PARIS'S BATTLE OF NEIGHBORHOODS

IBM Applied Data Science Capstone

Submitted By: Youssra Saadeddine

Introduction:

Paris, renowned as the city of light and love, is simply Europe's most enthralling capital. Beyond the boulevards and classical monuments, you'll find fascinating small museums, scores of family-run hotels and neighborhoods filled with hotel, food and drink shops, boutiques and bakeries.

Tourism in Paris is a major income source. In 2018, 17.95 million international, overnighting tourists visited the city, mainly for sightseeing and shopping (and estimated to be well over double if including domestic overnighting visitors).

Having a hotel at Paris is absolutely a good project but before having this you should do an analysis of location.

Business Problem:

The objective of this capstone project is to analyze and select the best locations in Paris to open a hotel. Using data science methodology and machine learning techniques like clustering.

if someone is looking to open a new hotel in the city of amour, where would you recommend that they open it?

Target Audience of this project

This project is useful to property developers or investors looking to open a hotel in Paris.

DataSet:

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 in Paris. We will use this data to perform clustering on the neighborhoods.

Sources of data

This Wikipedia page (https://en.wikipedia.org/wiki/Quarters_of_Paris) contains a list of neighborhoods in Paris.

Methodology

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).

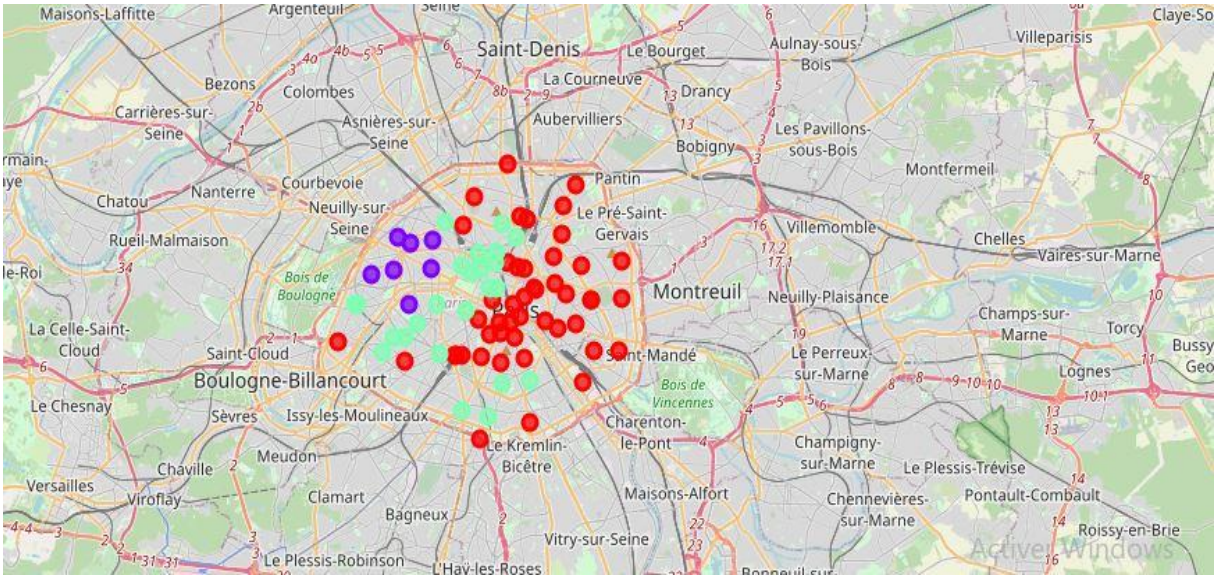
First, we will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and Beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.
Next, we will us the Foursquare API to get the top 100 venues and we need to register a Foursquare Developer Account in order to obtain the Foursquare Id and Foursquare secrete key. We then make an API calls to Foursquare passing the geographical coordinates of the neighborhood in a Python loop. Foursquare will return the venue data to the 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 these returned venues. Then, we will analyze 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 of use in clustering. Since we are analyzing the “Hotel” data, we will filter the “Hotel” as venue category for the neighborhoods.
Lastly, we will perform clustering on the data by using 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 particularly suited to solve the problem for this project.

We will cluster the neighborhoods in 3 cluster based on this frequency of occurrence for “Hotel”. The result will allow us to identify which neighborhood has fewer number of hotels.

Results

The results from the k-means clustering shows that we categorize the neighborhood into 3 clusters
based on the frequency of occurrence for “hotels”:

- Cluster 0: Neighborhoods with the high number of Hotels.
- Cluster 1: Neighborhoods with the low number of Hotels.
- Cluster 2: Neighborhoods with the moderate number of Hotels.



Discussion

most of the hotels are concentrated in the central area of Paris city, with the highest number in cluster 0 and moderate number in cluster 2, cluster 1 has no hotel in the neighborhood (central area).

Cluster 2 represent a great opportunity and high potential area to open new hotels as there is very little competition from existing hotels in **the west side**, Meanwhile, hotels in cluster 0 are likely suffering from the of intense competition in the central area of Pairs due to oversupply and high concentration of hotels ,but in the **Eastern side**, having an hotel in cluster 0 is the best choice.

From another perspective, the result also shows the oversupply of hotel mostly happened in the central area of the city, with the west area still have the few

hotel. Therefore, this project recommends property developers to capitalize on these new finding's tom open a new hotel in cluster 2 with the little to no competition.

Limitation and Suggestion for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of hotel, there are other factors such as population and tourism statistic which could influence the location decision of a new hotel.

In our data we have the population in each neighborhood so we can do another analysis to support our decision and make it clear.

In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with the limitation as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more result.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing and preparing the data, performing the machine learning by clustering the data into 3 cluster based on their similarities, and lastly providing recommendation to the relevant stakeholders i.e. property developers and investors regarding the best location to open a hotel. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhood in cluster 2 are the most prefer location to open a hotel. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential location while avoiding overcrowded areas in their decision to open a new hotel.