Finding the Best Hotel in Paris

Tristan Renaud

June 10, 2020

1 Introduction

1.1 Background

My wife and I would like to take a trip to Paris and need a hotel. This hotel needs to be well rated as well as within our budget and walking distance to our travel interests.

I tried using travel websites to solve this problem, but I struggled with narrowing down the hundreds of Parisian hotels. While I could filter by various attributes (rating, price, etc.), I could not filter by my travel interests. For example, I want to be within walking distance to cafés, pastry shops, museums and parks.

As a result, I decided to look to data science to help us with this problem. The problem will be generalized to work with anyone's travel interests.

1.2 Problem & Target Audience

What is the best Parisian hotel given a traveler's budget and travel interests? This project aims to answer this question using data science methods.

Specifically, we will use a combination of location data from Foursquare Place API, Yelp Fusion API and Nominatim.

To solve this problem, we expect a hotel or set of hotels that best align with a traveler's interests given a budget.

1.3 Target Audience

The target audience are travel companies. While travel companies may utilize personal experience and/or algorithms to offer up hotels, I find they do not necessarily consider what is in the immediate vicinity of every hotel in a city.

By integrating a customer's interests into the equation, they can improve their recommendations, in turn making customers happier and more likely to return for future travel.

2 Data acquisition and cleaning

2.1 Data Source

We will use the following data sources to solve our problem:

1. **Nominatim (GeoPy)** - We will use this to get the latitude and longitude values of Paris.
2. **Foursquare API** - We will use this to get hotel data (except pricing) and venues in proximity to each hotel.
3. **Yelp Fusion API** – Foursquare does not provide pricing data on hotels so we will use Yelp to acquire this information. Since we need to access this source, additional data will be brought in to strengthen our analysis.

2.2 Data Description

The data falls into three categories: coordinates, hotel data and venues in proximity to each hotel. This section summarizes these data and the table below provides a summary.

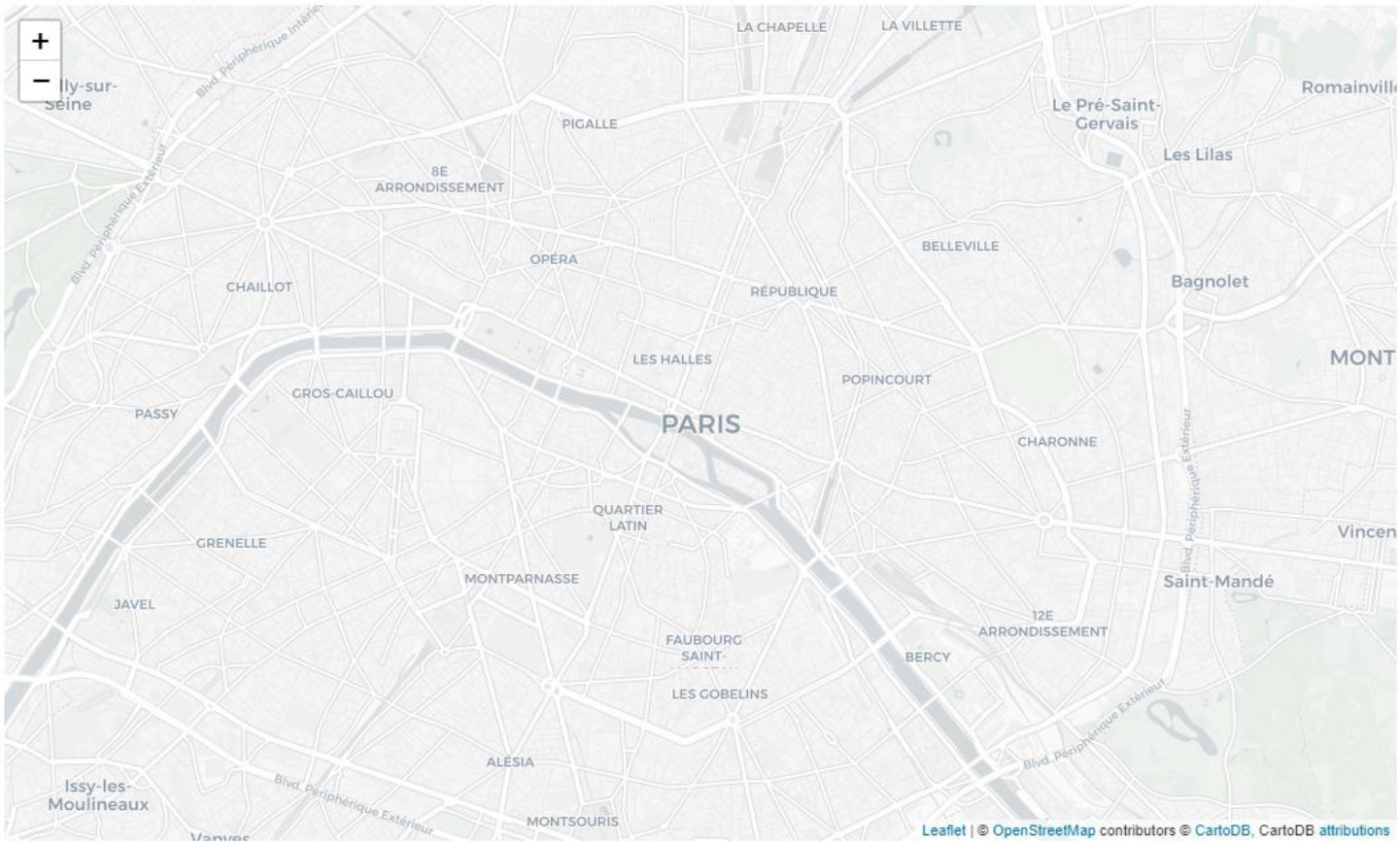
2.2.1 Latitude and Longitude value of Paris.

The geographical coordinates of Paris will be used to center maps (using Folium) and may be used as a parameter to determine a hotel's closeness to the city center.

Map example:

```
paris_map = folium.Map(location=[48.8566969, 2.3514616], zoom_start=13,tiles='CartoDB positron')

paris_map
```



2.2.2 Parisian hotel data

Hotel data is sourced from Foursquare and Yelp. I chose to include Yelp in this analysis because it has pricing information on hotels whereas Foursquare does not.

A combination of hotel information is included, such as ID (uniquifier), address, phone number and URL. This information will keep the data clean and allow others to easily research the hotel and make a reservation.

Summary of Hotel Data:

Source	Attributes	Metrics
Foursquare	id, name, formattedAddress, url, formattedPhone, address, postalCode, cc, city, state, country	latitude, longitude, verified, rating, ratingSignals, tipCount, count
Yelp	yelp.id, yelp.name	yelp.price, yelp.rating, yelp.review_count, yelp.is_claimed

Sample of hotel data:

```
df.T[[0,1]]
```

	0	1
id	4adcda00f964a520ba3021e3	4b817458f964a52024a730e3
name	Hôtel Four Seasons George V	Hôtel Plaza Athénée
latitude	48.8688	48.8662
longitude	2.30068	2.30437
formattedAddress	[31 avenue George V, 75008 Paris, France]	[25 avenue Montaigne (Rue Clément Marot), 7500...
verified	True	False
url	https://www.fourseasons.com/paris/	https://www.dorchestercollection.com/en/paris/...
rating	9.5	9.4
ratingSignals	912	496
formattedPhone	+33 1 49 52 70 00	+33 1 53 67 66 65
address	31 avenue George V	25 avenue Montaigne
postalCode	75008	75008
cc	FR	FR
city	Paris	Paris
state	Île-de-France	Île-de-France
country	France	France
tipCount	156	99
count	776	399
yelp.id	2d_5q0cr_bAA3yF38gpzrw	KlyztuvAktI7mXgn9Lqofw
yelp.name	Four Seasons Hôtel George V	Hôtel Plaza Athénée
yelp.price	4	4
yelp.rating	4.5	4.5
yelp.review_count	105	43
yelp.is_claimed	True	True

2.2.3 Venues within "walking distance" to each hotel.

For each hotel in the data, up to 100 nearby venues were pulled using Foursquare within a radius of 500 meters (0.3 miles) of each hotel. Each row includes the hotel and its latitude/longitude as well as a nearby venue's name, latitude/longitude and venue category.

Sample of venues near hotel:

```
hotel_venues.head(10)
```

	Hotel Name	Hotel Latitude	Hotel Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Hôtel Four Seasons George V	48.868849	2.300683	Hôtel Four Seasons George V	48.868849	2.300683	Hotel
1	Hôtel Four Seasons George V	48.868849	2.300683	Le Cinq	48.868798	2.300565	French Restaurant
2	Hôtel Four Seasons George V	48.868849	2.300683	Hôtel Prince de Galles	48.869096	2.300684	Hotel
3	Hôtel Four Seasons George V	48.868849	2.300683	The Spa	48.868283	2.300777	Spa
4	Hôtel Four Seasons George V	48.868849	2.300683	Le 39V	48.869834	2.300425	French Restaurant
5	Hôtel Four Seasons George V	48.868849	2.300683	Hôtel François 1er	48.870176	2.299917	Hotel
6	Hôtel Four Seasons George V	48.868849	2.300683	Le Bar	48.868727	2.301072	Hotel Bar
7	Hôtel Four Seasons George V	48.868849	2.300683	Hôtel Barrière Le Fouquet's	48.871035	2.301653	Hotel
8	Hôtel Four Seasons George V	48.868849	2.300683	Pret A Manger	48.870020	2.298433	Sandwich Place
9	Hôtel Four Seasons George V	48.868849	2.300683	Pret A Manger	48.868442	2.303393	Sandwich Place

2.3 Data Cleaning

2.3.1 Hotel Data

Foursquare identified 402 hotels in Paris, which were then crossmatched with Yelp. Any hotels that did not have a Yelp counterpart OR had neither a Foursquare nor Yelp rating were dropped.

This reduced the dataset down to 213 hotels.

Afterwards, upon closer inspection, I found that there were some rows that had the same address as well as some rows with the same Yelp ID. These were deduped, keeping the higher result (assuming Foursquare and Yelp provide more relevant/accurate results first).

The final dataset includes 208 hotels.

It is worth noting that some hotels did not have Paris as the city. After spot-checking a few, they all appeared to be within the Paris metropolitan area (https://en.wikipedia.org/wiki/Paris_metropolitan_area).

2.3.2 Nearby Venue Data

While the dataset does not have missing values, I noticed that the venue category 'hotel' consisted of about 10% of the venues.

I contemplated removing all hotel venues from this dataset since this dilutes the other venue categories which are associated with travel interests. However, I decided to keep these hotel venues because they make up a given hotel's neighborhood and may lead to deceptive results.

For example, suppose 50 out of 100 of the nearby venues for a given hotel are other hotels. As a tourist, I would want to know this, and I may choose to avoid areas with a high density of hotels.