Airbnb Listings in Paris

NGUYEN Do Hien Anh

Dec 2019

A project for **IBM Data Science Professional Certificate**

I. Introduction

1. Background

Just a few years ago, accommodations for traveling were largely limited to hotel, hostel, or the likes. However, homestay is gaining popularity these days with the introduction of intermediaries that connect travelers and hosts, the most notable being Airbnb. The company was founded in 2008 with a few listings in the U.S. Since then, it has expanded globally to over 190 countries and regions with more than seven million listings worldwide (Airbnb, 2019). Many travelers, especially young travelers, prefer Airbnb for its flexibility and various options regarding price, place, type of accommodation, etc. Even business travelers are starting to choose Airbnb: the number of companies using Airbnb had increased from only 250 in 2015 to over 250,000 in 2017 (Zaleski, 2017). Staying with Airbnb is a popular choice, thus knowing what listings to book is a common interest among many when visiting a new place.

Paris is one of the most visited capital in the world. Also known as the City of Light, the city prides itself on numerous monuments and architectures, from the Eiffel tower to Louvre palace, from Sacré-Cœur church to Arc de Triomphe. Beside the iconic places, Paris also offers great things to explore, from the museums and art galleries to the opera houses, from the diverse tastes of food and drinks to the bursting nightlife with bars and clubs. There are things to do no matter where one stays in Paris, but the venues are not evenly distributed, so selecting where to stay is a concern if one wants to be near certain types of venues.

2. Problem

This project was inspired by my deep-rooted wish to visit Paris. The first thing that came to my mind was where to stay and for how much. I love exploring the life of a local, so Airbnb is a great choice. I also want to stay within walking distance of restaurants, coffee houses, and big plazas to walk around and mingle to my heart's content. I was going to do a project to visualize on the map, for each quarter of Paris districts, the price ranges of Airbnb listings as well as what types of entertainment the quarter has.

3. Interest

People, especially young travelers, who want to explore Paris could be interested in this project.

II. Data

For this project, I used two (02) sources of public data and Foursquare API to get all data.

The first public source is for data of the price ranges of Airbnb listing in Paris. I could go on Airbnb's website and crawl all the publicly available data there, including price per night, listing location, condition, host's information, etc. But there is a great website called Inside Airbnb (https://insideairbnb.com/) that already did all of that. Data is updated monthly, and since I did this project in early August, I used data collected in July 09th, 2019. There are five (05) files available including the calendar, summary and details of listings, summary and details of reviews for all listings. In this website, there are also another two (02) files, one is the list of all Paris districts (here called neighborhood), and the other is the geoJSON file of the districts. For this project, I used the geoJSON file and the summary of listings, with one (01) line for each listing (each line with a unique listing ID). There are 16 fields including listing ID, name, type, price per night, latitude and longitude, neighborhood (district the listing belongs to), etc.

The second public source is for data of Paris quarters: correct names, latitude and longitude, which I got from Paris Data (https://opendata.paris.fr/explore/dataset/quartier_paris/information/), an open database created by the government of Paris city.

The Foursquare API is an application programming interface that allows interaction with the **Foursquare** platform (https://foursquare.com/) - which, in short, is like Yelp for entertainment. In details, Foursquare aggregates, by location, data of numerous venues including restaurants, museums, shopping malls, etc. The data available on Foursquare includes the basic information such as venue's name, category, location latitude and longitude, as well as the details such as menu, photos of the interior and products, user rating and review, reservation options and payment methods, etc.

III. Methodology

1. Data cleansing & transformation

a. For data of Airbnb listings

I fetched data from Inside Airbnb to data frame ***sum_listing*** with 16 columns and 64,293 rows.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
0	3109	zen and calm	3631	Anne	nan	Observatoire	48.8335	2.3185	Entire home/apt	60	3	8
1	5396	Explore the heart of old Paris	7903	Borzou	nan	Hôtel-de-Ville	48.8510	2.3587	Entire home/apt	115	1	180
2	7397	MARAIS - 2ROOMS APT - 2/4 PEOPLE	2626	Franck	nan	Hôtel-de-Ville	48.8576	2.3527	Entire home/apt	115	10	248
3	7964	Large & sunny flat with balcony !	22155	Anaïs	nan	Opéra	48.8746	2.3434	Entire home/apt	130	6	6
4	8522	GREAT FLAT w/ CITY VIEW	25435	Adrien	nan	Ménilmontant	48.8653	2.3933	Entire home/apt	90	3	1

Exploratory analysis showed that column *neighbourhood_group* was blank, so I removed it. Pivoting for minimum number of nights required to book revealed that there were multiple minimum requirements, yet I was only looking for short stays, so listings that needed more than seven (07) days of booking were also removed. It turned out most listings were good for short stays as more than 90% of them remained, my data frame now had 15 columns and 59,989 rows.

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_
0	3109	zen and calm	3631	Anne	Observatoire	48.8335	2.3185	Entire home/apt	60	3	8	2018-12-19	
1	5396	Explore the heart of old Paris	7903	Borzou	Hôtel-de-Ville	48.8510	2.3587	Entire home/apt	115	1	180	2019-06-21	
2	7397	MARAIS - 2ROOMS APT - 2/4 PEOPLE	2626	Franck	Hôtel-de-Ville	48.8576	2.3527	Entire home/apt	115	10	248	2019-06-25	
3	7964	Large & sunny flat with balcony !	22155	Anaïs	Opéra	48.8746	2.3434	Entire home/apt	130	6	6	2015-09-14	
4	8522	GREAT FLAT w/ CITY VIEW	25435	Adrien	Ménilmontant	48.8653	2.3933	Entire home/apt	90	3	1	2010-06-16	

Moving forward, I looked at the types of listings. There are three (03) listing types, identified in column *room_type*, which are: Entire home/apt, Private room, Shared room. Looking at the descriptive statistics of listings based on types, I saw that the prices shot up unusually high: the 75 percentile were only €50 - €126 per night, but the maximum values were from €1,000 to €9,379.

	price							
	count	mean	std	min	25%	50%	75%	max
room_type								
Entire home/apt	51,651.0000	118.0318	152.5522	0.0000	65.0000	89.0000	126.0000	9,379.0000
Private room	7,893.0000	116.4749	207.3384	8.0000	40.0000	59.0000	90.0000	5,000.0000
Shared room	445.0000	53.2180	97.7920	10.0000	25.0000	32.0000	50.0000	1,000.0000

Taking a closer look and sorting the the prices in descending order, there were only 9 listings with price per night above €5,000 and a little under 500 listings with price per night above €1,000. These listings accounted for less than 1% of listings, yet the values were high enough to skew the analysis, so I removed them as well. The final data frame, called ***inscope_listing*** had 15 columns and 59,518 rows.

With the final data frame for Airbnb listings, I calculated the average price per night for each district and checked the descriptive statistics. To get ready for visualization later, I stored this information in a new data frame called ***df_listing***.

	neighbourhood	number of listing	average price per night
0	Batignolles-Monceau	4066	100.6849
1	Bourse	1937	137.5581
2	Buttes-Chaumont	3444	74.9059
3	Buttes-Montmartre	6845	83.1445
4	Entrepôt	4205	99.3410
5	Gobelins	2173	80.1146

b. For data of Paris' quarters

The latitude and longitude of Paris' quarters were scraped from raw data in Paris Data as below.

	datasetid	fields	geometry	record_timestamp	recordid
0	quartier_paris	{'n_sq_qu': 750000014, 'perimetre': 2678.34092...	{'type': 'Point', 'coordinates': [2.3581623338...	2019-09-01T17:00:34.240+07:00	df2bb3a2f910c84c0b2117e82b1a50d492565d05
1	quartier_paris	{'n_sq_qu': 750000034, 'perimetre': 3133.58009...	{'type': 'Point', 'coordinates': [2.3322688688...	2019-09-01T17:00:34.240+07:00	10ad04c11f3966a5b537988cb88fb60d56909638
2	quartier_paris	{'n_sq_qu': 750000042, 'perimetre': 4052.56773...	{'type': 'Point', 'coordinates': [2.3761180559...	2019-09-01T17:00:34.240+07:00	a0e67334ab550e84ad3a951530f00722fd971178
3	quartier_paris	{'n_sq_qu': 750000058, 'perimetre': 5979.71146...	{'type': 'Point', 'coordinates': [2.3107774536...	2019-09-01T17:00:34.240+07:00	b09d7f570d6bf273458362fc925d308d19991c68
4	quartier_paris	{'n_sq_qu': 750000012, 'perimetre': 1861.80411...	{'type': 'Point', 'coordinates': [2.3548515182...	2019-09-01T17:00:34.240+07:00	9b4b8f64d46994bda1a666d41824000700eaff4b

After getting the data I need - basic information of all 80 quarters (04 quarters for each of 20 of Paris' districts), I stored this in ***paris_quarters***.

	quarter	latitude	longitude
0	Saint-Gervais	48.8557	2.3582
1	Chaussée-d'Antin	48.8735	2.3323
2	Saint-Ambroise	48.8623	2.3761
3	Necker	48.8427	2.3108
4	Sainte-Avoie	48.8626	2.3549

c. For data of venues in Paris

To get data of venues in Paris, I connected to the Foursquare API.

Starting by defining a function to fetch data, I collected for each quarter 75 venues within 1,500 meters from the quarter's identifying latitude and longitude. The data, stored on a data frame called ***paris_venues***, included the venue's name, category, location, and the quarter it belonged to.

	Quarter	Quarter Latitude	Quarter Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Saint-Gervais	48.8557	2.3582	Miznon	48.8572	2.3590	Israeli Restaurant
1	Saint-Gervais	48.8557	2.3582	Aux Merveilleux de Fred	48.8557	2.3564	Dessert Shop
2	Saint-Gervais	48.8557	2.3582	Les Mots à la Bouche	48.8577	2.3573	Bookstore
3	Saint-Gervais	48.8557	2.3582	L'As du Fallafel	48.8574	2.3591	Falafel Restaurant
4	Saint-Gervais	48.8557	2.3582	Le Peloton Café	48.8555	2.3560	Coffee Shop

There were 254 unique categories in this data frame. I then proceeded to encode the quarters' data to get information on mean of frequency of occurrence of each category (e.g. out of 75 venues fetched, how many percentages were venues in category "Art gallery", in "Used Bookstores", etc.) This data would be used for clustering later and was stored in ***paris_grouped***.

	Quarter	Accessories Store	African Restaurant	Alsatian Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Used Bookstore	Vegetarian / Vegan Restaurant	Vietnamese Restaurant
0	Amérique	0.0000	0.0133	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000	0.0000	0.0133
1	Archives	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0400	0.0267	0.0000	...	0.0000	0.0000	0.0133
2	Arsenal	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0133	0.0133	0.0000	...	0.0000	0.0133	0.0000
3	Arts-et-Métiers	0.0000	0.0000	0.0000	0.0000	0.0000	0.0133	0.0000	0.0133	0.0000	...	0.0000	0.0133	0.0000
4	Auteuil	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0133	0.0133	0.0000	...	0.0000	0.0000	0.0000

From this data frame, I also ranked the five (05) most common venue categories of a quarter and created data frame ***quarter_venues_sorted*** to store the data on the name of these 05 categories and number of venues in each category.

	Quarter	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Number of 1st Most Common Venue	Number of 2nd Most Common Venue	Number of 3rd Most Common Venue	Number of 4th Most Common Venue	Number of 5th Most Common Venue
0	Amérique	French Restaurant	Park	Restaurant	Café	Bar	11	4	4	4	4
1	Archives	French Restaurant	Café	Clothing Store	Wine Bar	Cocktail Bar	8	4	4	3	3
2	Arsenal	French Restaurant	Cocktail Bar	Hotel	Pizza Place	Coffee Shop	10	3	3	3	3
3	Arts-et-Métiers	Wine Bar	Italian Restaurant	Sandwich Place	Japanese Restaurant	Cocktail Bar	5	4	4	4	4
4	Auteuil	Tennis Court	French Restaurant	Italian Restaurant	Market	Japanese Restaurant	15	8	6	3	3

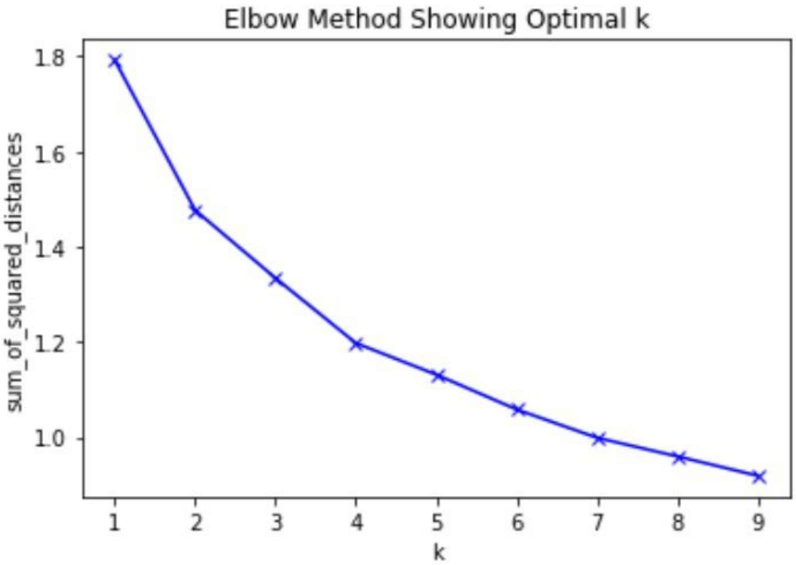
2. Data clustering with K-mean

a. Find the optimal K

I used the data in ***paris_grouped*** to cluster the quarters together with K-mean clustering - an unsupervised machine learning algorithm to divide a dataset into K clusters.

To find the optimal K, I used the elbow method, which took all K in a predefined range, clustered the data with each of the K, and calculated the sum of squared distances for each K value (sum of squared distances indicates errors from predicted values to actual values, so the smaller the better).

Since there were only a total of 80 quarters, I chose K range from 1 to 10 and plotted the results of elbow method on a graph as belows. The graph showed that optimal K was 4.



b. Fit model with optimal K found

I fitted the K-mean clustering model with K = 4 and cluster the data of ***paris_grouped*** (transformed to ***data_to_cluster*** by dropping the quarter labels). After fitting model, I had the cluster number (from 0 to 3) for each quarter. By inserting the cluster number to data in ***paris_venues_sorted*** and merging this data frame with ***paris_quarters*** to ***paris_merged***, I finally had a data frame with 20 rows (01 for each quarter) and 14 columns (including quarter name, latitude, longitude, cluster number, names of 05 most common venue categories and number of venues in each categories).

	quarter	latitude	longitude	cluster number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Number of 1st Most Common Venue	Number of 2nd Most Common Venue	Number of 3rd Most Common Venue	Number of 4th Most Common Venue	Number of 5th Most Common Venue
0	Saint-Gervais	48.8557	2.3582	1	French Restaurant	Ice Cream Shop	Art Gallery	Art Museum	Plaza	8	4	4	3	3
1	Chaussée-d'Antin	48.8735	2.3323	2	Hotel	French Restaurant	Plaza	Chocolate Shop	Deli / Bodega	13	8	4	2	2
2	Saint-Ambroise	48.8623	2.3761	0	French Restaurant	Cocktail Bar	Coffee Shop	Wine Bar	Pizza Place	12	5	4	4	4
3	Necker	48.8427	2.3108	3	French Restaurant	Hotel	Japanese Restaurant	Bakery	Korean Restaurant	18	9	4	3	3
4	Sainte-Avoie	48.8626	2.3549	1	Japanese Restaurant	Bookstore	French Restaurant	Italian Restaurant	Cocktail Bar	4	4	4	3	3

c. Examine data to name the clusters

I got the quarters clustered into four (04) different groups, but with cluster labels as numbers the results didn't have much meaning. Therefore, I extracted the data of each cluster to examine and name the clusters based on most common venues. The first cluster named *Restaurant & Bar* comprised of quarters with mostly venues for a variety of food and drink. The second cluster was named *Diverse Entertainment* for the numerous and diverse categories of venues it offered (in short, it had everything). The third cluster, with high hotel density, went with a very self-explanatory name, *Lots of Hotel*. Finally, there were many hotels and French/ Italian restaurants in the fourth cluster, earning it the name of *Hotel & French/Italian Food*.

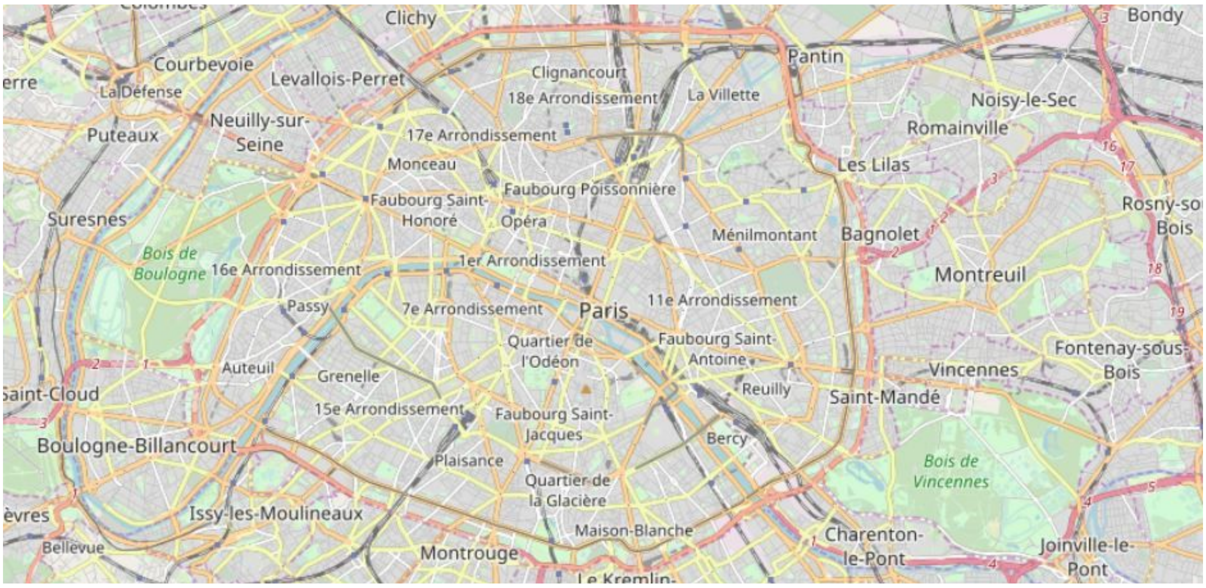
	quarter	latitude	longitude	cluster name	cluster number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Number of 1st Most Common Venue	Number of 2nd Most Common Venue	Number of 3rd Most Common Venue	Number of 4th Most Common Venue
0	Saint-Gervais	48.8557	2.3582	Diverse Entertainment	1	French Restaurant	Ice Cream Shop	Art Gallery	Art Museum	Plaza	8	4	4	3
1	Chaussée-d'Antin	48.8735	2.3323	Lots of Hotel	2	Hotel	French Restaurant	Plaza	Chocolate Shop	Deli / Bodega	13	8	4	2
2	Saint-Ambroise	48.8623	2.3761	Restaurant & Bar	0	French Restaurant	Cocktail Bar	Coffee Shop	Wine Bar	Pizza Place	12	5	4	4
3	Necker	48.8427	2.3108	Hotel & French/Italian Food	3	French Restaurant	Hotel	Japanese Restaurant	Bakery	Korean Restaurant	18	9	4	3
4	Sainte-Avoie	48.8626	2.3549	Diverse Entertainment	1	Japanese Restaurant	Bookstore	French Restaurant	Italian Restaurant	Cocktail Bar	4	4	4	3

I also added a point label to each quarter for result visualization later.

	quarter	latitude	longitude	cluster name	point label	cluster number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Number of 1st Most Common Venue	Number of 2nd Most Common Venue	Number of 3rd Most Common Venue
0	Saint-Gervais	48.8557	2.3582	Diverse Entertainment	8 French Restaurant, 4 Ice Cream Shop, 4 Art G...	1	French Restaurant	Ice Cream Shop	Art Gallery	Art Museum	Plaza	8	4	4
1	Chaussée-d'Antin	48.8735	2.3323	Lots of Hotel	13 Hotel, 8 French Restaurant, 4 Plaza	2	Hotel	French Restaurant	Plaza	Chocolate Shop	Deli / Bodega	13	8	4
2	Saint-Ambroise	48.8623	2.3761	Restaurant & Bar	12 French Restaurant, 5 Cocktail Bar, 4 Coffee...	0	French Restaurant	Cocktail Bar	Coffee Shop	Wine Bar	Pizza Place	12	5	4
3	Necker	48.8427	2.3108	Hotel & French/Italian Food	18 French Restaurant, 9 Hotel, 4 Japanese Rest...	3	French Restaurant	Hotel	Japanese Restaurant	Bakery	Korean Restaurant	18	9	4
4	Sainte-Avoie	48.8626	2.3549	Diverse Entertainment	4 Japanese Restaurant, 4 Bookstore, 4 French R...	1	Japanese Restaurant	Bookstore	French Restaurant	Italian Restaurant	Cocktail Bar	4	4	4

IV. Results

To visualize all results, I first created a map of Paris with *folium* library.

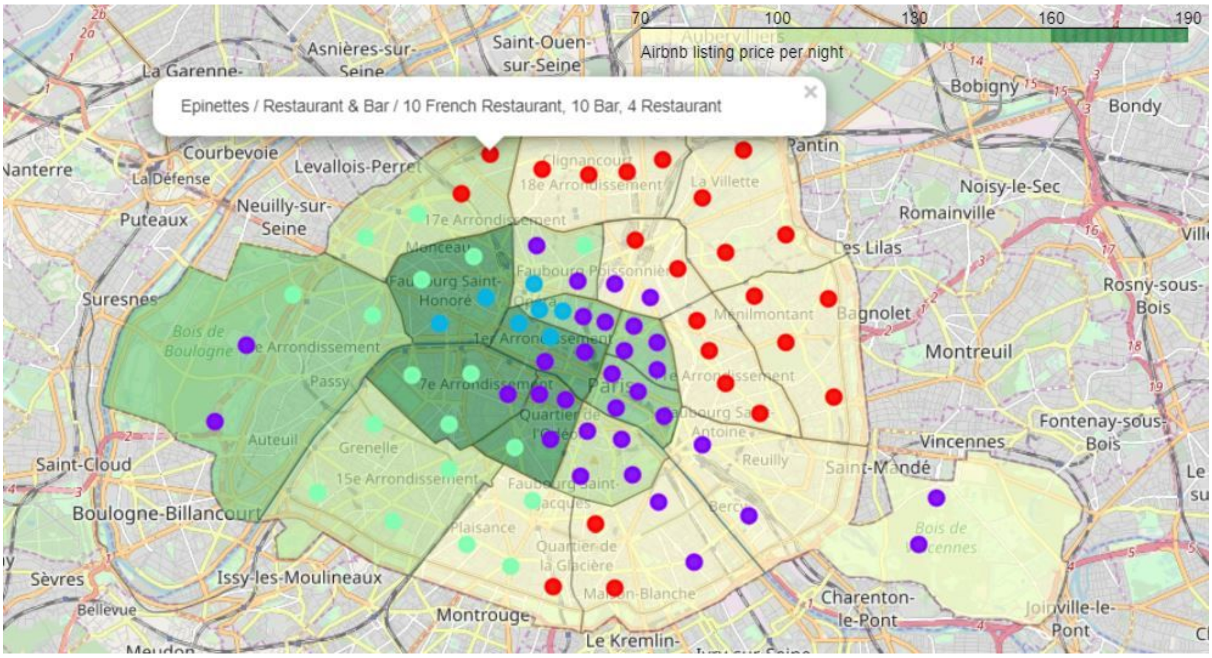


For the average price per night of listings in each district, I used choropleth map - a map that used different shadings to indicate the values of properties in the shaded areas. In this case, I chose yellow to green, with yellow being the districts with cheapest average price per night and green the highest. I didn't want my choropleth map to cut off at random values, so I'd want to set bin values for it. From the analysis conducted in previous section for data frame ***df_listing***, I knew that the minimum average price per night was €73.45 and the maximum was €184.94.

	number of listing	average price per night
count	20.0000	20.0000
mean	2,975.9000	118.9079
std	1,509.1577	35.2606
min	1,163.0000	73.4490
25%	1,877.7500	89.1986
50%	2,594.0000	110.1100
75%	3,664.0000	148.3367
max	6,845.0000	184.9382

Therefore, a bin of €30 with the minimum value of €70 and maximum of €190 would be appropriate: four (04) different shades for four (04) price ranges: yellow (€70 - €100), yellowish green (€100 - €130), green (€130 - €160), and dark green (€160 - €190). After setting the bin values, I drew the choropleth map with geodata from the geoJSON file from Inside Airbnb and ended up with the below map.

For the clustering of quarters, I used dots of different colors to represent different clusters. I picked a rainbow colors scheme, so the first group *Restaurant & Bar* would be red dots, second group *Diverse Entertainment* would be purple, third group *Lots of Hotel* would be blue, and forth group *Hotel & French/Italian Food* would be green. Since I also added a label for each quarter in previous steps, the new map now had different shadings which indicate average price per night for Airbnb listings, and different colored dots showing the entertainment choices around it. For example, pick a red dot in Batignolles-Monceau district and click on it, I can see that it is Epinettes quarter, Airbnb price per night ranges from €100 to €130, the main outlets are restaurants and bars with 10 French restaurants, 4 restaurants (of other cuisine), and 10 bars within a 1,500-meter radius of walking distance.



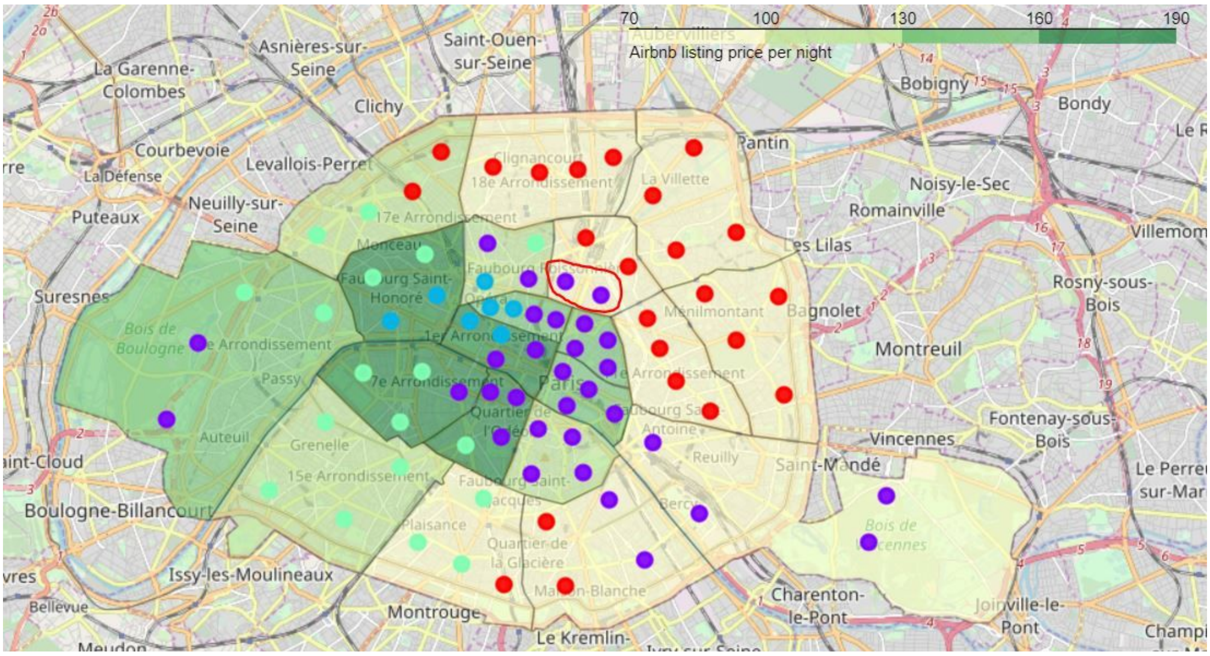
V. Discussion

From the map of final results, I can see the whole picture of Paris.

Generally the inner districts have higher average price per night for each booking (greener color) and more diverse entertainment options (purple dots) - which makes sense since districts are at the city center. Most hotels are located on the west side of the city (represented by blue and green dots), with hotel density also concentrated in the center (blue dots). The outer districts have mostly restaurants and bars.

Through this map, Opéra district stands out as a center district with reasonable price range (under €130 per night) and various forms of entertainment, since there are purple, green, and blue dots in it. It's a good place for people on a tighter budget. On the other hand, Passy may seem a bit unattractive as the district has higher price while being farther from the center. However, there are 15 tennis courts in the vicinity of Auteuil quarter in this district, so maybe tennis players would love to stay there. The art lovers might prefer either Saint-Gervais quarter of Hôtel-de-Ville district or Saint-Thomas-d'Aquin quarter of Palais-Bourbon district, as around these two quarters are multiple art galleries and museums. Regarding those who want to have Asian cuisines, look no further than the south of Gobelins district: both Maison-Blanche and Gare quarters have many Vietnamese and Thai restaurants. Vegans and vegetarians might discover some great places in Saint-Vincent-de-Paul quarter of Entrepôt district.

In my case, I want to stay where it is as cheap as possible while being near to the center. I also want to have lots of options, so I'd choose either quarters Porte-Saint-Denis or Porte-Saint-Martin of Entrepôt district as circled in red.



VI. Conclusion

From the discussion section above, this project might be beneficial for travelers new to Paris, helping to reduce the headache when booking accommodation.

The same analysis could be done for other cities or towns, and further analysis might be done on deeper levels. For example, there could be details on the venues such as name, reviews and pictures, links to preview, etc. In addition, there could be advanced filtering on seasonality of the Airbnb listings (prices during the holidays could be different from off-peak periods), accommodation types (shared/private room or whole apartment), host rating (verified host or not), the list goes on.

Another aspect to look at is the data input. For this project, I used static data collected from predefined sources, which is not the most updated information. Dynamic data fetched directly from Airbnb could provide real-time data and reflect better on the fluctuation of prices. This project is just a starting point, the exploration is endless.

VII. Reference

Airbnb. (2019, September 22). *Airbnb Newsroom Fast Facts*. From Airbnb: https://press.airbnb.com/fast-facts/

Zaleski, O. (2017, April 28). *Bloomberg Technology*. From Bloomberg: https://www.bloomberg.com/news/articles/2017-04-28/airbnb-goes-after-business-travelers-with-new-booking-tool

Get the Data. (2019, July 9). Retrieved from Inside Airbnb: http://insideairbnb.com/get-the-data.html

Quartiers administratifs. (2013, September 30). Retrieved from Paris Data: https://opendata.paris.fr/explore/dataset/quartier_paris/information/

Foursquare developer. (2019, September 22). Retrieved from Foursquare: https://developer.foursquare.com/