**Clustering the Cities of Grand Paris**

**IBM Data Science Professional Certificate**

**Capstone Project**

Introduction / Business Problem………………………………………………………………………………………….…….p. 2

Data……………………………………………………………………………………………………………………………….…….…….p. 3

**Introduction / Business Problem**

With the covid-19 crisis, more and more people want to move from Paris, where the housing prices are very high and therefore the house surfaces small, either to a different region or to the area around Paris where the real estate is less expensive.

For the people who want to move from Paris to the area around Paris but who don’t know accurately all cities around Paris, it would be convenient to have a tool capable of proposing some areas corresponding to their needs (transports, employment, education, etc) and their centers of interest (entertainment, sport, etc).

This tool could be used by individuals, could be helpful for estate agents as well.

The idea is to cluster the cities around Paris (officially grouped under the name “Grand Paris”) by the categories of venues they have the most. The people who want to relocate to this area would simply have to choose the cluster that corresponds the best to their needs and center of interests to have the list of cities best suited for them

**Data**

For this study, we will need to list all venues available in the neighborhood of each city of the Grand Paris area. This requires the following steps of data acquisition:

* List the cities of Grand Paris.
* Get their geospatial data.
* From these geospatial data, list the venues available in each city.

**List of cities of Grand Paris.**

A survey in a search engine led to different possibilities, but the most obvious one is the use of open data provided by the French Government:

<https://www.data.gouv.fr/fr/datasets/communes-de-la-metropole-du-grand-paris-par-ept/>

These data can conveniently be downloaded as an excel file, which contains:

* The city names (column "Libellé géographique").
* Some city codes (column"code géographique"), which are not the postal code therefore not directly usable by interfaces using the postal code.
* The codes "région", "département" and "EPT", which are not relevant for our study.

**Geospatial data**

We will use the geocoder python package, which takes the city names in string format as inputs and sends out the longitude and latitude of the city center.

**Venues in each city**

In order to get a list of the venues available in each city, I will use the Foursquare API. Foursquare is a technology company that built a massive dataset of crowd-sourced location data.

By using the Foursquare API, one can get the location, the category, and many information of the venues located in a specified radius around a geospatial position (defined by a longitude and a latitude).