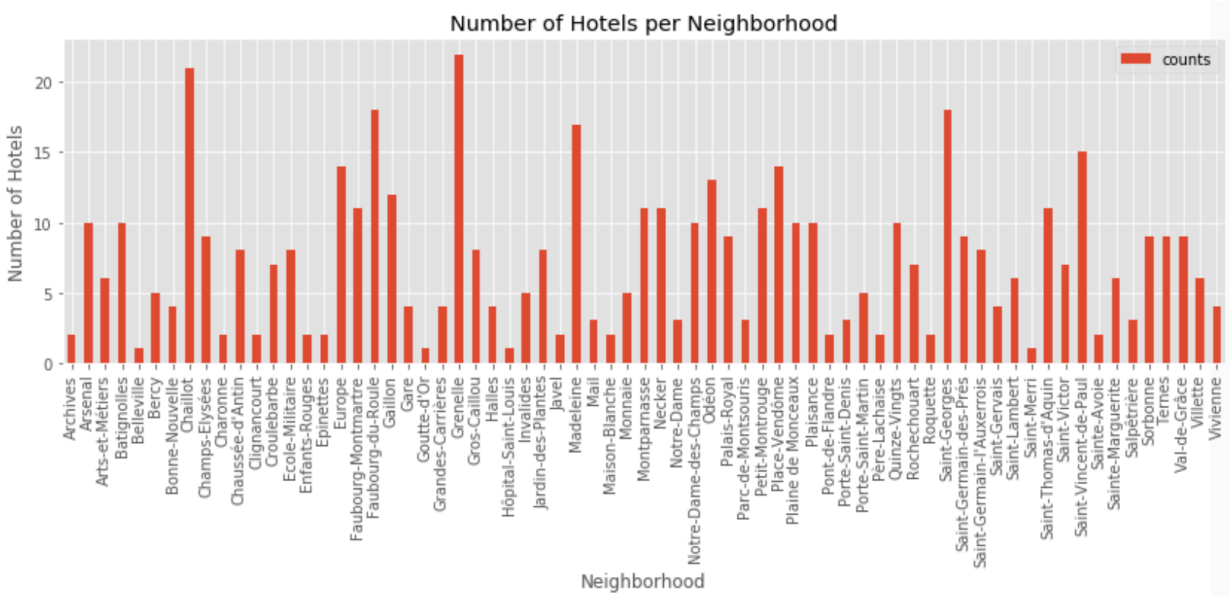
Paris has a total of 20 boroughs and 80 neighborhoods. In this project neighborhoods will be clustered following exploratory data that will be discovered in the next part.

3.3 Exploratory Data Analysis

To better understand the data and to check if the data is correct we should locate neighborhoods on a Paris map using Folium maps.

The dataset is correct and we can see all the Paris neighborhoods in a map.

If we want to explore neighborhoods for choosing the best hotel location we should first check if all the neighborhoods have hotels. Total hotels downloaded from Foursquare API is 503, which are located in 69 neighborhoods. Thus we should exclude 11 neighborhoods from our further analysis. As we can see from the graph below, the greatest number of hotels obtained per neighborhood is in Chaillot and Grenelle neighborhoods.



As we are using categorical variable (venue category) for analysis, we should first transform data into dummy variables using one hot encoding. Then we group rows by

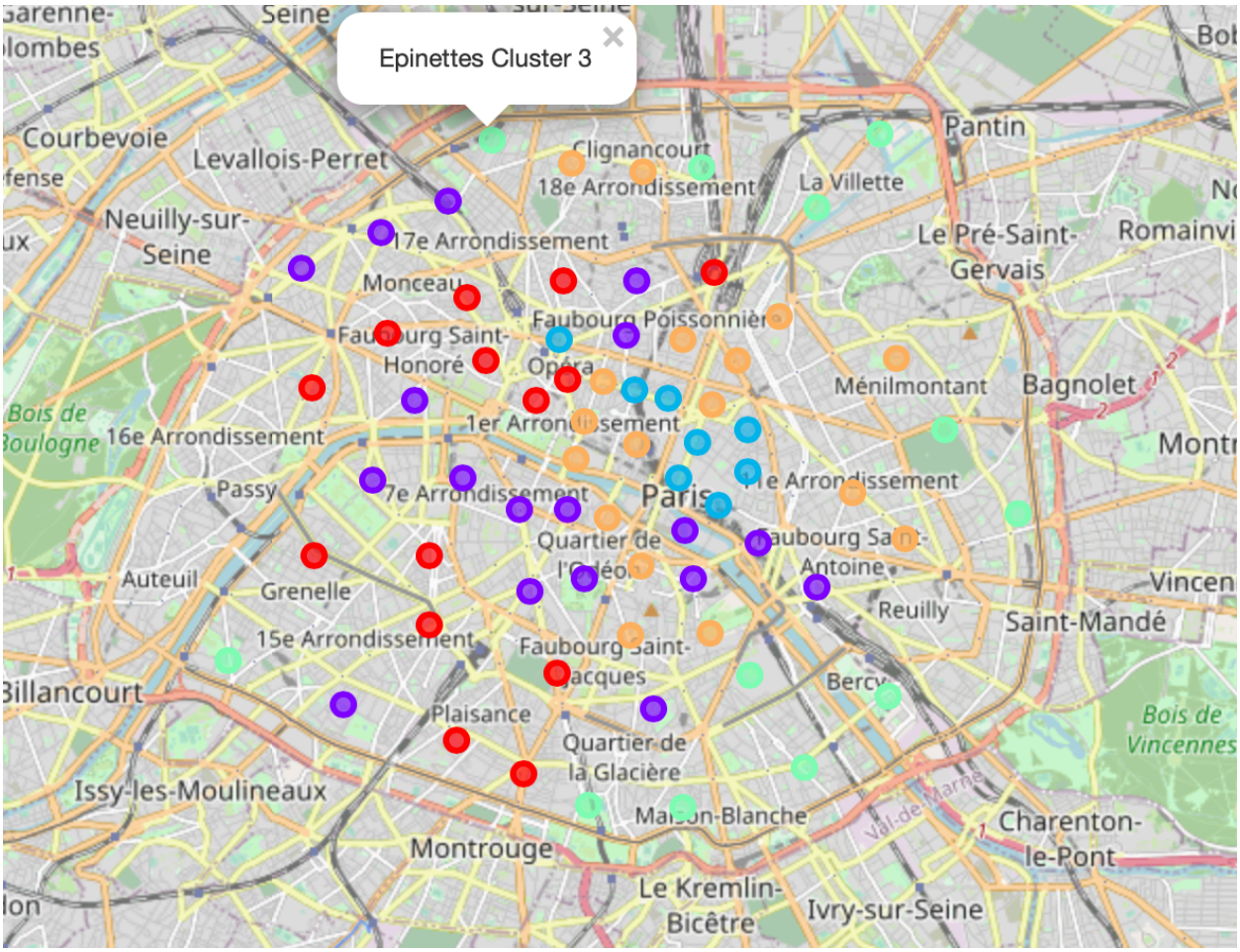
neighborhood and by taking the mean of the frequency of occurrence of each unique venue category. After doing so we get 345 unique venue categories for 80 neighborhoods. We will perform further neighborhood segmentation based on top-10 most common venues in each neighborhood.

3.4 Machine Learning Algorithm

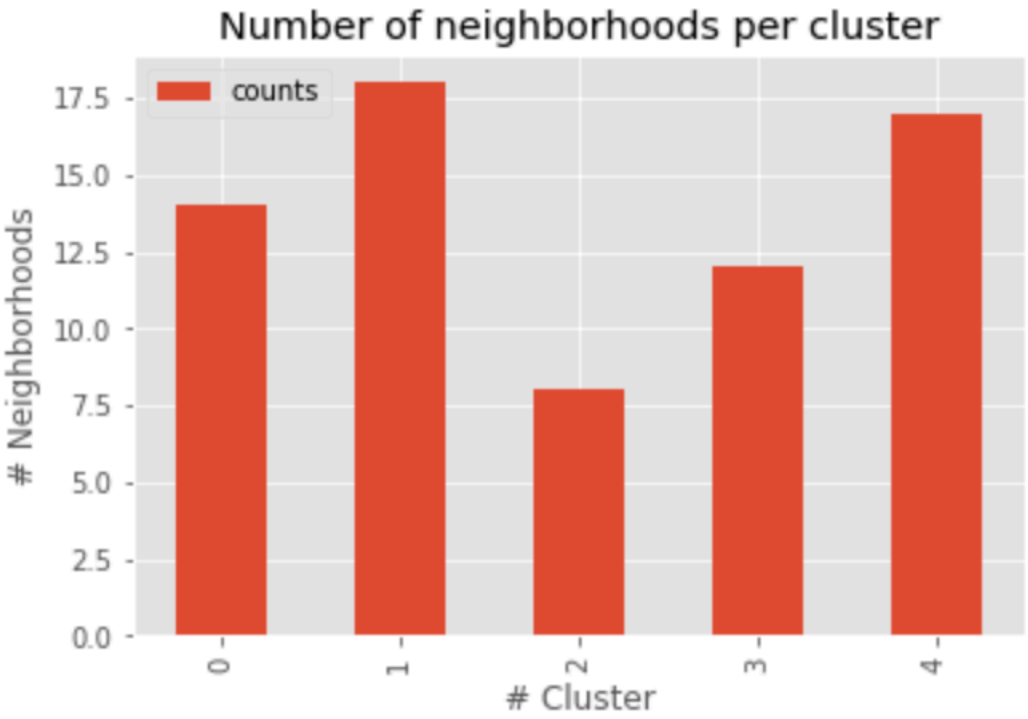
We will use the *k*-means clustering algorithm to segment the neighborhoods based on nearby venues categories. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes [5]. We will use k=5 to obtain 5 clusters from 69 neighborhoods.

4. Results

After performing cluster segmentation based on top-10 most common venues in each neighborhood we got 5 clusters. We use Folium maps to visualize obtained clusters on a Paris map.



We can see from the map that number of neighborhoods per cluster are not the same. Let's confirm it by plotting the bar chart:



The greatest number of neighborhoods is in cluster 1 and cluster 4. Let's analyze each cluster in details.

Cluster 0 (red markers)

Neighborhoods located in different borough of Paris. First most common venues are Hotels and French restaurants. There are also Japanese and Italian restaurants, shops, cafes and some bars. We use word clouds to visualize most common places in a cluster.

Cluster 1 (purple markers)

Neighborhoods located in different borough of Paris, usually near Cluster 0. First most common venue is French restaurants, second most common venue is Hotel. This cluster is very similar to Cluster 0, but it also has many Bakeries, some galleries and museums, parks and wine places.

Cluster 2 (blue markers)

This cluster consists only of 8 neighborhoods and 4 boroughs, that are located closely together in the geographical center. 1st most common venues are: French restaurants, cocktail bars, clothing stores and art galleries.

Cluster 3 (bright green markers)

Neighborhoods in this cluster are located far from geographical center of Paris and far from the most of touristic places. 1st most common venues are French restaurants, bike rentals, hotels, bars, grocery stores, bakeries, Japanese restaurants.

Cluster 4 (orange markers)

Neighborhoods in this cluster are located mainly in the city center. The most common venues are French restaurants, coffee shops, bars and Japanese restaurants. However there are also many places with Italian food, wine places, some shops and museums.

5. Discussion

As we can see on the above map, for some clusters there is a dependence between most common venues and distance from the city center. For example, neighborhoods from Cluster 3 are located far from center and neighborhoods from Cluster 2 (blue markers) are all located nearby in a geographical center of Paris.

So, let's model a situation when you and you partner are choosing the best neighborhoods to book a hotel for your next vacation. For example, you both want the hotel to be in the geographical center not far from the main touristic places – then you should exclude hotels cluster 3 immediately. You also want to try some local French food, but you like Italian as well. Your partner like cocktails, but you prefer wine – Cluster 0 is

out. After lunch you thinking of visiting some museums and art galleries, and your partner wants to go for some shopping – hotels in Cluster 2 is the best choice for you.

6. Conclusion

Based on your preferences on what places you like to visit during your vacation and with the help of machine learning algorithms you may choose the best suitable hotel in any city.

Therefore, using clustering algorithm based on nearby places you can analyze neighborhoods for many different purposes, for example it may help you to choose the location for your new business or exploring the boroughs for buying or renting an apartment.

7. References

[1] https://en.wikipedia.org/wiki/Paris
[2] https://www.statista.com/statistics/630596/number-of-hotels-iin-greater-paris-by-hotel-rating/
[3] https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e
[4] https://opendata.paris.fr/explore/dataset/quartier_paris/download/?format=json&timezone=Europe/Berlin')
[5] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1