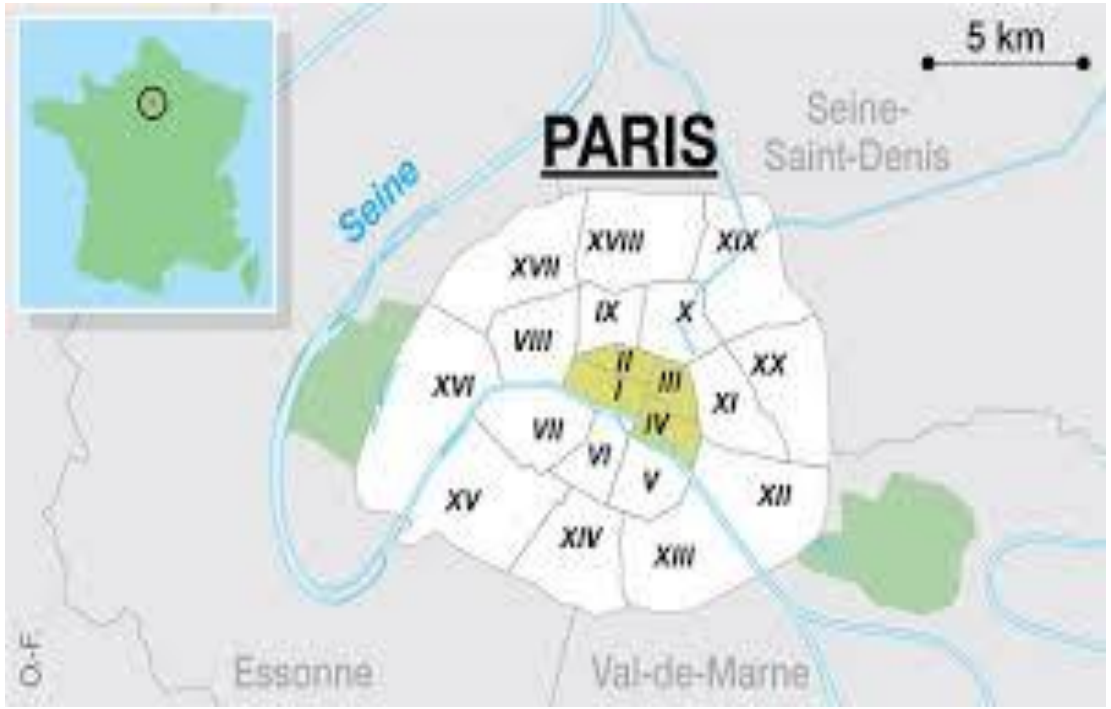
Introduction & issue to be solved

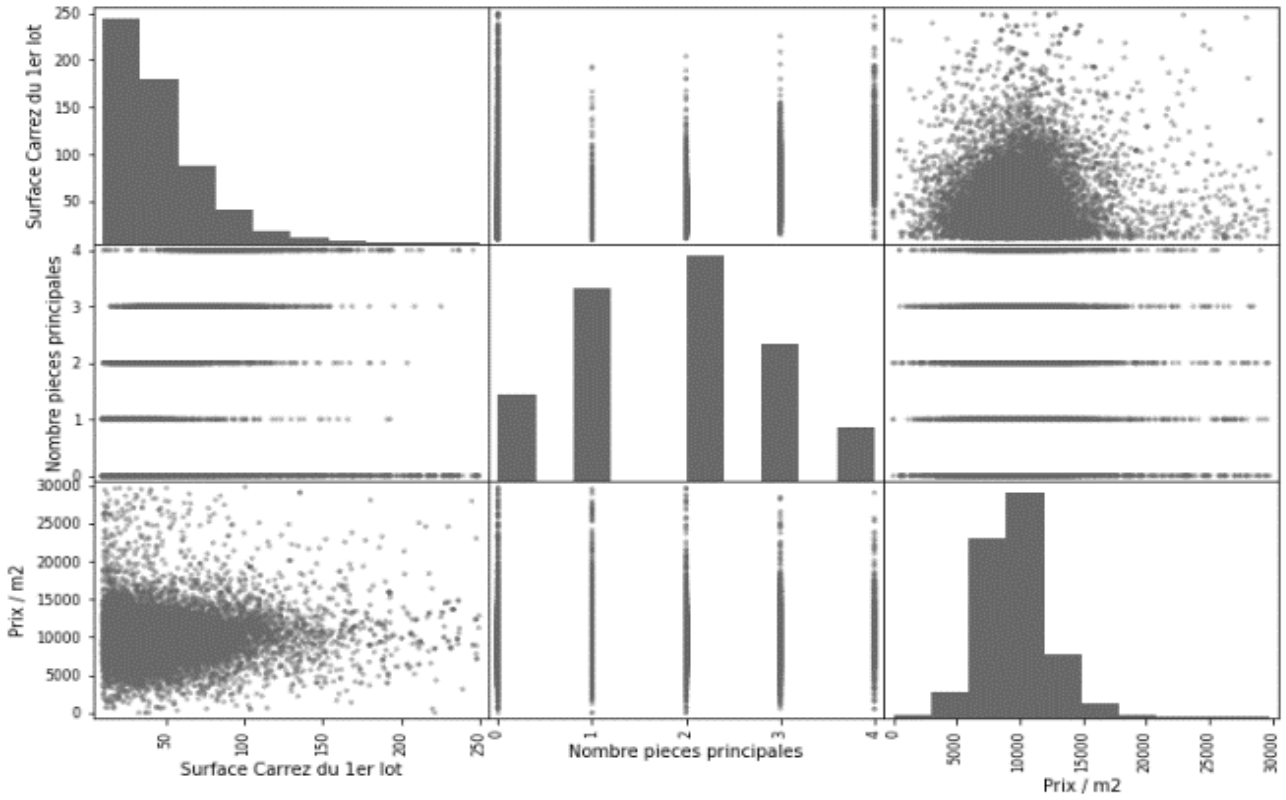


In a heterogeneous and pricy city such as Paris, France, it might be complicated for buyers to estimate if the price of the apartment they wish to acquire is fair, or for seller to estimate the correct price to which advertise their good.

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.

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 used to support the project



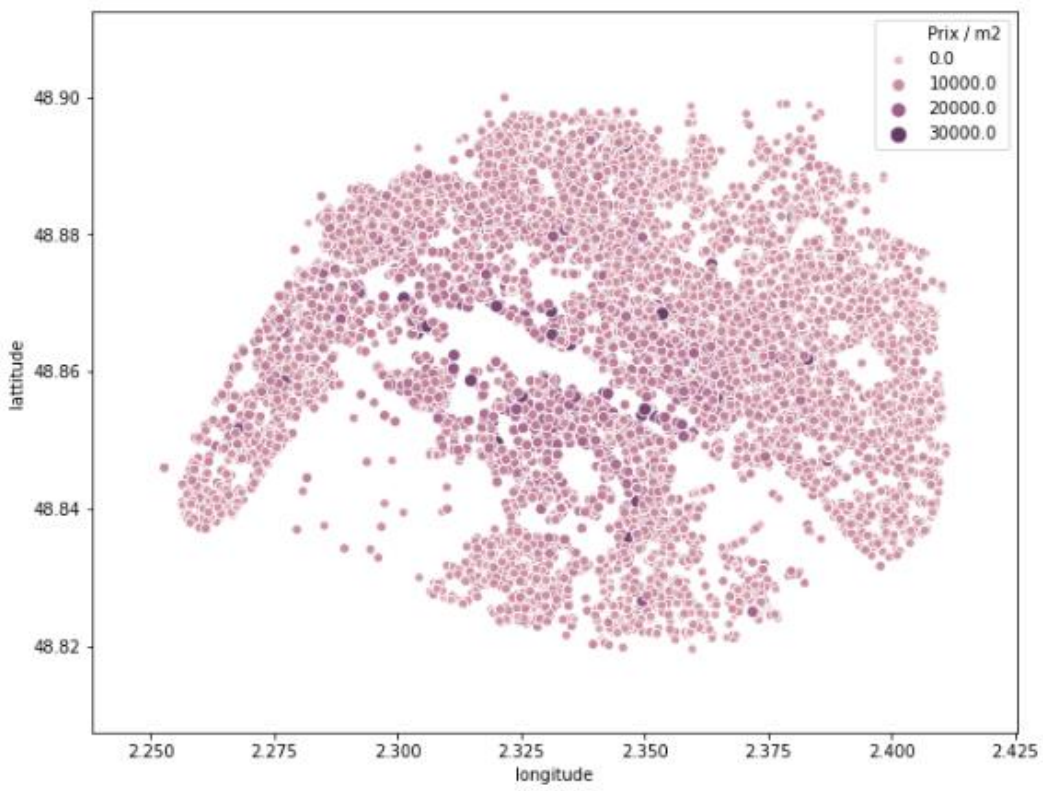
In order to setup this estimator, the model will rely on a few publicly available data sources.

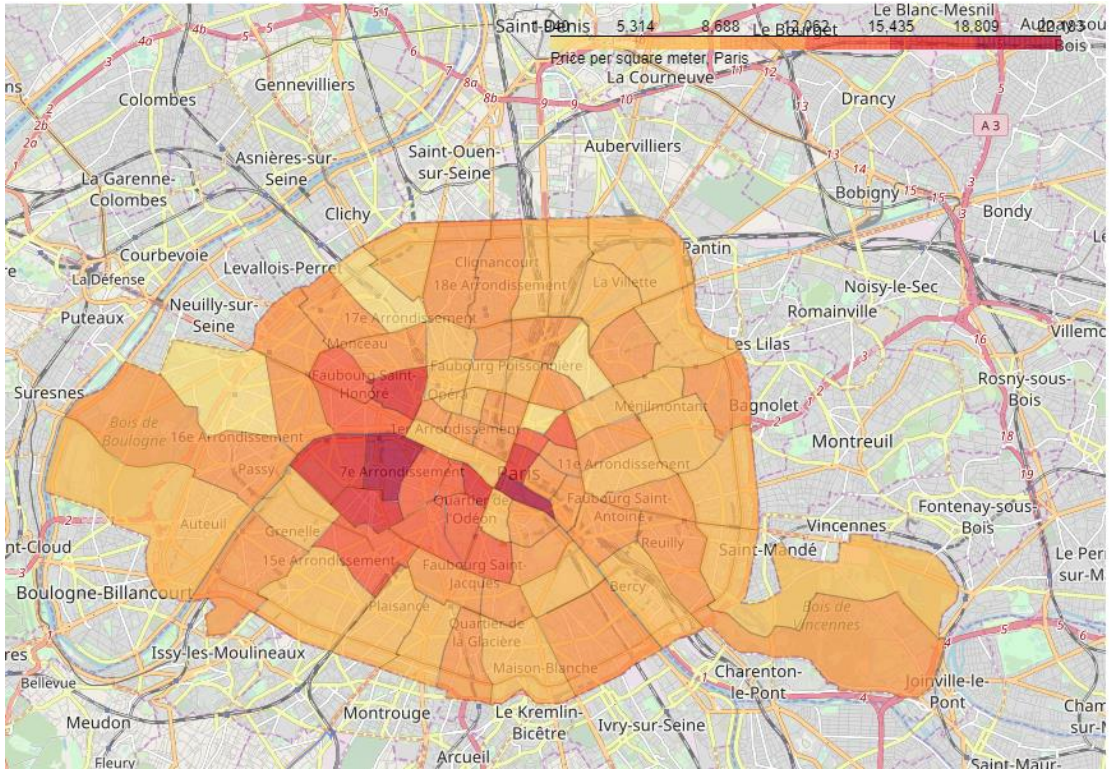
The most significant one is the list of all transactions completed in France for the year 2018. This dataset is publicly available and counts a few millions real estate transactions. This will be the primary source of information used to build the model.

The chart besides represent the scatter matrix of significant data column of this dataset once filter to retain only the data points located in Paris

Each data point corresponds to a good that has been sold in 2018. They are referenced by address. We cannot use the address to properly localize a good, thus we need to use the openstreetmap API to get the latitude and longitude for each of the address of the filtered dataset.

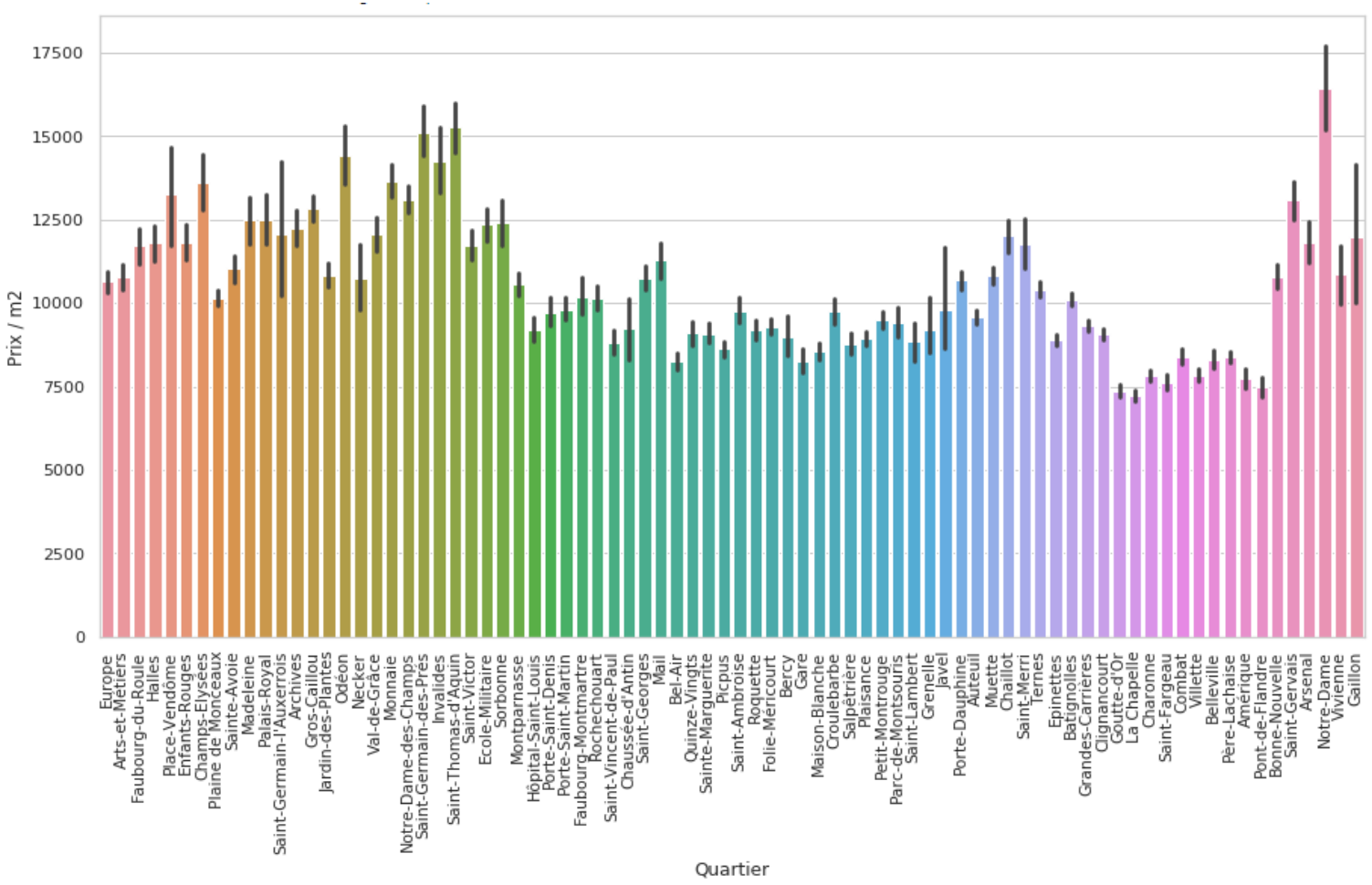
Data maping using openstreetmap API





After downloading, filtering and cleaning, the real estate sales data was transformed using the openstreetmap API to get the latitude and longitude for each of the address of the filtered dataset. Further postprocessing allowed to generate maps showing the distribution of the real estate price

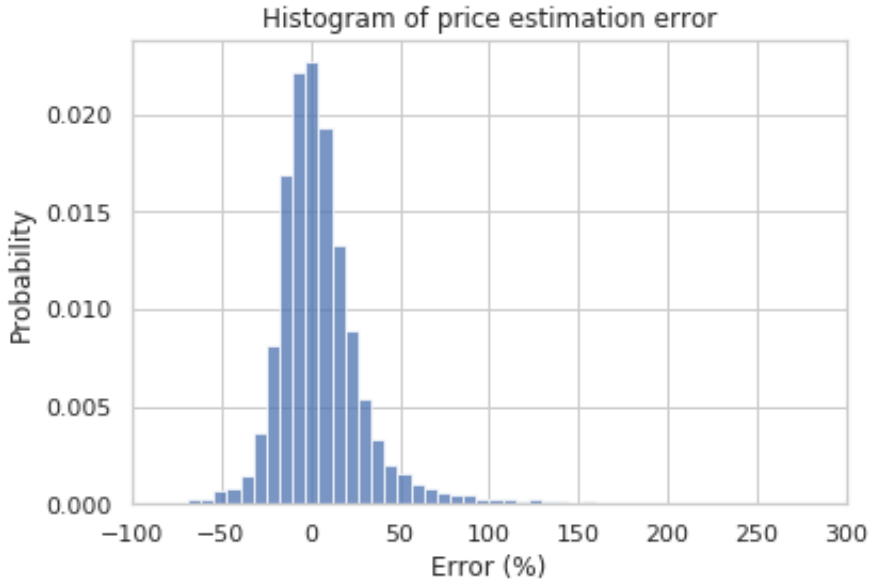
Price distribution per *Quartier* & starting hypothesis



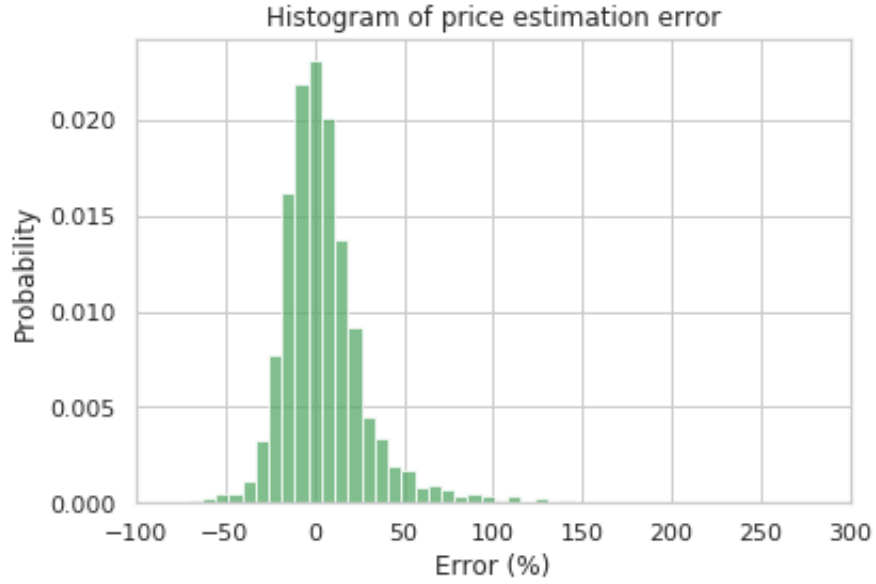
The whole project relies on the fact there are significant variation of price within each large Parisian district. There are four *quartiers* per district, the quartiers are grouped by district on the beside chart.

We can see that there are a lot of discrepancies in the average housing price per square meters for quartiers within the same district. The base assumption seems to be justified.

Price estimation using a multidimensional linear model



Error	
count	10728.000000
mean	5.382573
std	26.773932
min	-69.144668
25%	-9.623880
50%	1.495266
75%	14.640377
max	298.451804



Error	
count	3541.000000
mean	5.344953
std	26.968491
min	-70.624525
25%	-9.800899
50%	1.434652
75%	14.205802
max	301.295419

The following pre-processing was applied to the final dataset:

- One hot encoding the quartiers in which the goods are located (as linear regression cannot deal with categorical variables)
- Normalize the surface and price values using standard scaling

By first training a multidimensional linear model on the full dataset, we end up with the blue error distribution (in % of the actual price).

We can see that the results that we got with the train / test split (in green) are similar. Hence it looks like this model is quite robust and could be deemed as acceptable.

In both case the average error is around 5%, which is not bad.