Claire Le Cam 05/22/2020

Search for locations for the establishment of a new yoga studio in Paris

Data acquisition

1. Data sources

The first grid to identify the areas of interest being the Parisian districts, we will use the open data from the data.gouv.fr site (postcode, surface area, coordinates of the centre and its perimeter, INSEE code). The Foursquare API enabled us to obtain all the places (name, category, coordinates) of Paris present in their database. As the API limit is 100 elements per query, we will have to divide the betting area into circles with a radius of 500 m, to be sure to extract the available data set. These data will enable us to identify the yoga studios but also the profile of each district. The number of yoga studios obtained being unrealistic because it is too small, it will be completed thanks to the API of the Sirene V3.0 database of the website ODS (opendatasoft)

The data on the distribution of the population (number, age groups and types of jobs by district) has been downloaded beforehand and is available on the INSEE website. Transport data (names of stations, lines, coordinates) are available on the « iledefrance mobilités » website. Affiliation to the boroughs of these stations will be possible thanks to the Address API of geo.api.gouv.fr, which makes it possible to obtain an address based on geographical coordinates. Finally, the average rental prices come from the geolocaux site. As the site is protected against code scrapping, but the effort being reasonable, the data was manually taken from the source code.

2. Data cleaning

The data obtained through the Foursquare API required first of all the removal of the duplicates obtained because of the methodology used. In addition, the postal code field was filled in more or less rigorously, which necessitated correcting the format of certain elements by deleting strings of characters. For 2-3 items, the postal code was incorrect and had to be changed. Finally, the places with the postcode 'Paris' or 'France' were simply deleted because they were few and irrelevant. Finally, the data were restricted to the districts of Paris (postal code 75001-75020) in order to remove unnecessary data.

Concerning the data from the Sirene V3.0 database, the query was made on the name 'yoga', which necessitated the deletion of data that did not correspond to the type of activity entered. Similarly, it was found that some establishments were duplicated, probably as a result of a move. Studios not currently active were therefore also deleted from the table. In total, half of the studios obtained via the Sirene V3.0 database were deleted.

Concerning the population distribution data, the document provides a large and rather redundant amount of information: e.g. on an age group, then on the same gender age group. After a quick data exploration, I identified the redundancies. For a question of efficiency it was simpler and quicker not to remove them but I did not use them.

The rest of the imported data required no correction.