COURSERA/IBM DATA SCIENCE CAPSTONE PROJECT USING PYTHON

Searching for a neighborhood in Paris to open a virtual vegan restaurant in the context of the Covid-19

Using Foursquare location data, data visualization with Folium and machine learning with the k-means algorithm

Author: Oriane LESIAK

Photo by Ella Olsson @ellaolsson on Unsplash

This article corresponds to the final project of the IBM data science professional certificate. This project offers a good picture of what a data scientist does in real life. The goal is to identify a business problem and to solve it with location data, using specifically the platform of Foursquare, and machine learning. In this case, I will respond to a business client issue who would like to open a virtual vegan restaurant in Paris in the context of the Covid-19 crisis. To do that, I want to make a strategic recommendation on which Parisian neighbourhood is the best choice for starting one. My analysis process will be in 5 axes. I will firstly define the business problem more precisely. Then, I will describe the data sources that I use. After that, I will structure and clean the data in dataframes, and analyze them. Finally, I will make a final recommendation to the client based on the results of my analysis.

Summary:

I. INTRODUCTION: Description of the business problem

I.A. Problem definition: Looking for a neighborhood to open a virtual vegan restaurant in Paris in 2021

As the French political and economic capital, Paris concentrates almost 20% of the French population and 30% of the wealth created in France. As the city of lights, Paris also centralizes culture, arts and gastronomy. In this specific case, I would like to estimate if Paris represents the ideal city to start a foodservice business even in the current context. Thus, I have been contacted by an entrepreneur who would like to open a virtual vegan restaurant in Paris which would propose home delivery and click & collect services. As an influencer, he has developed a social media activity based on vegan, healthy, rapid and affordable cooking. Due to high demand, he is thinking about extending his activity in 2021 by delivering meals to his followers based in Paris. In the context of the Covid-19 crisis, he would like to know if this project is profitable and what are the most interesting neighborhoods in Paris to open his business.

The home delivery and the click & collect solutions are different digital solutions business people have massively developed during the crisis. The business model of the entrepreneur will be as follows: my client will rely on online pre-commands via his own website that will be promoted and relayed by his social media accounts (Instagram and Youtube). He will use a solution like Clickeat that provides a system of online commands for delivery and click & collect. Then, he will prepare the meals and finally will deliver it via outsourced partners, so his clients can eat their food at home or at work. He will follow and adapt the model of "Out Fry" from Tasty in Paris. Also, with this agile model, he will optimize the management of his stock and will prevent waste. So, the idea will be to create a “ghost kitchen”: a place where the meals will be prepared and then delivered to the consumer.

Photo by Louis Hansel @shotsoflouis on Unsplash

Furthermore, his ambition presents different challenges. He would like to find a place in the center of highly dense Paris in order to optimize the access to every district and the time of delivery. However, in the geographical target, the real estate cost and the global foodservice competition are very high. Consequently, he would like to know precisely with the data analysis what would be the best neighborhood to target.

I.B. Target audience & interest for the case

Besides, different people could be also interested in this project:

- Business people or entrepreneurs specialized in the catering and the foodservice industry, that want to see the viability of this type of project, and potentially, to adapt their business model taken into question by the current crisis,
- Investors interested about this industry and new type of business emerging with the Covid-19 context,
- Data scientists and business analysts who want to learn how to use Foursquare and machine learning techniques.

II. DATA SOURCES: Description of the data and how it will be used to solve the problem

For this project, I will collect, analyze and compare specific information about Parisian neighborhoods in order to build my final recommendation. I will mainly combine data of Paris' boroughs and neighborhoods extracted from different official sources and data about venues using the Foursquare API.

II.A. Data about Parisian boroughs and neighborhoods: creation of 2 datasets in Excel based on official data extracted from Paris Data, DataFrance and the Paris Chamber of Notaries

It is important to note that Paris is organized in 20 boroughs that are administrative districts. Each of these boroughs are divided in 4 neighborhoods. So, we will look at the 80 neighborhoods of Paris.

- **Dataset 1 entitled "1. Paris Boroughs & Neighborhoods Data.xlsx":** this dataset gives the list of neighborhoods for each Parisian borough (designated by a postal code and a proper name). It also indicates the geographical coordinates of each neighborhood (latitude and longitude). The data have been extracted from the Paris Data website and rearranged for the project (cleaning and merger of data tables, translation of columns' titles...). More specifically, two main original datasets have been used for the creation of the dataset 1: a dataset giving information about boroughs (source: https://opendata.paris.fr/explore/dataset/arrondissements/export/?disjunctive.c_ar&disjunctive

.c_arinsee&disjunctive.l_ar) and another one about neighborhoods (source: https://opendata.paris.fr/explore/dataset/quartier_paris/export/).

- **Dataset 2 entitled "2. Paris Boroughs Price and Population Data.xlsx":** this other dataset gives specific data about Parisian boroughs: the price per square meter, the percentage of people aged 15-29 years per borough, the percentage of people aged 30-44 years per borough, and the percentage of executives and higher intellectual professions per borough. I have chosen these data because, according to several marketing studies, people which are more interested by vegan and vegetarian food are young people and adults, with high level of education. It is important to notice that data have been limited to the borough level because of the lack of information at a higher scale. Furthermore, a classification of the boroughs has been made for each of this four variables. Indeed, in function of quartiles computed for each variable, I have determined three levels: "Low level", "Mid level" and "High level". I have built this dataset by extracting manually and associating the data taken from DataFrance on population per borough (source: http://map.datafrance.info/population?coords.lat=48.86098807882853&coords.lng=2.3166561126708984&zoom=13, date of the data: 2012), and also from the Paris Chamber of Notaries for the price per square meter per borough (date of the data: end of 2019).

II.B. Data about Parisian venues: use of the Foursquare API

The API gives information about location and different venues in Paris. I will obtain names, categories and locations (longitude and latitude) for each venue.

III. METHODOLOGY OF THE DATA ANALYSIS

III.A. Overall methodology

My overall goal is to identify the best neighborhood where to open the virtual restaurant. In order to identify it, I will compare the different neighborhoods with 3 main types of variables:

- **A cost variable: the price per square meter per borough,** in order to take into account the price of acquisition of a place for the "ghost kitchen".

- **Marketing multiple variables: the percentage of targeted consumers.** As I have previously explained, I will use the 3 following variables for my comparative analysis:
  - The percentage of people aged 15-29 years per borough,
  - The percentage of people aged 30-44 years per borough,
  - The percentage of executives and higher intellectual professions per borough.

- **A competition variable: the portion of similar healthy vegan and vegetarian restaurants that propose home delivery and click & collect services.** With Foursquare, combined to a list of businesses proposing these services, I will determine the number of venues per neighborhood that could compete with the business that my client wants to launch.

I will begin the comparison analysis with the cost and marketing variables, and then, I will use the competition variable.

III.B. First selection of neighborhoods with the cost and marketing multiple variables

III.B.1. Data import and cleaning

Firstly, I will import Pandas and Numpy libraries, and also, datasets 1 and 2. I will then merge these two datasets in order to create my initial dataframe. You can see my resulting dataframe below.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions
0	75001	Louvre	Halles	48.862289	2.344899	12840	High level	23	Low level	26	High level	36	High level
1	75001	Louvre	Palais-Royal	48.864660	2.336309	12840	High level	23	Low level	26	High level	36	High level
2	75001	Louvre	Saint-Germain-l'Auxerrois	48.860650	2.334910	12840	High level	23	Low level	26	High level	36	High level
3	75001	Louvre	Place-Vendôme	48.867019	2.328582	12840	High level	23	Low level	26	High level	36	High level
4	75002	Bourse	Mail	48.868008	2.344699	11250	Mid level	27	High level	30	High level	37	High level
...	...	...	...	...	...	...	...	...	...	...	...	...	...
75	75019	Buttes-Chaumont	Combat	48.878639	2.380127	8490	Low level	22	Low level	23	Mid level	19	Low level
76	75020	Ménilmontant	Père-Lachaise	48.863719	2.395273	8560	Low level	21	Low level	24	Mid level	21	Low level
77	75020	Ménilmontant	Belleville	48.871531	2.387549	8560	Low level	21	Low level	24	Mid level	21	Low level
78	75020	Ménilmontant	Saint-Fargeau	48.871035	2.406172	8560	Low level	21	Low level	24	Mid level	21	Low level
79	75020	Ménilmontant	Charonne	48.854760	2.407430	8560	Low level	21	Low level	24	Mid level	21	Low level

80 rows × 13 columns

Then, in order to narrow my data, I will suppress the cells that don't match my targets.

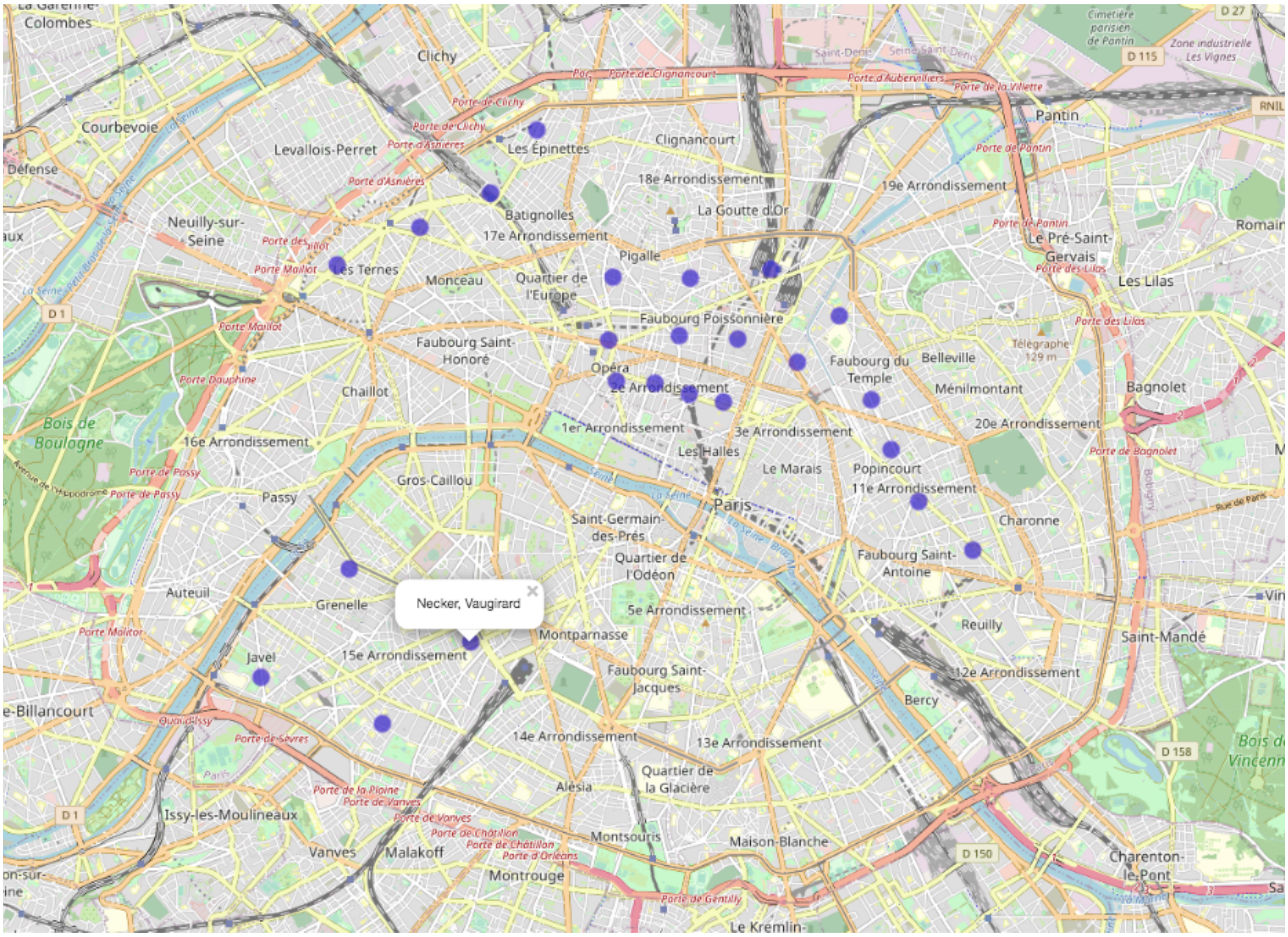
- For the level of % of people aged 15-29 years, I suppress rows with low level, as I am targeting boroughs/neighborhoods with high or mid-levels.
- For the level of % of people aged 30-44 years, I suppress rows with low level, as I am targeting boroughs/neighborhoods with high or mid levels.
- For the level of % of executives and higher intellectual professions, I suppress rows with low level, as I am targeting boroughs/neighborhoods with high or mid-levels.
- For the level of price per square meter, I suppress all rows that present a high level, as I am looking for affordable locations.

As a result, I obtain a final dataframe with 24 potential neighborhoods to target, which are concentrated in 6 main boroughs: Bourse, Opéra, Entrepôt, Popincourt, Vaugirard and Batignolles-Monceau.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions
4	75002	Bourse	Mail	48.868008	2.344699	11250	Mid level	27	High level	30	High level	37	High level
5	75002	Bourse	Bonne-Nouvelle	48.867150	2.350080	11250	Mid level	27	High level	30	High level	37	High level
6	75002	Bourse	Gaillon	48.869307	2.333432	11250	Mid level	27	High level	30	High level	37	High level
7	75002	Bourse	Vivienne	48.869100	2.339461	11250	Mid level	27	High level	30	High level	37	High level
32	75009	Opéra	Rochechouart	48.879812	2.344861	10730	Mid level	24	Mid level	26	High level	36	High level
33	75009	Opéra	Saint-Georges	48.879934	2.332850	10730	Mid level	24	Mid level	26	High level	36	High level
34	75009	Opéra	Chaussée-d'Antin	48.873547	2.332269	10730	Mid level	24	Mid level	26	High level	36	High level
35	75009	Opéra	Faubourg-Montmartre	48.873935	2.343253	10730	Mid level	24	Mid level	26	High level	36	High level
36	75010	Entrepôt	Hôpital-Saint-Louis	48.876008	2.368123	9730	Low level	24	Mid level	28	High level	31	Mid level
37	75010	Entrepôt	Porte-Saint-Denis	48.873618	2.352283	9730	Low level	24	Mid level	28	High level	31	Mid level
38	75010	Entrepôt	Saint-Vincent-de-Paul	48.880735	2.357471	9730	Low level	24	Mid level	28	High level	31	Mid level
39	75010	Entrepôt	Porte-Saint-Martin	48.871245	2.361504	9730	Low level	24	Mid level	28	High level	31	Mid level
40	75011	Popincourt	Sainte-Marguerite	48.852097	2.388765	9980	Mid level	25	High level	27	High level	31	Mid level
41	75011	Popincourt	Saint-Ambroise	48.862345	2.376118	9980	Mid level	25	High level	27	High level	31	Mid level
42	75011	Popincourt	Folie-Méricourt	48.867403	2.372965	9980	Mid level	25	High level	27	High level	31	Mid level
43	75011	Popincourt	Roquette	48.857064	2.380364	9980	Mid level	25	High level	27	High level	31	Mid level
56	75015	Vaugirard	Grenelle	48.850172	2.291853	10030	Mid level	24	Mid level	23	Mid level	32	Mid level
57	75015	Vaugirard	Necker	48.842711	2.310777	10030	Mid level	24	Mid level	23	Mid level	32	Mid level
58	75015	Vaugirard	Saint-Lambert	48.834294	2.296920	10030	Mid level	24	Mid level	23	Mid level	32	Mid level
59	75015	Vaugirard	Javel	48.839060	2.278076	10030	Mid level	24	Mid level	23	Mid level	32	Mid level
64	75017	Batignolles-Monceau	Batignolles	48.888482	2.313856	10210	Mid level	24	Mid level	24	Mid level	31	Mid level
65	75017	Batignolles-Monceau	Epinettes	48.894943	2.321119	10210	Mid level	24	Mid level	24	Mid level	31	Mid level
66	75017	Batignolles-Monceau	Ternes	48.881178	2.289964	10210	Mid level	24	Mid level	24	Mid level	31	Mid level
67	75017	Batignolles-Monceau	Plaine de Monceaux	48.885044	2.302910	10210	Mid level	24	Mid level	24	Mid level	31	Mid level

III.B.2. Visualization of the first selection of neighborhoods in Paris, using Folium

Then, I want to visualize my first results on a map of Paris with the Folium library. I add the geographical coordinates of Paris by referring to the GeoPy library, and combine them with the coordinates of the Parisian neighborhoods that I have selected. The map represents the neighborhoods in purple points. For each point, a data label indicates the names of the neighborhood and of the borough.



III.C. Second selection of neighborhoods with the competition variable

III.C.1. Import and structure of the data from Foursquare

Now, I will narrow the analysis adding the competition variable. I start using the Foursquare API to retrieve nearby venues information in the selected neighborhoods (that is to say, names, locations and category types). My research is limited to 100 venues for each neighborhood and within a radius of 500 meters. The API returns a JSON file with all venues data that I transform into a dataframe. In this dataframe, I add the information about the corresponding neighborhood for each venue: name and geographical coordinates (see below).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Mail	48.868008	2.344699	Hoppy Corner	48.867726	2.347375	Beer Bar
1	Mail	48.868008	2.344699	L'Appartement Sézane	48.869574	2.345060	Women's Store
2	Mail	48.868008	2.344699	Lockwood	48.867727	2.346945	Cocktail Bar
3	Mail	48.868008	2.344699	Le Moderne	48.868856	2.342142	French Restaurant
4	Mail	48.868008	2.344699	Boneshaker Doughnuts	48.867857	2.347341	Donut Shop
...	...	...	...	...	...	...	...
1725	Plaine de Monceaux	48.885044	2.302910	Restaurant Lyna	48.887904	2.306638	Italian Restaurant
1726	Plaine de Monceaux	48.885044	2.302910	Franprix	48.888840	2.303700	Supermarket
1727	Plaine de Monceaux	48.885044	2.302910	Coccodrillo	48.887823	2.306940	Italian Restaurant
1728	Plaine de Monceaux	48.885044	2.302910	Hotel Mercure Paris 17 Batignolles	48.888060	2.306772	Hotel
1729	Plaine de Monceaux	48.885044	2.302910	Pavillon Pereire	48.887687	2.307940	Hotel

1730 rows × 7 columns

In total, I obtain 1730 venues for all targeted neighborhoods.

III.C.2. Narrowing of the results to vegetarian / vegan restaurants and addition of the criteria of Covid-19 adapted services

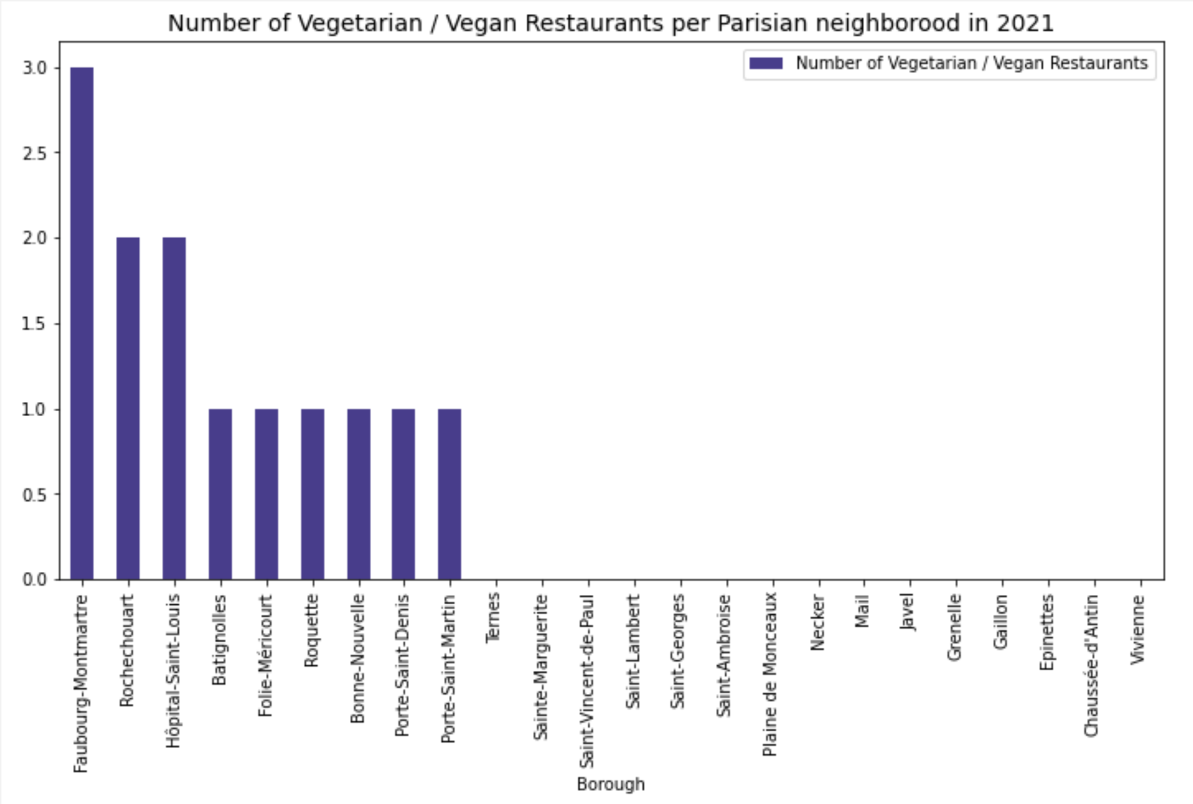
Then, I narrow venues information to a specific venue category: vegetarian and vegan restaurants. I also don’t forget to drop potential duplicates. Indeed, some venues appear twice in different bordering neighborhoods. In total, I obtain 14 results of vegetarian / vegan restaurants in Paris.

In the context of the Covid-19, I also add two dining options criteria to measure if the venues have adapted their services. Both options are the possibility of home delivery (by the restaurant or specialized firms such as Uber Eats, Deliveroo and Just Eat) and takeaway/click & collect. Referring to local files of Google My Business, I have consolidated an Excel file with the information for each restaurant: it's called "3. Paris List of Vegan & Vegetarian restaurants with Covid-19 adapted services". Thus, I can see that all restaurants propose at least a take-away and/or click & collect dining option. Only 4 venues don't also guarantee a delivery service. Consequently, as every restaurant have adapted its services, I won't consider dining options as discriminatory criteria.

III.C.3. Preparation for data processing with machine learning

I will now obtain and/or compute 5 sub-variables for each neighborhood:

- The list of names of vegetarian / vegan restaurants per neighborhood for possible benchmark,
- The absolute number of vegetarian / vegan restaurants per neighborhood that we can visualize below:



- The number of restaurants with a delivery service per neighborhood,
- The number of restaurants with take-away and/or click & collect services per neighborhood.
- The more precise percentage of vegetarian / vegan restaurants per neighborhood that I will use after for data processing in machine learning. I compute this percentage with the method of "one-hot encoding". The idea is to convert the data obtained for all category types into numerical and binary vectors that we called "dummies". For that, I compute the frequency of occurrence of all category types for each venue, and then, I group the results by computing the mean of the frequencies for each neighborhood. Then, I only keep the data obtained for the category of "vegetarian / vegan restaurant". As a consequence, I can use these final results as inputs for the application of the k-means algorithm.

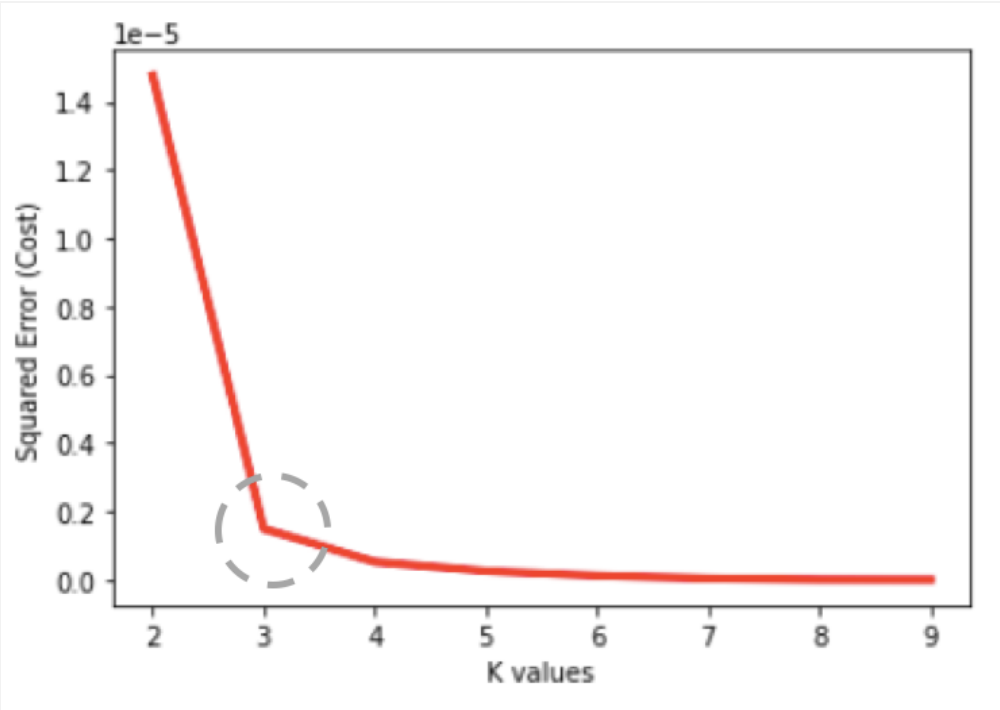
Then, I merge the 5 sub-variables that I have obtained to my initial dataframe in order to have an exhaustive picture of the situation. You can see the final dataframe below.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
7	75009	Opéra	Faubourg-Montmartre	48.873935	2.343253	10730	Mid level	24	Mid level	26	High level	36	High level	0.050847	3.0	2.0	3.0	So Nat,Le Tricycle,La Palanche D'âulac
4	75009	Opéra	Rochechouart	48.879812	2.344861	10730	Mid level	24	Mid level	26	High level	36	High level	0.043478	2.0	2.0	2.0	Le Potager de Charlotte,42 Degrés
8	75010	Entrepôt	Hôpital-Saint-Louis	48.876008	2.368123	9730	Low level	24	Mid level	28	High level	31	Mid level	0.020202	2.0	2.0	2.0	Bodhi Vegan,Sol Semilla
20	75017	Batignolles-Monceau	Batignolles	48.888482	2.313856	10210	Mid level	24	Mid level	24	Mid level	31	Mid level	0.010204	1.0	0.0	1.0	My Kitch'n
9	75010	Entrepôt	Porte-Saint-Denis	48.873618	2.352283	9730	Low level	24	Mid level	28	High level	31	Mid level	0.019231	1.0	0.0	1.0	Jah Jah
11	75010	Entrepôt	Porte-Saint-Martin	48.871245	2.361504	9730	Low level	24	Mid level	28	High level	31	Mid level	0.023256	1.0	1.0	1.0	Elaichi
1	75002	Bourse	Bonne-Nouvelle	48.867150	2.350080	11250	Mid level	27	High level	30	High level	37	High level	0.017544	1.0	0.0	1.0	Kitchen
14	75011	Popincourt	Folie-Méricourt	48.867403	2.372965	9980	Mid level	25	High level	27	High level	31	Mid level	0.013514	1.0	1.0	1.0	Soya Cantine Bio
15	75011	Popincourt	Roquette	48.857064	2.380364	9980	Mid level	25	High level	27	High level	31	Mid level	0.013889	1.0	1.0	1.0	Aujourd'hui & Demain
0	75002	Bourse	Mail	48.868008	2.344699	11250	Mid level	27	High level	30	High level	37	High level	0.000000	0.0	0.0	0.0	0
16	75015	Vaugirard	Grenelle	48.850172	2.291853	10030	Mid level	24	Mid level	23	Mid level	32	Mid level	0.000000	0.0	0.0	0.0	0
22	75017	Batignolles-Monceau	Ternes	48.881178	2.289964	10210	Mid level	24	Mid level	24	Mid level	31	Mid level	0.000000	0.0	0.0	0.0	0
21	75017	Batignolles-Monceau	Epinettes	48.894943	2.321119	10210	Mid level	24	Mid level	24	Mid level	31	Mid level	0.000000	0.0	0.0	0.0	0
19	75015	Vaugirard	Javel	48.839060	2.278076	10030	Mid level	24	Mid level	23	Mid level	32	Mid level	0.000000	0.0	0.0	0.0	0
18	75015	Vaugirard	Saint-Lambert	48.834294	2.296920	10030	Mid level	24	Mid level	23	Mid level	32	Mid level	0.000000	0.0	0.0	0.0	0
17	75015	Vaugirard	Necker	48.842711	2.310777	10030	Mid level	24	Mid level	23	Mid level	32	Mid level	0.000000	0.0	0.0	0.0	0
12	75011	Popincourt	Sainte-Marguerite	48.852097	2.388765	9980	Mid level	25	High level	27	High level	31	Mid level	0.000000	0.0	0.0	0.0	0
13	75011	Popincourt	Saint-Ambroise	48.862345	2.376118	9980	Mid level	25	High level	27	High level	31	Mid level	0.000000	0.0	0.0	0.0	0
10	75010	Entrepôt	Saint-Vincent-de-Paul	48.880735	2.357471	9730	Low level	24	Mid level	28	High level	31	Mid level	0.000000	0.0	0.0	0.0	0
6	75009	Opéra	Chaussée-d'Antin	48.873547	2.332269	10730	Mid level	24	Mid level	26	High level	36	High level	0.000000	0.0	0.0	0.0	0
5	75009	Opéra	Saint-Georges	48.879934	2.332850	10730	Mid level	24	Mid level	26	High level	36	High level	0.000000	0.0	0.0	0.0	0
3	75002	Bourse	Vivienne	48.869100	2.339461	11250	Mid level	27	High level	30	High level	37	High level	0.000000	0.0	0.0	0.0	0
2	75002	Bourse	Gaillon	48.869307	2.333432	11250	Mid level	27	High level	30	High level	37	High level	0.000000	0.0	0.0	0.0	0
23	75017	Batignolles-Monceau	Plaine de Monceaux	48.885044	2.302910	10210	Mid level	24	Mid level	24	Mid level	31	Mid level	0.000000	0.0	0.0	0.0	0

III.C.4. First clustering of the neighborhoods with the k-means algorithm according to the percentage of vegetarian / vegan restaurants

As an unsupervised learning, the k-means algorithm is used for clustering. I will use it to segment the neighborhoods of my dataframe into several groups, called “clusters”, in function of the competition variable. Indeed, the neighborhoods will be grouped in function of their similarities, in terms of percentage of vegetarian and vegan restaurants.

First of all, I need to determine the best number of clusters with the “Elbow method”. With this method, I determine a range of potential values for “k”: between 2 and 10. For each of this value, I compute the total within-cluster sum of squared errors. Then, I plot it as a curve on a graph in function of the number of clusters. As a result, I can determine the best “k” value where the curve represents an elbow. Here, I can see that I have obtained the number 3 as the best "k" value.



Consequently, I can cluster the Parisian neighborhoods into 3 groups using the k-means algorithm. I will create a new dataframe, by adding a column with cluster labels to my previous dataframe, and I will visualize the results on a new map.

Details of the 3 clusters of neighborhoods according to the competition variable (percentage of vegetarian / vegan restaurants):

- Cluster 0: it contains all the neighborhoods which have none vegan / vegetarian restaurants. It is shown in red color in the map.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
0	75002	Bourse	Mail	48.868008	2.344699	11250	Mid level	27	High level	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0
16	75015	Vaugirard	Grenelle	48.850172	2.291853	10030	Mid level	24	Mid level	23	Mid level	32	Mid level	0.0	0	0.0	0.0	0.0	0
22	75017	Batignolles-Monceau	Ternes	48.881178	2.289964	10210	Mid level	24	Mid level	24	Mid level	31	Mid level	0.0	0	0.0	0.0	0.0	0
21	75017	Batignolles-Monceau	Epinettes	48.894943	2.321119	10210	Mid level	24	Mid level	24	Mid level	31	Mid level	0.0	0	0.0	0.0	0.0	0
19	75015	Vaugirard	Javel	48.839060	2.278076	10030	Mid level	24	Mid level	23	Mid level	32	Mid level	0.0	0	0.0	0.0	0.0	0
18	75015	Vaugirard	Saint-Lambert	48.834294	2.296920	10030	Mid level	24	Mid level	23	Mid level	32	Mid level	0.0	0	0.0	0.0	0.0	0
17	75015	Vaugirard	Necker	48.842711	2.310777	10030	Mid level	24	Mid level	23	Mid level	32	Mid level	0.0	0	0.0	0.0	0.0	0
12	75011	Popincourt	Sainte-Marguerite	48.852097	2.388765	9980	Mid level	25	High level	27	High level	31	Mid level	0.0	0	0.0	0.0	0.0	0
13	75011	Popincourt	Saint-Ambroise	48.862345	2.376118	9980	Mid level	25	High level	27	High level	31	Mid level	0.0	0	0.0	0.0	0.0	0
10	75010	Entrepôt	Saint-Vincent-de-Paul	48.880735	2.357471	9730	Low level	24	Mid level	28	High level	31	Mid level	0.0	0	0.0	0.0	0.0	0
6	75009	Opéra	Chaussée-d'Antin	48.873547	2.332269	10730	Mid level	24	Mid level	26	High level	36	High level	0.0	0	0.0	0.0	0.0	0
5	75009	Opéra	Saint-Georges	48.879934	2.332850	10730	Mid level	24	Mid level	26	High level	36	High level	0.0	0	0.0	0.0	0.0	0
3	75002	Bourse	Vivienne	48.869100	2.339461	11250	Mid level	27	High level	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0
2	75002	Bourse	Gaillon	48.869307	2.333432	11250	Mid level	27	High level	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0
23	75017	Batignolles-Monceau	Plaine de Monceaux	48.885044	2.302910	10210	Mid level	24	Mid level	24	Mid level	31	Mid level	0.0	0	0.0	0.0	0.0	0

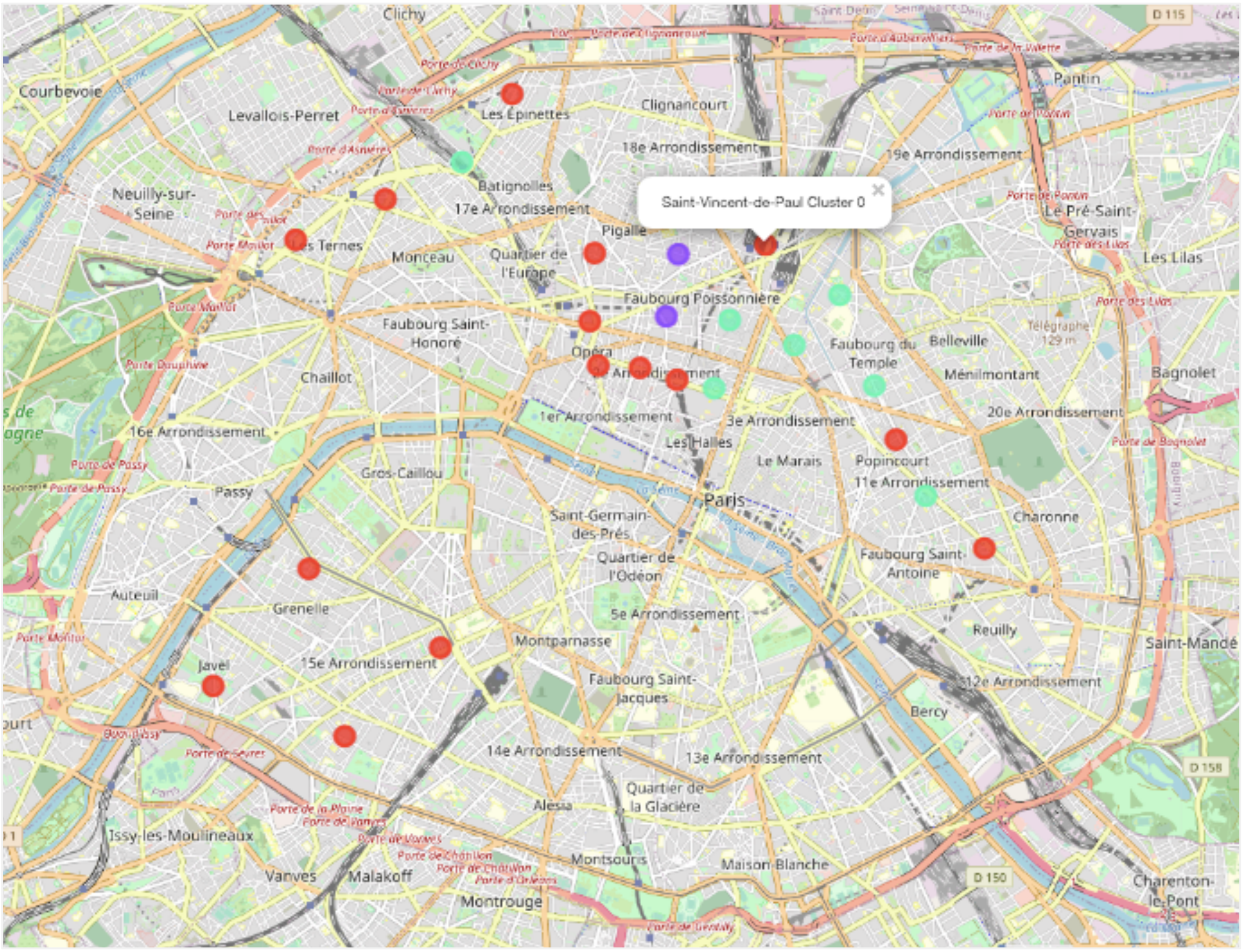
- Cluster 1: it contains all the neighborhoods which have the highest percentages of vegan / vegetarian restaurants. It is shown in purple color in the map.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
7	75009	Opéra	Faubourg-Montmartre	48.873935	2.343253	10730	Mid level	24	Mid level	26	High level	36	High level	0.050847	1	3.0	2.0	3.0	So Nat,Le Tricycle,La Palanche D'âulac
4	75009	Opéra	Rochechouart	48.879812	2.344861	10730	Mid level	24	Mid level	26	High level	36	High level	0.043478	1	2.0	2.0	2.0	Le Potager de Charlotte,42 Degrés

- Cluster 2: it contains all the neighborhoods which have medium percentages of vegan / vegetarian restaurants. It is shown in turquoise blue color in the map.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
20	75017	Batignolles-Monceau	Batignolles	48.888482	2.313856	10210	Mid level	24	Mid level	24	Mid level	31	Mid level	0.010204	2	1.0	0.0	1.0	My Kitch'n
9	75010	Entrepôt	Porte-Saint-Denis	48.873618	2.352283	9730	Low level	24	Mid level	28	High level	31	Mid level	0.019231	2	1.0	0.0	1.0	Jah Jah
11	75010	Entrepôt	Porte-Saint-Martin	48.871245	2.361504	9730	Low level	24	Mid level	28	High level	31	Mid level	0.023256	2	1.0	1.0	1.0	Elaichi
1	75002	Bourse	Bonne-Nouvelle	48.867150	2.350080	11250	Mid level	27	High level	30	High level	37	High level	0.017544	2	1.0	0.0	1.0	Kitchen
14	75011	Popincourt	Folie-Méricourt	48.867403	2.372965	9980	Mid level	25	High level	27	High level	31	Mid level	0.013514	2	1.0	1.0	1.0	Soya Cantine Bio
15	75011	Popincourt	Roquette	48.857064	2.380364	9980	Mid level	25	High level	27	High level	31	Mid level	0.013889	2	1.0	1.0	1.0	Aujourd'hui & Demain

Visualization of the results:



Finally, I choose to keep only the Cluster 0 for the pursuit of the analysis because it encompasses all neighborhoods with none vegan / vegetarian restaurants. I make the choice to target these neighborhoods because of the inexistent competition, but without taking into account the delivery perimeter of restaurants of other neighborhoods. I will now reuse the clustering algorithm to narrow the results. Indeed, I will reapply it to obtain new clusters in function of the population variables, and finally, in function of the property price variable.

III.C.5. Second clustering with the Cluster 0 in function of targeted populations data

I will reuse the k-means algorithm to create sub-clusters in the Cluster 0 in function of targeted populations data. These data refer to the percentages per borough of:

- People aged 15-29 years and 30-44 years,
- Executives and higher intellectual professions.

With the “Elbow method”, I find a new best “k” value which is 4.

Details of the 4 clusters of neighborhoods according to the population variable:

- Cluster 00: it contains all the neighborhoods with high levels for all targeted populations.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	Target Population Cluster Label	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
0	75002	Bourse	Mail	48.868008	2.344699	11250	Mid level	0	27	High level	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0
3	75002	Bourse	Vivienne	48.869100	2.339461	11250	Mid level	0	27	High level	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0
2	75002	Bourse	Gaillon	48.869307	2.333432	11250	Mid level	0	27	High level	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0

- Cluster 01: it contains all the neighborhoods with mid levels for all targeted populations.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	Target Population Cluster Label	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
16	75015	Vaugirard	Grenelle	48.850172	2.291853	10030	Mid level	1	24	Mid level	23	Mid level	32	Mid level	0.0	0	0.0	0.0	0.0	0
22	75017	Batignolles-Monceau	Ternes	48.881178	2.289964	10210	Mid level	1	24	Mid level	24	Mid level	31	Mid level	0.0	0	0.0	0.0	0.0	0
21	75017	Batignolles-Monceau	Epinettes	48.894943	2.321119	10210	Mid level	1	24	Mid level	24	Mid level	31	Mid level	0.0	0	0.0	0.0	0.0	0
19	75015	Vaugirard	Javel	48.839060	2.278076	10030	Mid level	1	24	Mid level	23	Mid level	32	Mid level	0.0	0	0.0	0.0	0.0	0
18	75015	Vaugirard	Saint-Lambert	48.834294	2.296920	10030	Mid level	1	24	Mid level	23	Mid level	32	Mid level	0.0	0	0.0	0.0	0.0	0
17	75015	Vaugirard	Necker	48.842711	2.310777	10030	Mid level	1	24	Mid level	23	Mid level	32	Mid level	0.0	0	0.0	0.0	0.0	0
23	75017	Batignolles-Monceau	Plaine de Monceaux	48.885044	2.302910	10210	Mid level	1	24	Mid level	24	Mid level	31	Mid level	0.0	0	0.0	0.0	0.0	0

- Cluster 02: it contains all the neighborhoods with at least 2 high levels for people aged 30-44 years and executives / higher intellectual professions.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	Target Population Cluster Label	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
6	75009	Opéra	Chaussée-d'Antin	48.873547	2.332269	10730	Mid level	2	24	Mid level	26	High level	36	High level	0.0	0	0.0	0.0	0.0	0
5	75009	Opéra	Saint-Georges	48.879934	2.332850	10730	Mid level	2	24	Mid level	26	High level	36	High level	0.0	0	0.0	0.0	0.0	0

- Cluster 03: it contains all the neighborhoods with at least 1 high level for people aged 30-44 years.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	Target Population Cluster Label	% of people aged 15-29 years per borough	Level of % of people aged 15-29 yrs	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
12	75011	Popincourt	Sainte-Marguerite	48.852097	2.388765	9980	Mid level	3	25	High level	27	High level	31	Mid level	0.0	0	0.0	0.0	0.0	0
13	75011	Popincourt	Saint-Ambroise	48.862345	2.376118	9980	Mid level	3	25	High level	27	High level	31	Mid level	0.0	0	0.0	0.0	0.0	0
10	75010	Entrepôt	Saint-Vincent-de-Paul	48.880735	2.357471	9730	Low level	3	24	Mid level	28	High level	31	Mid level	0.0	0	0.0	0.0	0.0	0

Considering these results, I can limit my analysis to the Cluster 00 which includes neighborhoods with the highest levels of targeted populations. However, I can see that the property price is the also highest of the 4 clusters. As a consequence, I will do a last clustering with all neighborhoods that present at least 2 high levels for targeted populations. For that, I will make a concatenation of the neighborhoods of clusters 00, 02 and 03. For the cluster 03, I will specifically suppress the data of the neighborhood “Saint-Vincent-de-Paul”.

III.C.6. Final clustering with a concatenation of clusters in function of the property price data

For the last clustering, I will reapply the k-means algorithm to segment the obtained neighborhoods in function of the borough price per square meter. In this case, I choose to define myself a value for “k”, that is to say 3, in order to have the most precise results.

Details of the 3 clusters of neighborhoods according to the real estate price variable:

- Cluster 0C0: it contains the neighborhoods of the Opéra borough which presents the medium level of price per square meter in comparison to the 2 other clusters. It is shown in red on the map.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	Property Price Cluster Label	Target Population Cluster Label	% of people aged 15-29 years per borough	...	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
6	75009	Opéra	Chaussée-d'Antin	48.873547	2.332269	10730	Mid level	0	2	24	...	26	High level	36	High level	0.0	0	0.0	0.0	0.0	0
5	75009	Opéra	Saint-Georges	48.879934	2.332850	10730	Mid level	0	2	24	...	26	High level	36	High level	0.0	0	0.0	0.0	0.0	0

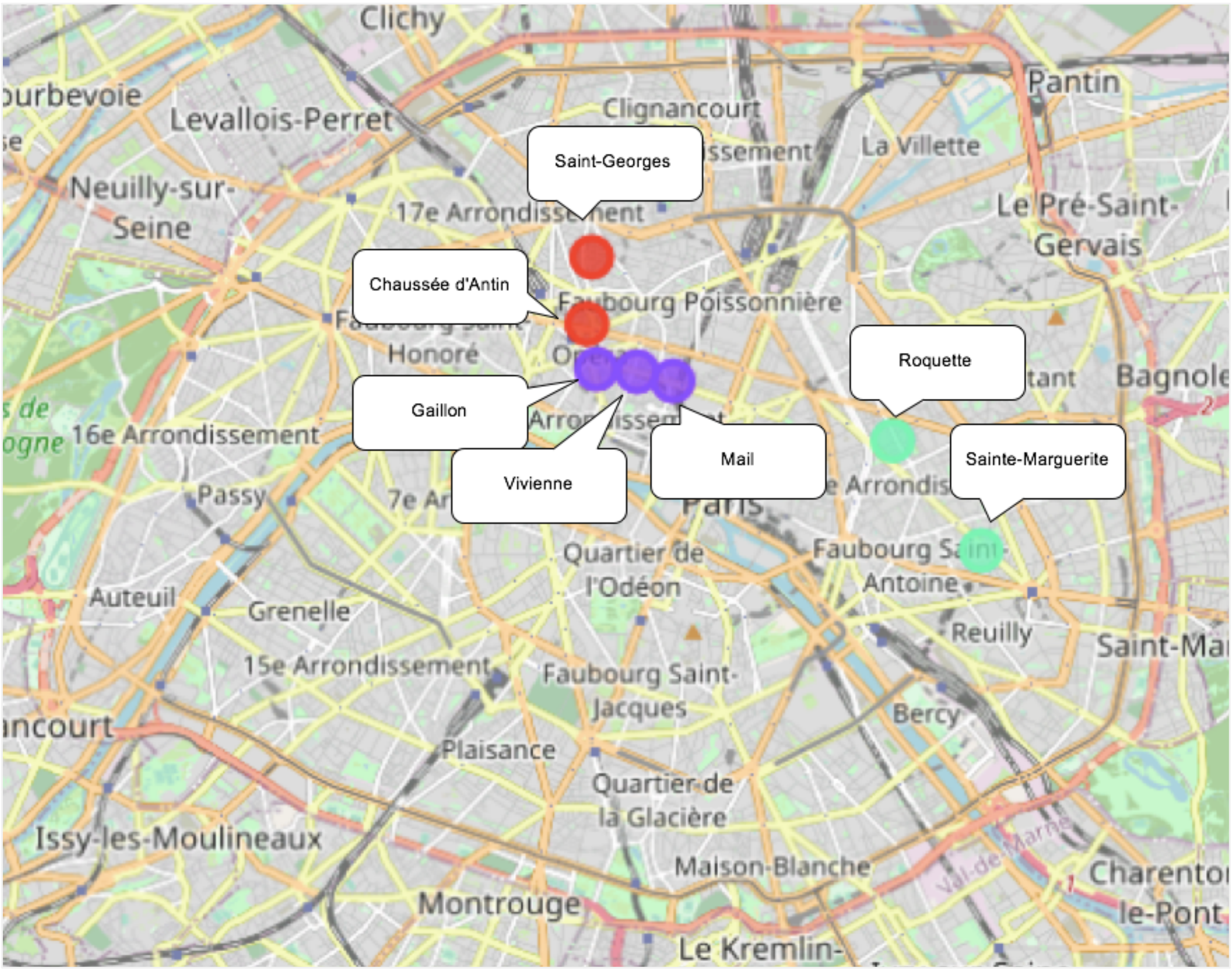
- Cluster 0C1: it contains the neighborhoods of the Bourse borough which presents the highest level of price per square meter in comparison to the 2 other clusters. It is shown in purple on the map.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	Property Price Cluster Label	Target Population Cluster Label	% of people aged 15-29 years per borough	...	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
0	75002	Bourse	Mail	48.868008	2.344699	11250	Mid level	1	0	27	...	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0
3	75002	Bourse	Vivienne	48.869100	2.339461	11250	Mid level	1	0	27	...	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0
2	75002	Bourse	Gaillon	48.869307	2.333432	11250	Mid level	1	0	27	...	30	High level	37	High level	0.0	0	0.0	0.0	0.0	0

- Cluster 0C2: it contains the neighborhoods of the Popincourt borough which presents the lowest level of price per square meter in comparison to the 2 other clusters. It is shown in turquoise blue on the map.

	Postal code	Borough	Neighborhood	Latitude	Longitude	Borough's price per square metre (in €)	Level of price per square metre	Property Price Cluster Label	Target Population Cluster Label	% of people aged 15-29 years per borough	...	% of people aged 30-44 years per borough	Level of % of people aged 30-44 yrs	% of executives and higher intellectual professions per borough	Level of % of executives and higher intellectual professions	% of Vegetarian / Vegan restaurants	Competition Cluster Label	Number of Vegetarian / Vegan Restaurants	Number of V/V restaurants with a delivery service	Number of V/V restaurants with take-away and/or click & collect services	List of Vegetarian / Vegan Restaurants
12	75011	Popincourt	Sainte-Marguerite	48.852097	2.388765	9980	Mid level	2	3	25	...	27	High level	31	Mid level	0.0	0	0.0	0.0	0.0	0
13	75011	Popincourt	Saint-Ambroise	48.862345	2.376118	9980	Mid level	2	3	25	...	27	High level	31	Mid level	0.0	0	0.0	0.0	0.0	0

Visualization of the results:



IV. RESULTS OF THE DATA ANALYSIS

Thus, I have obtained 3 final clusters. So, I want to determine in which of them we can find the best neighborhood to invest in. The 3 clusters correspond to:

- Cluster 0C2: in the Popincourt borough, the neighborhoods Sainte-Marguerite and Roquette present the lowest real estate price, but also the lowest mean of targeted population percentages. As a result, I decide to exclude them from the scope.
- Cluster 0C1: in the Bourse borough, the neighborhoods Gaillon, Mail et Vivienne present the highest levels of targeted population in comparison to the 2 other clusters. However, the price per square meter is also the most important. Consequently, these neighborhoods appear to be more expensive compared to the others.
- Cluster 0C0: in the Opéra borough, the neighborhoods Chaussée d'Antin and Saint-Georges seem to offer a compromise between the two previous clusters. Indeed, they present the second highest level of targeted populations and the second lowest price per square meter.

V. DISCUSSION

Moreover, the map gives additional information. As we can see, clusters 0C1 and 0C0 are bordering in the northern center of Paris, whereas the cluster 0C2 is far away from them in the eastern mid-center. Focusing on the clusters 0C1 and 0C0, I can see that the neighborhood Chaussée d'Antin from the cluster 0C0 is the only one at the junction between the 2 boroughs of Opéra and Bourse, and is adjacent to neighborhoods Gaillon and Vivienne (and knowing that Mail is bordering Vivienne). So, its position would be very interesting for the delivery and click & collect services. However, the Opéra borough encompasses also the two neighborhoods that present the highest percentage of vegan / vegetarian restaurants in Paris. So, an important competition will be closed.

In summary, I recommend to the entrepreneur to consider the neighborhood of Chaussée d'Antin or the neighborhood of Gaillon:

- The entrepreneur can target the Chaussée d'Antin neighborhood, more affordable but with very high nearby competition. This competition could also be seen as a factor of success because the client would be sure to find a high level of consumers.
- Or, he can prefer the Gaillon neighborhood, more expensive, but which is located a little far away from the competitors and which concentrate a high level of targeted populations.

In the next step, the entrepreneur could look at specific premises to install his kitchen in both neighborhoods and compare prices. He could also make a benchmark of the nearby competition to analyze the best practices of the vegetarian / vegan restaurants.

VI. CONCLUSION

During this project, I have used the Foursquare API and different methods of data science and machine learning to obtain a final recommendation on where to open a virtual restaurant. My analysis can be compared to a "Russian doll" as I have conducted a more and more precise and narrowed research using the k-means clustering algorithm.

To conclude, this final project has permitted to develop a realistic data science project leading to a final recommendation. Moreover, I have measured the great interest of what an algorithm can bring to an analysis.