**Real estate pricing in Paris**

A data-oriented price estimation approach

Augustin CURLIER

## Introduction/Business Problem

In a heterogeneous and pricy city such as Paris, France, it might be complicated for buyers to estimate if the price of the apartment they wish to acquire is fair, or for seller to estimate the correct price to which advertise their good.

Price per square meter provided by real-estate professionals is often an averaged value over large administrative districts 'arrondissement'. These estimations do not reflect the fine particularities of neighborhoods, which might have a large impart on the price.

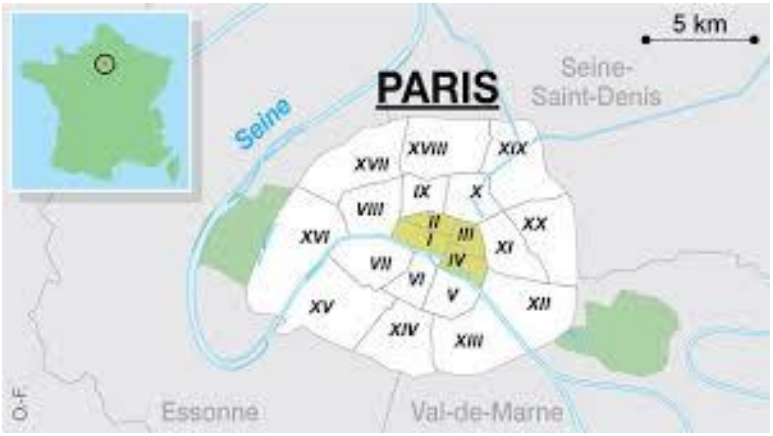


*Figure 1: Paris districts (20 in total)*

The idea here is to develop a model based on publicly available data and machine learning algorithms to come up with a predictor able to yield an estimation of the price per square meter for a Parisian apartment, taking into account more element than just the district onto which the apartment is located.

## Data

In order to setup this estimator, the model will rely on a few publicly available data sources.

First is the list of all transactions completed in France for the year 2018. This dataset is publicly available and counts a few millions real estate transactions. This will be the primary source of information used to build the model :https://www.data.gouv.fr/fr/datasets/r/1be77ca5-dc1b-4e50-af2b-0240147e0346

The second data source used is the official sub-division of paris into smaller administrative districts (Quartiers). This data is used in the form of a geoJSON file and will allow us to present the results using a more interesting meshing than the coarse administrative districts maps : https://opendata.paris.fr/explore/dataset/quartier_paris/download/?format=geojson&timezone=Europe/Berlin

The third data source is OpenStreetMaps, through its tool to search OSM data by name and address (geocoding), called Nominatim. It can be found at nominatim.openstreetmap.org. The GeoPy library, which is a Python 2 and 3 client for several popular geocoding web services allows to access Nominatim

though just a few lines of code. This data will allow to acquire the position of the apartments which have been sold in 2018 using their postal addresses as provided in the transaction dataset.

The fourth data source is Foursquare, accessible through its Python API, which will provide information about the location of popular venues in Paris on a *Quartier*-refined mesh. This will allow to identify 'hot spots' by category (family-friendly, clubs & bar, business, cultural, …) that will necessarily have an impact on the price of the transaction.