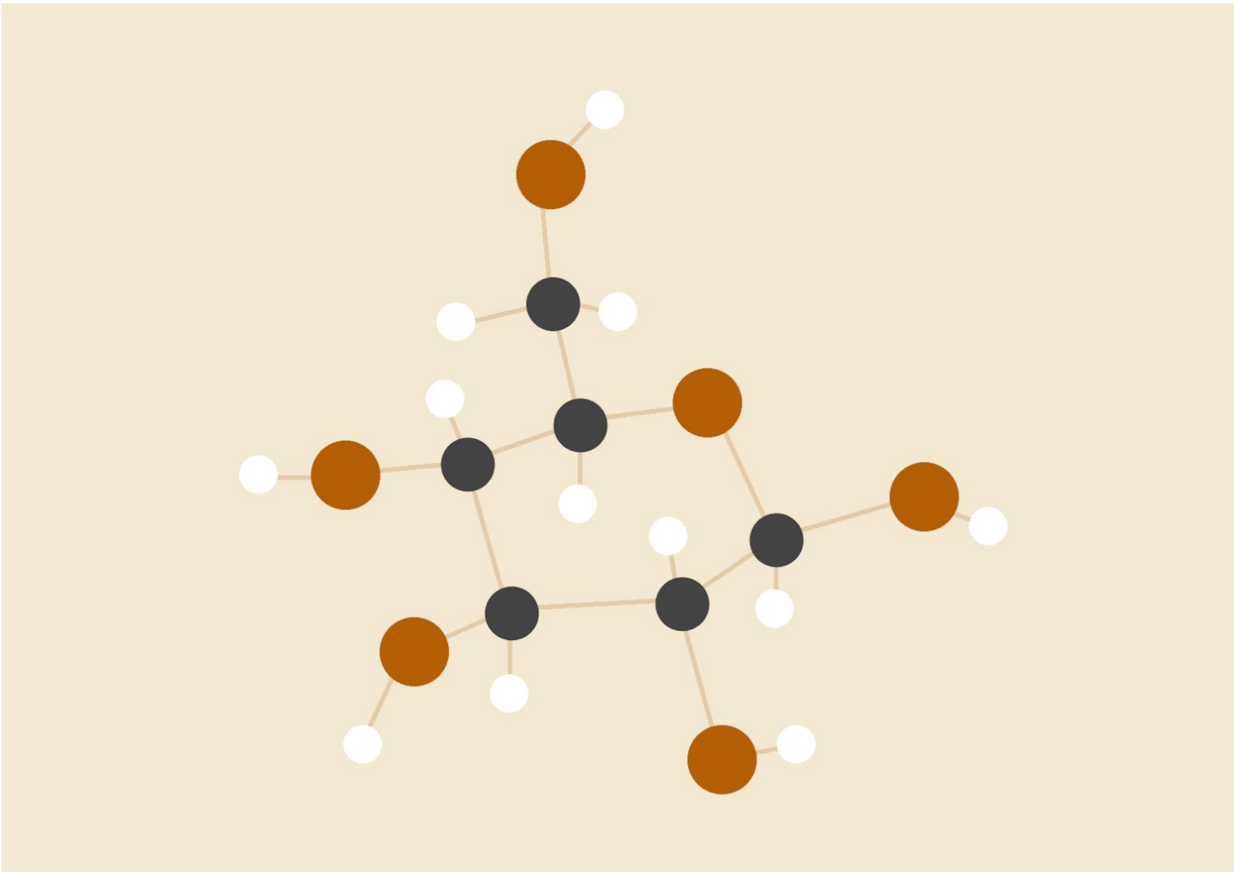
The Battle of Neighborhoods

Applied Data Science Capstone Project



Guillaume RUE

25/06/2020

TABLE OF CONTENT

INTRODUCTION **2**

DATA **2**

DATA CLEANING **4**

Methodology **4**

Exploration of Murray Hill 4

Exploration of Paris neighborhoods 5

Creation of the model 6

RESULTS **8**

CONCLUSION **9**

FUTURE DIRECTIONS **9**

INTRODUCTION

Paris, capital of France, is a worldwide known city. It is composed of twenty boroughs and eighty neighborhoods. For someone outside, it is very complicated to know each neighborhood and what is it specificity. In order to make it easier, this Capstone Project simulates the move of somebody from Murray Hill at Manhattan to Paris. Move inside a country is not always easy, so we can imagine between two differents countries it is even more complicated. If you live somewhere for many years, you will develop habits, depending on where you live. Based on this principle, I decided to compare Murray Hill to each neighborhoods of Paris. To do this, I used the Foursquare API which allows to get every venues of a place.

DATA

Before I can start this project, I needed some data of Paris and Manhattan. First, for Manhattan, I used the dataset of the course which is free and available at this link :

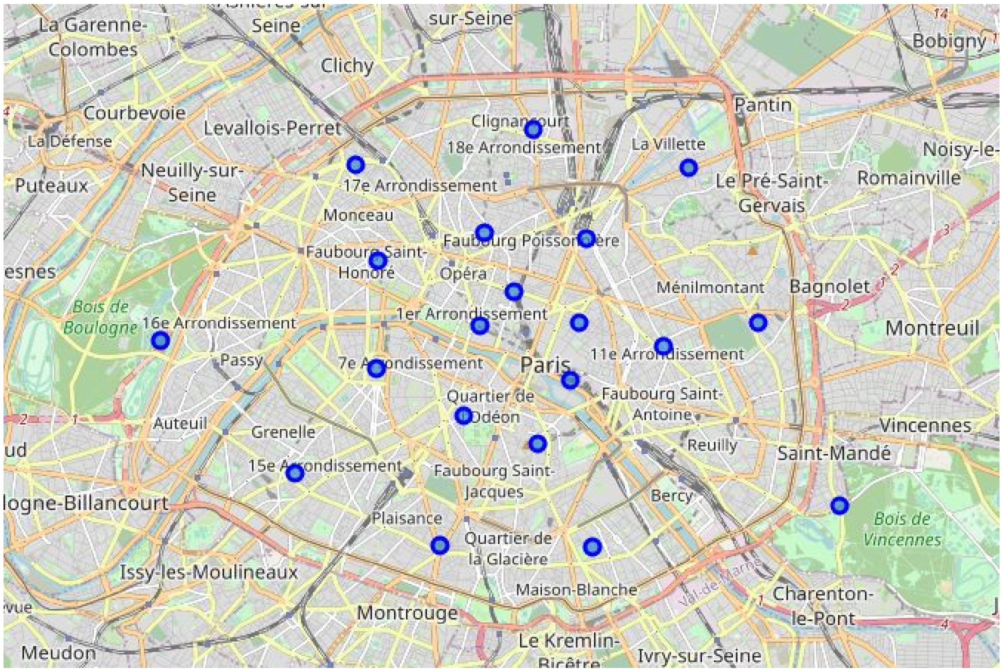
- https://geo.nyu.edu/catalog/nyu_2451_34572

For Paris, I used two free datasets you can download by following next links :

- for boroughs :
  https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e
- for neighborhoods :
  https://opendata.paris.fr/explore/dataset/quartier_paris/download/?format=json&timezone=Europe/Berlin

Those datasets contain names and GPS coordinates of boroughs and neighborhoods so it allowed me to create maps of Paris and get venues with Foursquare API. I think it is easier for someone who does not know a place to be able to locate it instead of just having the name. So, thanks to those datasets I could create two maps :

Paris boroughs



Paris neighborhoods

DATA CLEANING

For this project, the data cleaning part was not very complicated. In fact, from the three datasets, I only needed few information such as :

- name of the borough/neighborhood
- postal code
- GPS coordinates

So I had to determine where those information were and then just create my final dataset. Luckily, there were no missing values. At the end, I had a total of eigthy-one lines in the dataset (eighty from paris neighborhoods + the one for Murray Hill).

Methodology

For this part, I used the lab “Segmenting and Clustering Neighborhoods in New York City”. All the required function were already implemented and it was not necessary for this project to try modify them.

Exploration of Murray Hill

First, to have a better idea of my goal, I decided to have a look at the top 100 venues of Murray Hill and see if some categories were more dominant.

	name	categories	lat	lng
0	Ippodo Tea Co.	Tea Room	40.749757	-73.977733
1	Kajitsu	Japanese Restaurant	40.749763	-73.977688
2	Sons of Thunder	Hawaiian Restaurant	40.747970	-73.975751
3	Perk Kafe	Coffee Shop	40.747768	-73.977363
4	The Renwick Hotel, Curio Collection by Hilton	Hotel	40.750184	-73.977604

As dominant results, I got Hotels, coffee shops, japanese restaurants, sandwich/pizza places, so I was able to think similar neighborhoods should also contain those types of venues.

Exploration of Paris neighborhoods

As I did just before for Murray Hill, I get the top 100 venues for each neighborhood of Paris. Then, I calculated the frequency of each category of venue. For example, if we look at the neighborhood "Champs-Elysées", wa have those results :

```
----Champs-Elysées----
                venue  freq
0   French Restaurant  0.15
1               Hotel  0.10
2            Boutique  0.08
3      Clothing Store  0.05
4 Japanese Restaurant  0.04
```

Then, it was easy to create a dataframe of the 10th most common venues of each neighborhood. This dataframe will be useful in the results section.

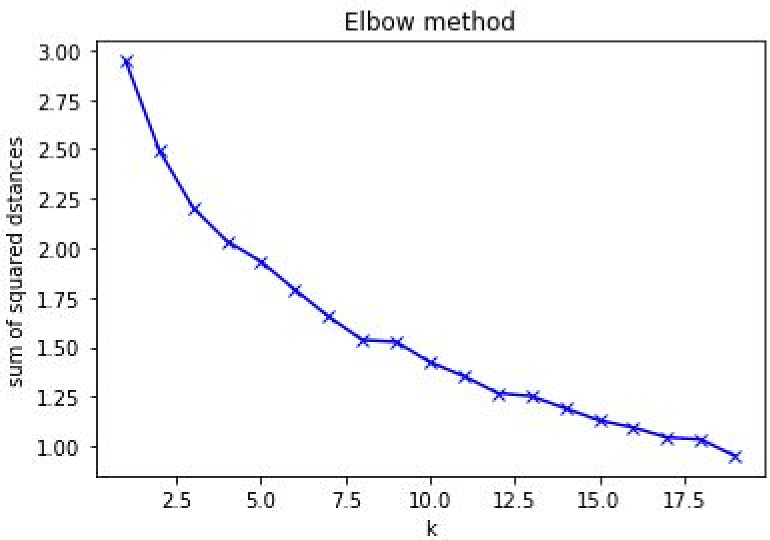
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Amérique	French Restaurant	Supermarket	Bistro	Health Food Store	Café	Bed & Breakfast	Tram Station	Park	Pool	Plaza

Creation of the model

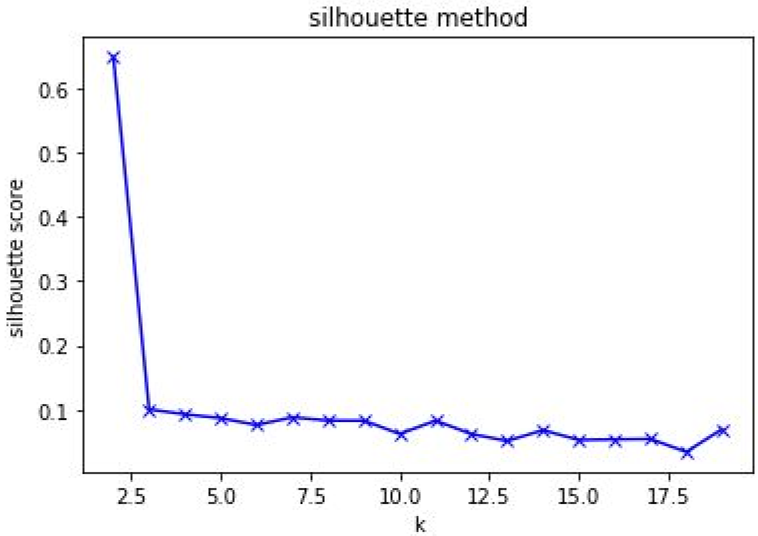
To segment neighborhoods of Paris, based on venues of each, the learning method was unsupervised. In fact, we never knew which neighborhood was similar to another. So, to determine this, I used K-means algorithm.

Find the right value of k :

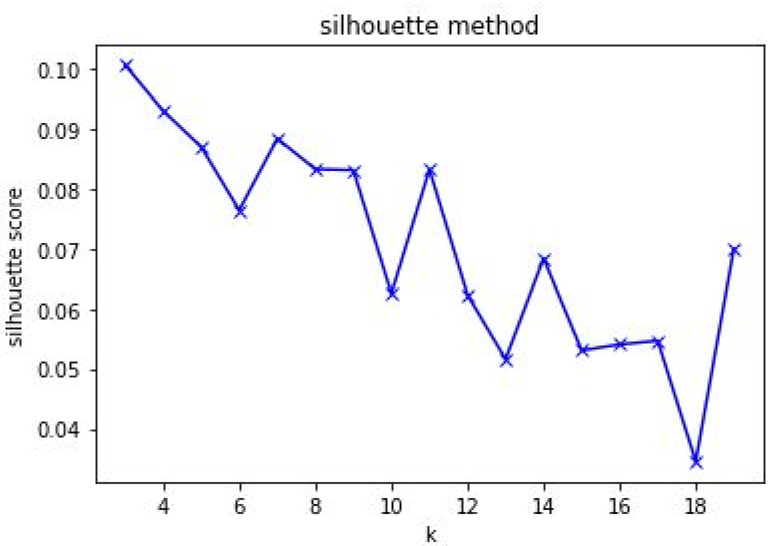
To find k, I started with the elbow method which compute an average score for all clusters. As distance calculation I used the sum of square distances.



Unfortunately, I was not really satisfied with the results. So, I decided to use the silhouette method.



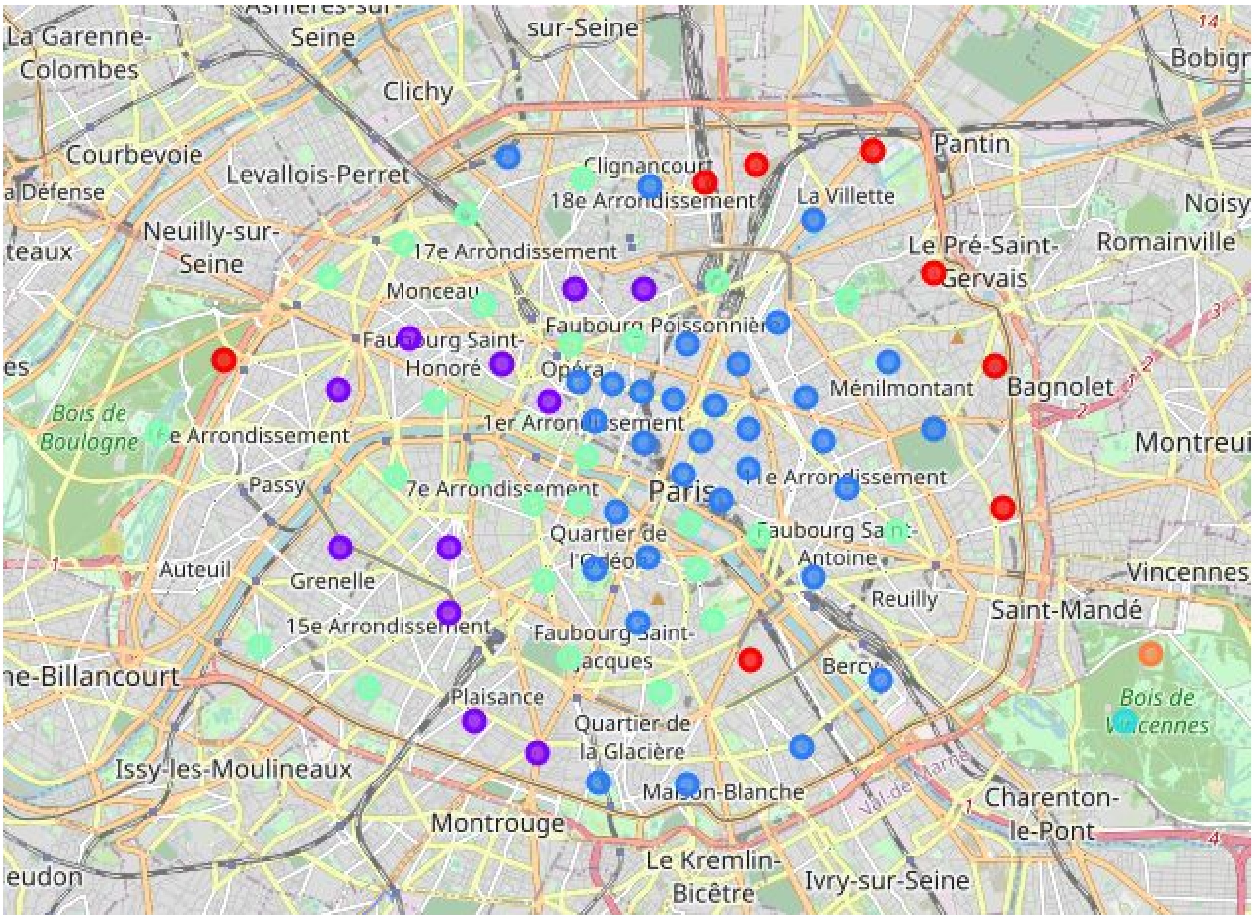
One again, the plot was not really helpful, and I supposed that take k=2 was not necessary the good choice. I wanted to have a wider segmentation of the neighborhoods. So, I just removed k=2 of the plot to have a better visualisation of the plot and see where I found peaks between k=3 and k=19.



So we can see a peak at k=7 and another at k=11, but I decided to choose 7 for the value of k.

RESULTS

After I created my K Means model with k=7, I was able to see which neighborhoods of Paris were in the sme cluster. So, once again I created a map a chose a different color for each cluster :



Finally, I used the model to predict the cluster of Murray Hill, and according to the data I used, it mostly similar to the cluster with blue circles on the map. So I looked at quickly what were top 10 venues of some neighborhoods. As I said before, in the top 5 venues of Murray Hill I found ones likes Japanese restaurants, hotels, coffee shops, pizza/sandwich places. And as you can see in the table below (which is only a part of the entire table), we can see some of the venues I talked just before.

	num_neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	2.0	2	French Restaurant	Coffee Shop	Ice Cream Shop	Bakery	Chinese Restaurant
1	3.0	2	Japanese Restaurant	French Restaurant	Hotel	Coffee Shop	Plaza
2	5.0	2	Japanese Restaurant	Hotel	French Restaurant	Wine Bar	Jewelry Store
3	6.0	2	Japanese Restaurant	French Restaurant	Wine Bar	Hotel	Bistro
4	7.0	2	French Restaurant	Cocktail Bar	Wine Bar	Bakery	Coffee Shop
5	8.0	2	French Restaurant	Cocktail Bar	Hotel	Bakery	Coffee Shop
6	9.0	2	French Restaurant	Hotel	Italian Restaurant	Bar	Chinese Restaurant
7	10.0	2	French Restaurant	Hotel	Japanese Restaurant	Italian Restaurant	Wine Bar
8	11.0	2	French Restaurant	Hotel	Italian Restaurant	Japanese Restaurant	Coffee Shop
9	12.0	2	French Restaurant	Art Gallery	Hotel	Café	Chinese Restaurant

CONCLUSION

So as we just see, if someone would like to move from Murray Hill to Paris and find some similar venues, this person should try to find a place in one of the neighborhood with a blue circle on the last map. It is mostly neighborhoods closed from the city center of Paris with a majority in the 2nd and 3rd borough (8 neighborhoods in total in those two boroughs).

FUTURE DIRECTIONS

The study is only focused on the venues of each neighborhoods. However, to move from a place to another some other points could be importants. For example the house prices is probably a huge point and all neighborhoods of Paris are not equals. Then, maybe it is possible to find a dataset with population repartition such as french, english, german, etc. I suppose it would be easier for an american to live close to others americans who already know the city. Finally, use the proximity of public transport could also be interesting, because Paris is not the easiest city of France to drive so be closed to transport is pretty important for many people.