Coursera Capstone

IBM Applied Data Science Capstone

The Battle of Neighborhoods

Opening a New Hotel in Paris, France

By Racha Khairallah
September 2019

Contents

1- Business Problem

For many tourists, visiting Paris is a great way to relax and enjoy themselves during weekends and holidays. They can do shopping, dine at restaurants, shop at the various fashion outlets, watch movies , visit many historical sites and organize many others activities.

Paris is one of the busiest cities in Europe, appear as the most visited city around the world with 17.95 million tourists visitors in 2018.

Building hotels allow property developers to earn consistent rental income. Of course, as with any business decision, opening a new hotel requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the hotel is one of the most important decisions that will determine whether the hotel will be a success or a failure.

By using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the following business question:

In the city of Paris, France, if a property developer is looking to open a new hotel, where would you recommend that they open it?

This project is particularly useful to property developers and investors looking to open or invest in new hotel in the capital city of France, Paris.

2- Data

To solve the problem, we will need the following data:

• List of neighborhoods in Paris.
• Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
• Venue data, particularly data related to hotels. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

1- We will use web scraping techniques to extract the data from the Wikipedia page, (https://en.wikipedia.org/wiki/Category:Districts_of_Paris) with the help of Python requests and beautifulsoup packages.

2- We will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

3- We will use Foursquare API to get the venue data for those neighborhoods**.**

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

3- Methodology

Firstly, we need to get the list of neighborhoods in the city of Paris. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Districts_of_Paris). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Paris.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return

the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Hotels” data, we will filter the “Hotels” as venue category for the neighborhoods.

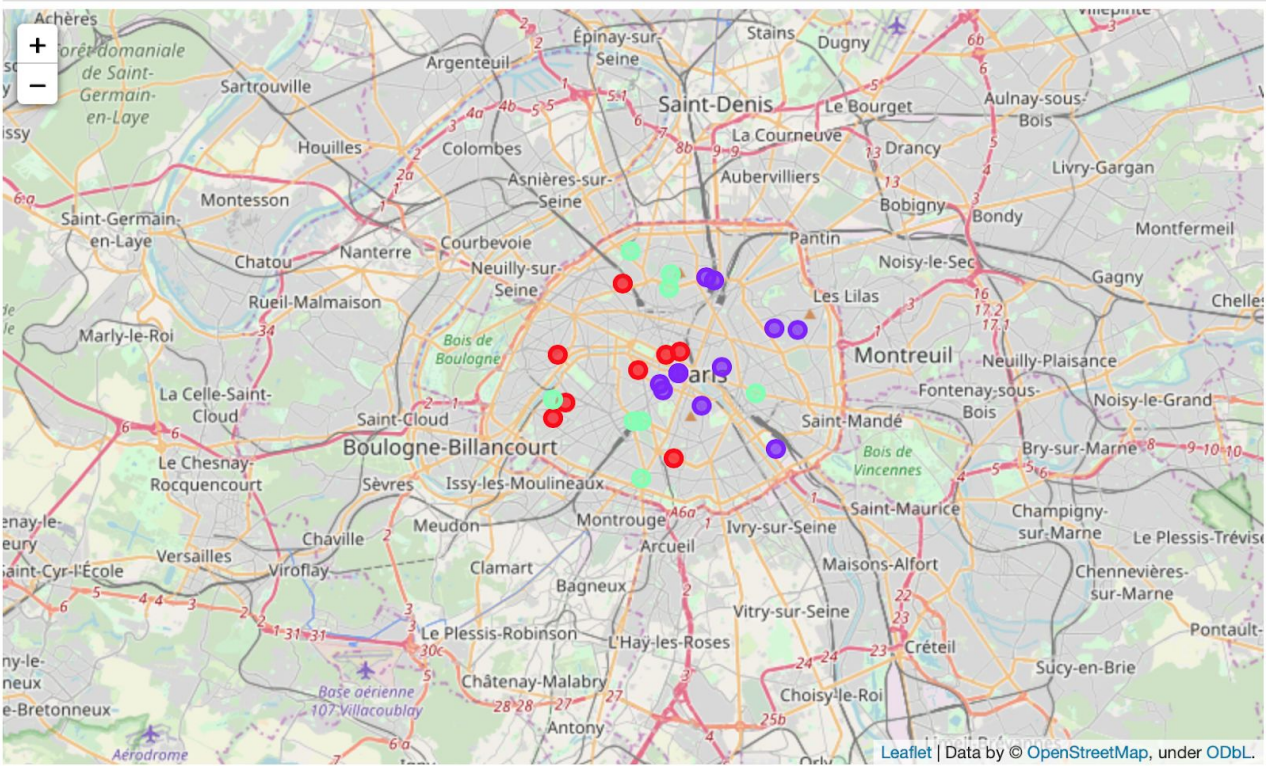
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Hotels”. The results will allow us to identify which neighborhoods have higher concentration of hotels while which neighborhoods have fewer number of hotels. Based on the occurrence of hotels in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new hotels.

4- Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Hotels”:

- Cluster 0: Neighborhoods with low number of Hotels
- Cluster 1: Neighbourhoods with moderate number of Hotels
- Cluster 2: Neighbourhoods with low number of hotels

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour**.**



5- Discussion

A very small number of hotels are concentrated in the central area of Paris city. As showing the results of cluster 0 and 2 with low number of Hotels and cluster 2 with moderate number of Hotels. This represents a great opportunity and high potential areas to open new hotel as there is very little to no competition from existing hotels.

This project recommends property developers to capitalize on these findings to open new hotel in neighborhoods in cluster 0 or cluster 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new hotel in neighborhoods in cluster 1 with moderate competition.

6- Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing

recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new hotel. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 and 2 are the most preferred locations to open a new hotel. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new hotel.

7- References:

Category:Districts_of_Paris: https://en.wikipedia.org/wiki/Category:Districts_of_Paris

Foursquare Developers Documentation: https://developer.foursquare.com/docs

8- Appendix:

Cluster 0:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
0	Batignolles	0.14	0	48.883330	2.316670
18	Paris Rive Gauche	0.10	0	48.832060	2.339660
29	Trocadéro	0.16	0	48.862340	2.288070
14	Les Halles	0.10	0	48.863190	2.342010
12	Javel, France	0.11	0	48.843870	2.286130
9	Grenelle	0.11	0	48.848420	2.291640
11	Jardin de Tivoli, Paris	0.13	0	48.862586	2.335726
6	Faubourg Saint-Germain	0.10	0	48.857807	2.323826

Cluster 1:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
25	Quartier des Grandes-Carrières	0.00	1	43.760100	5.750995
8	Goutte d'Or	0.03	1	48.885010	2.354130
26	Revolutionary sections of Paris	0.02	1	48.857170	2.341400
10	Historical quarters of Paris	0.02	1	48.857170	2.341400
21	Quarters of Paris	0.02	1	48.857170	2.341400
27	Saint-Germain-des-Prés	0.02	1	48.853770	2.333310
13	Latin Quarter, Paris	0.01	1	48.847750	2.351800
28	The Marais	0.02	1	48.858760	2.360870
2	Bercy	0.04	1	48.834880	2.384590
24	Quartier de La Chapelle	0.02	1	48.884150	2.357093
17	Ménilmontant	0.01	1	48.869881	2.394115
1	Belleville, Paris	0.00	1	48.870180	2.384230
19	Passy	0.00	1	46.540550	4.540730
22	Quartier Asiatique	0.03	1	48.852030	2.334490

Cluster 2:

:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
23	Quartier Pigalle	0.070000	2	48.882031	2.337576
15	Montmartre	0.060000	2	48.886150	2.337970
16	Montparnasse	0.080000	2	48.843130	2.321290
7	Front de Seine	0.080000	2	48.849310	2.285780
5	Faubourg Saint-Antoine	0.050000	2	48.850940	2.375670
4	Cour des miracles	0.074074	2	45.748650	4.881890
3	Cité des Fleurs	0.060000	2	48.892612	2.320326
20	Petit-Montrouge	0.090000	2	48.826420	2.325200
30	Épinettes	0.070000	2	48.842962	2.325298