IBM DATA SCIENCE

PROFESSIONAL CERTIFICATE

CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

# Apartments Sales Prices & Venues Data Analysis of la Métropole Européenne de Lille (MEL)


*Autor :*

Janvion LOUKOU

*Pairs :*

...


17 juin 2020

# Table des matières

# 1 Introduction

Since the 2000s, the price of housing in the French market, in terms of income and rents, is undeniably very high.

To this end, we can emphasize that from 2000 to 2010, on average and across France, the housing price index increased by 107% while the rent index and income per household did not only increased by 27% and 25% respectively.

In 2011, and this for almost 70 years, the price of housing in France was at its peak.

In 2015, 68% of French people lived in a house. Definitely higher than the European average of 58%.

During the years 2016 and 2017, the prices undoubtedly increase. We say more than 0.8% in 2016, and an increase of almost 1.7% in 2017. Not exempt from this increase, 2018 is the year of continuity !

The sales volume remains stable at its highest level. There were 960,000 transactions in 2017, compared to 948,000 in September 2018.

In conclusion, this is a very fluid market which represents more than 3 sales per year for 100 households.

Note that in the first half of 2018, 35% of the 50 largest cities in France saw their average prices decrease.

The Métropole Européenne de Lille (MEL) is not exempt from this increase. Real estate investors are therefore faced with the problem of the profitability of investments in the center of large cities. Thus, the average price per m2 in Lille center is 3,201 euros. Under these conditions, investments are no longer profitable and investors generate zero or even negative cashflows. To overcome this phenomenon, one idea is to make investments in neighboring cities while remaining close to all amenities. This study therefore aims to explore the neighborhoods of Lille to identify small towns in which the cost of real estate transactions would be lower and therefore generate more profit.

# 2 Data available

Data are from the France government site : Demandes de valeurs foncières.

We have two databases :

The first database concerns real estate transactions in France in 2019. This database contains 40 variables and 2,535,791 rows. Among the variables, we find : the nature of the transaction, the value of the property, the name of the town in which the transaction takes place, the surface area of the property, the longitude and latitude of the property, etc... A first choice of variables will therefore be performed for Exploratory Data Analysis (EDA).

The second database from the Wikipedia site concerning information on the municipalities of la Métropole Européenne de Lille (MEL) will be used to extract real estate transactions from Lille. The Jupyter notebook presents all the stages of this data selection.

# 3 Methodology

The data available come from the French government website and relate to real estate transactions all over France. I focus in this study on the transactions carried out in *la Métropole Européenne de Lille (MEL)*, in the north of France. The objective is to explore the neighborhood of Lille and determine the municipalities which are interesting for a real estate investment.

## 3.1 Select data of la MEL

To select the transactions of la MEL, I download information about the municipalities of Lille from the Wikipedia site la MEL municipalities. From this information I collect the data concerning Lille in our database on real estate transactions.

```
(95, 6)
```

| | Nom | Code Insee | Gentilé | Superficie (km2) | Population (dernière pop. légale) | Densité (hab./km2) |
|---|---|---|---|---|---|---|
| 0 | Lille(siège) | 59350 | Lillois | 3483 | 232 787 (2017) | 6 684 |
| 1 | Allennes-les-Marais | 59005 | Allennois | 555 | 3 460 (2017) | 623 |
| 2 | Annœullin | 59011 | Annœullinois | 901 | 10 428 (2017) | 1 157 |
| 3 | Anstaing | 59013 | Anstinois | 23 | 1 469 (2017) | 639 |
| 4 | Armentières | 59017 | Armentiérois | 628 | 24 882 (2017) | 3 962 |

FIGURE 1 – *Data about la MEL municipalities head.*

(42395, 40)

| date_mutation | numero_disposition | nature_mutation | valeur_fonciere | adresse_numero | adresse_suffixe | adresse_nom_voie | adresse_code_voie | code_postal | c |
|---|---|---|---|---|---|---|---|---|---|
| 2019-01-04 | 1 | Vente | 190500.0 | 1.0 | NaN | ALL DU GENERAL KOENIG | 0488 | 59130.0 | |
| 2019-01-04 | 1 | Vente | 190500.0 | 1.0 | NaN | ALL DU GENERAL KOENIG | 0488 | 59130.0 | |
| 2019-01-04 | 1 | Vente | 116000.0 | 26.0 | NaN | AV DU MARECHAL DE LATTRE | 0650 | 59350.0 | |
| 2019-01-04 | 1 | Vente | 116000.0 | 26.0 | NaN | AV DU MARECHAL DE LATTRE | 0650 | 59350.0 | |
| 2019-01-04 | 1 | Vente | 8000.0 | 14.0 | NaN | AV DE LA ROSERAIE | 7609 | 59000.0 | |

FIGURE 2 – *Data about real estate transactions head.*

## 3.2 Features selection

The variables of interest for the analysis are : *nature_mutation* : the type of the transaction, *valeur_fonciere* : the value of the property, *nom_commune* : the name of the municipality, *type_local* : the type of premises, the surface of the property, the number of rooms, the longitude and the latitude. The figure (3) shows the DataFrame after *feature selction*.

| | nature_mutation | valeur_fonciere | nom_commune | nombre_lots | type_local | surface_reelle_bati | nombre_pieces_principales | longitude | latitude |
|---|---|---|---|---|---|---|---|---|---|
| **1432150** | Vente | 190500.0 | Lambersart | 2 | Appartement | 81.0 | 4.0 | 3.015595 | 50.651965 |
| **1432151** | Vente | 190500.0 | Lambersart | 1 | Dépendance | NaN | 0.0 | 3.015007 | 50.652149 |
| **1432152** | Vente | 116000.0 | Saint-André-lez-Lille | 1 | Appartement | 42.0 | 2.0 | 3.045852 | 50.655648 |

FIGURE 3 – *Data about real estate transactions after feature selection.*

## 3.3 Data cleanning

### 3.3.1 NAs imputation

Three columns contain many missing values (4) : local type (27%), area (45%) and number of main rooms (27%). The figure (5) shows the percentage of missing value per variable. We nevertheless keep all the columns because the surface of the property for example is important in determining the price of the property per m2. We therefore delete the lines containing missing values.

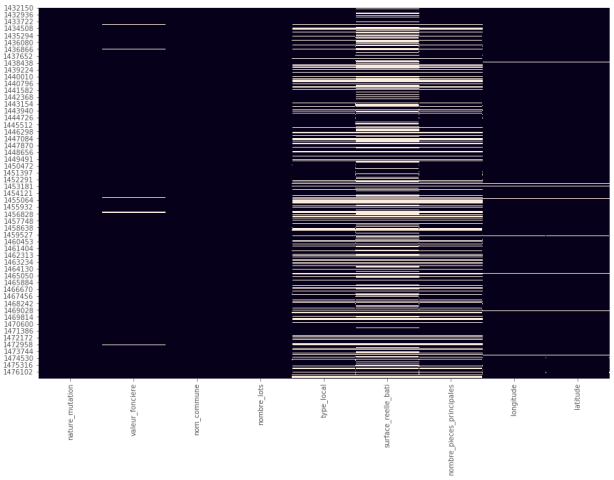


FIGURE 4 – *Missing value by column.*

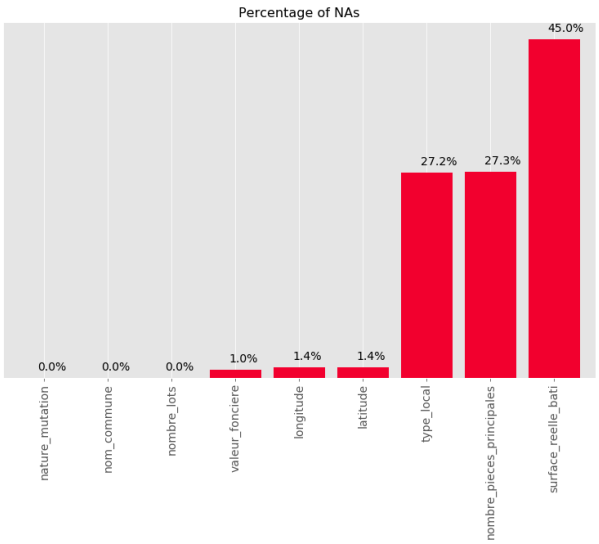


FIGURE 5 – *Rate of missing value by column.*

### 3.3.2 Outliers imputation

The column value of transaction (*valeur_fonciere*) contains many outliers. However a transaction can be considered as an outlier depending on the local type. Indeed we know for example that on average houses are expensive than apartments. So to deal with outliers, we also consider the local type. This is why we are going to remove the outliers by type of local. So we remove the outliers associated with the house, the outliers associated with the apartment and finally the outliers associated with industrial premises.

## 3.4 Exploratory Data Analysis

### 3.4.1 Transaction type

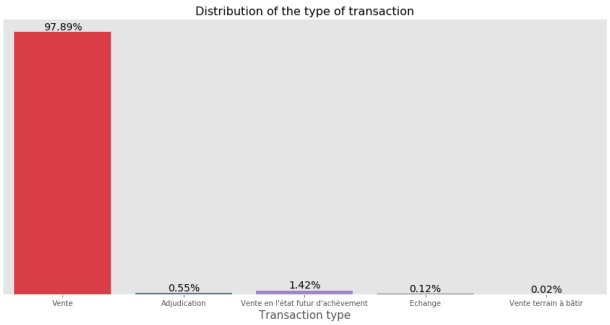


FIGURE 6 – *Rate of transaction type.*

The most successful transaction is the *sale* (98%). We also notice that there are fewer transactions concerning the *sale of building land*. People no longer buy land to build houses. They prefer to buy an already built house directly.

### 3.4.2 Local type

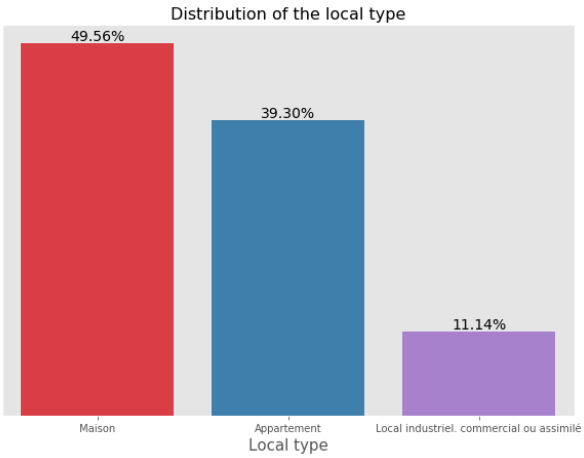


FIGURE 7 – *Rate of local type.*

Half of the transactions carried out in the MEL concern houses (49.56%). The apartments come in second place, followed by industrial premises (11%).

### 3.4.3 Municipalities

Unsurprisingly, the City of Lille is the one with the most transactions (around 6,000), followed by Tourcoing and Roubaix.

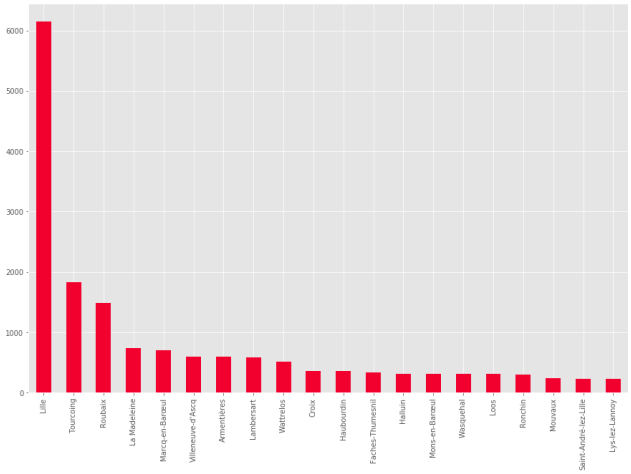


FIGURE 8 – *20 city with the most transactions in la MEL.*

## 3.5 Explore Neighborhoods in Lille

We get the coordinates of Lille using the geopy library. We can then represent the MEL maps.

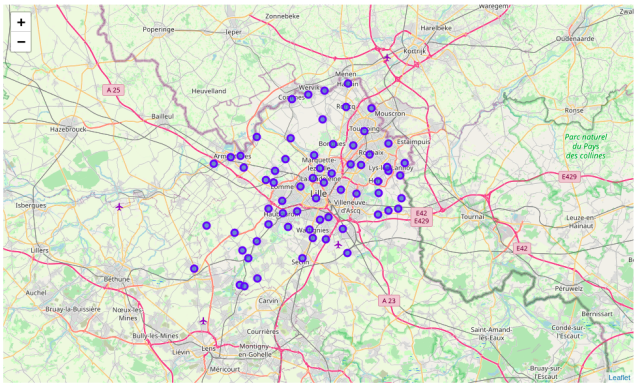


FIGURE 9 – *Maps of the Lille Neighborhoods.*

By grouping the venues by municipality, we can observe the cities with the most venues in the table. In total we have 126 different venues.

| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| **Lille** | 44 | 44 | 44 | 44 | 44 | 44 |
| **Roubaix** | 25 | 25 | 25 | 25 | 25 | 25 |
| **Tourcoing** | 20 | 20 | 20 | 20 | 20 | 20 |
| **Mouvaux** | 18 | 18 | 18 | 18 | 18 | 18 |
| **Halluin** | 16 | 16 | 16 | 16 | 16 | 16 |
| **Croix** | 12 | 12 | 12 | 12 | 12 | 12 |
| **Saint-André-lez-Lille** | 11 | 11 | 11 | 11 | 11 | 11 |
| **Comines** | 10 | 10 | 10 | 10 | 10 | 10 |
| **Armentières** | 9 | 9 | 9 | 9 | 9 | 9 |
| **Villeneuve-d'Ascq** | 8 | 8 | 8 | 8 | 8 | 8 |
| **Roncq** | 8 | 8 | 8 | 8 | 8 | 8 |

FIGURE 10 – *Municipalities with the most venues in la MEL.*

## 3.6 Clustering

We use the Kmeans algorithm (suitable for clustering) to create clusters of the municipalities of la MEL using the frequency table. Using the Elbow method, we choose 5 clusters.

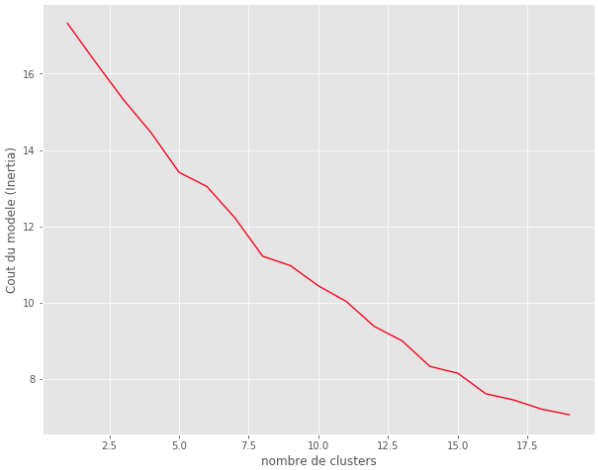


FIGURE 11 – *Elbow method for the choice ef the number of clusters.*

# 4 Results

## 4.1 Clusters maps

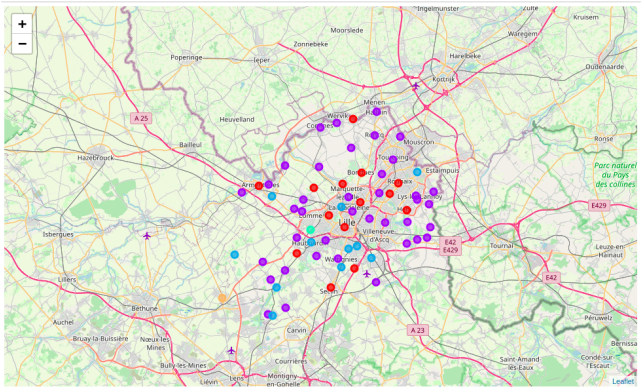


FIGURE 12 – *Clusters Maps of la MEL.*

## 4.2 Examine clusters

| | Venues | price per m2 (euros) | Number of rooms | Comment |
|---|---|---|---|---|
| Cluster 0 | French restaurant | 3073 | 2.5 | cheapest \| venues |
| . | | . | . | |
| Cluster 1 | Pizza place<br>Supermarket | 2816<br>. | 2.5<br>. | cheapest \| venues |
| . | Women's stores | . | . | |
| Cluster 2 | Train station | 3330 | 2.5 | average price |
| . | | . | . | |
| Cluster 3 | | 3340 | 3 | average price \| no venues |
| . | | . | . | |
| Cluster 4 | . | 3769 | 3 | expensive \| no more venues |
| . | | . | . | |

TABLE 1 – Clusters analysis

# 5 Discussion

To carry out this study, we focused on la MEL, in the Nord department. We could extend the study to other departments to compare prices by department. Our study was also based on the sale of

apartments. We could also carry out a similar study on the sale of houses for people who are thinking of buying family houses.

# 6 Conclusion

Taking into account the results of the study, Cluster 1 is more advantageous for someone looking to invest in real estate by buying apartments. Apartment prices are cheaper there, and there is enough venues for people to settle there. The municipalities of Clusters 0 also seem to adapt for this type of project. On the other hand, the municipalities of Clusters 3 and 4 are not suitable for rental property investment.