Choosing the perfect neighborhood

Pierre Beylard

11 May 2020

Introduction

Problem background

Considering some time limitation and Foursquare API requests limit, we will focus our work on the city of Bordeaux in France.

Bordeaux metropole is constituted of 28 towns dispatched on an area of 578 square kilometers[1]. The city is close to the Atlantic Ocean on the west part of France

Bordeaux is a city with a relatively rapid demographic growth around 1.5% per year during the previous 7 years. In 2017, Bordeaux had 760 000 inhabitants with an objective of 1 million inhabitants in 2030.

To achieve this objective, new neighborhoods are constructed. Furthermore, transport within the city are developed.

Of course, this growth is not only natural, and immigration plays an important role: A lot of people are starting to arrive in the city from other French cities. This growth also attracts new companies with an average of 2000 new jobs created per year thanks to these new companies[2].

Nowadays, in Bordeaux, when we talk about real estate business, it seems clear that is at the center of a lot of expectations:

- For project owners:
    - The choice of a neighborhood to implement a business will be critical for its future success (competition, potential customers, attractivity of the neighborhood…) ;
    - When it comes to invest in a house, the decision of the neighborhood must be thoroughly deliberate as it will impact people quality of life (activities in the neighborhoods, school, shops, parks, distance to their work…) .
- For investors, this choice will have a direct impact on their investments and the profitability.

In other terms, real estate has a direct financial impact combined with a strong social impact on population lives.

Objectives – Analytical approach

Our goal here is to provide insight on the different neighborhoods that will help decider to choose the correct place to invest.

[1] Source : http://www.bordeaux.fr/p287/bordeaux-en-chiffres

[2] Source : https://www.20minutes.fr/bordeaux/2248127-20180403-bordeaux-projets-entreprises-metropole-creent-plus-plus-emplois

Choosing a correct business emplacement will depend on several aspects: the type of business, the target, the population density, the competition, price per square meter of the local …).

Choosing a correct place to live will depend also on several aspects : the age of the buyer, the family structure (single, couple, kids…) , their hobbies, the place of their work, commodities, transport services, price per square meter, type of housing facilities ….

Finally, investors will be mainly interested in the capacity of the borrowers to pay of their loan, but they will also be interested in the potential price trends of the neighborhood in order to secure their investment.

In order to achieve our goal of showing relationships between neighborhoods, a descriptive approach will be conduct. We will aggregate neighborhoods in clusters depending on the following information:

- Real estate price
- Most common type of real estate properties (apartments, houses…)
- Principal venues of the neighborhoods

Data

Data requirements & collection:

As told above, we will need some critical information to construct our model:

- **Foursquare API** to find the main venues of given Bordeaux Metropole borough or towns
- **OpenData** [3] Bordeaux provides some accurate information on Bordeaux Metropole towns, as location of borough and towns in GeoJSON format. Other information as locations of main public transport stop or companies are available. But we will use Foursquare to achieve this objective.
- **DVF tool** [4] is a dataset provide by the French government regrouping the real estate transactions intervened during the last five years on the French metropolitan territory and the DOM-TOM, except for Alsace-Moselle and Mayotte. This dataset is not exhaustive as it regroups the information only on the transactions passed during the 5 last years. Even if all the information will not be available, it will give us a pretty good insight on the structure of Bordeaux real estate market (*price, house/flat surface area, number of rooms in the transaction, address of the building*….). in order to access this information, we have two choices: download txt files per year or use an unofficial API. We will use the API.

Methodology:

Data understanding and preparation:

We divided our work in 3 parts:

- Obtain the correct and accurate coordinates of the neighborhoods constituting the metropole.
- Obtain information on real estate market

[3] https://opendata.bordeaux-metropole.fr/

[4] https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres/

- Obtain local venues for each neighborhood.

Bordeaux Neighborhoods:

When started to work on the coordinates of neighborhoods, we rapidly faced an issue on accuracy of the data. In France, every town is represented by a unique INSEE Code. As we know that the Bordeaux metropole is constituted of 28 towns, we count the different number of INSEE codes contained in our dataset to see if we were accurate:

```
Entrée [345]: 1 bordeaux_neighborhoods['INSEE'].value_counts()

Out[345]: 33318    17
          33281    10
          33063     8
          33119     3
          Name: INSEE, dtype: int64
```

We only had 4 towns on a total of 28.

We had to find other datasets to have an accurate vision of metropole towns and neighborhoods.

We first scraped a governmental [5]page to check which towns were not included in this first data set by taking only the town name(communes) and its associated INSEE code:


```
[12]: 1 df

Out[12]:
```

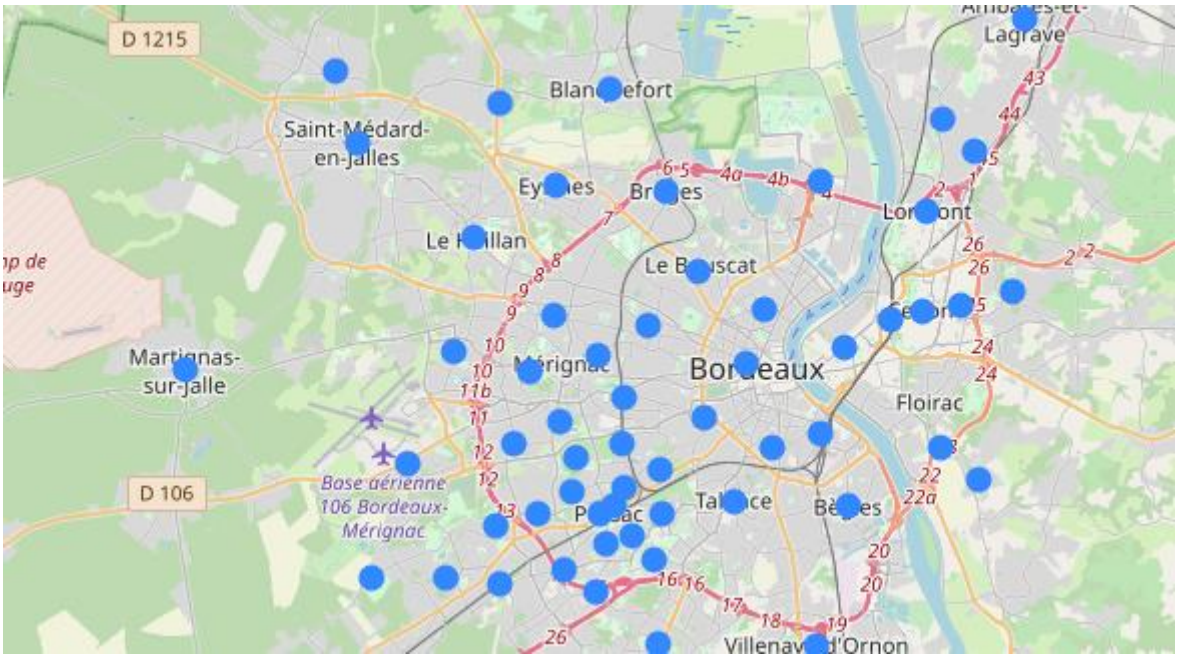
	Communes	INSEE
0	Ambarès-et-Lagrave	33003
1	Ambès	33004
2	Artigues-près-Bordeaux	33013
3	Bassens	33032
4	Bègles	33039
5	Blanquefort	33056
6	Bordeaux	33063

Then, we found another dataset [6]containing all the towns in the county of Gironde (where Bordeaux Metropole is located). We add only the towns of the metropole that were not in the first data set thanks to the INSEE Code (present in the first dataset and scraped from the webpage).

Thanks to these three steps we had an accurate dataset of Bordeaux metropole and neighborhoods.

[5] https://www.insee.fr/fr/metadonnees/cog/intercommunalite-metropole/EPCI243300316-bordeaux-metropole

[6] https://github.com/gregoiredavid/france-geojson/tree/master/departements/33-gironde



Real estate information:

Here again, we perform some basics data analysis to explore the dataset:

- See the shape of the data frame

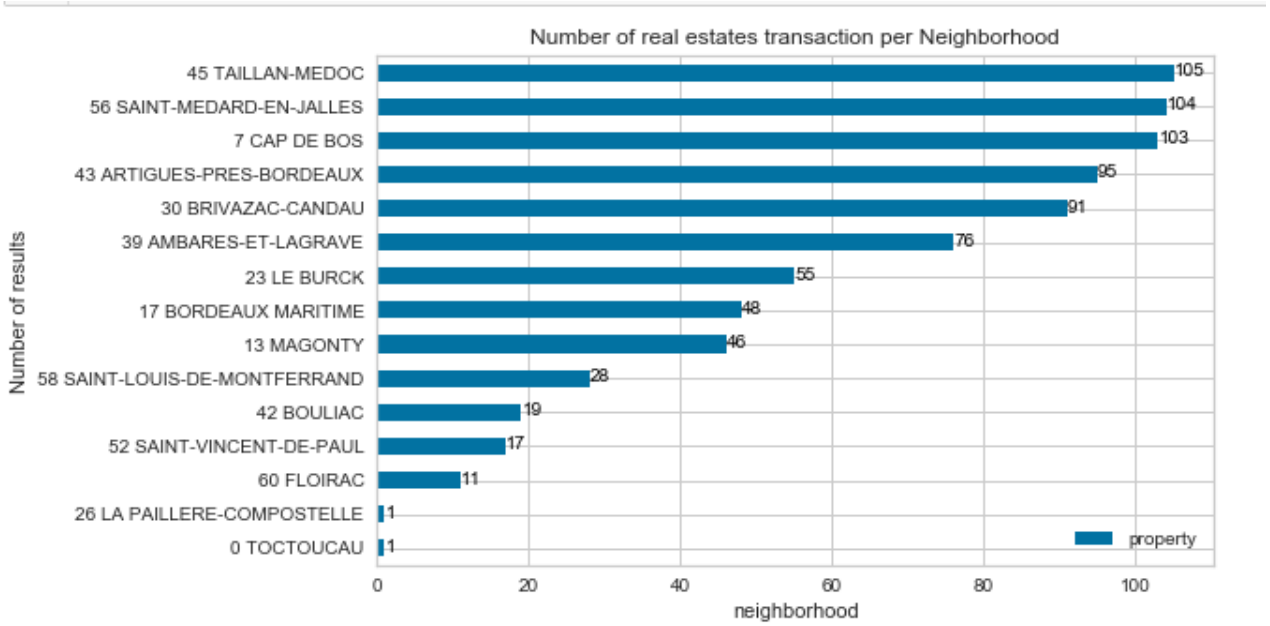
```
Entrée [369]:  1 dvf_values.shape
Out[369]: (154966, 11)
```

- Check presence of null values:

Our purpose is to obtain an average price per square meter. To achieve this, we need to drop all potential rows without information on 'area'. After achieving that, we still have more than 80 000 results.

```
Entrée [199]:  1 dvf_values_clean.shape
Out[199]: (81164, 11)
```

- Check the relevance of data, with this amount of results for only 62 neighborhoods, we may think that the data is accurate. However, after a short analysis, we found that 2 neighborhoods had only one result:

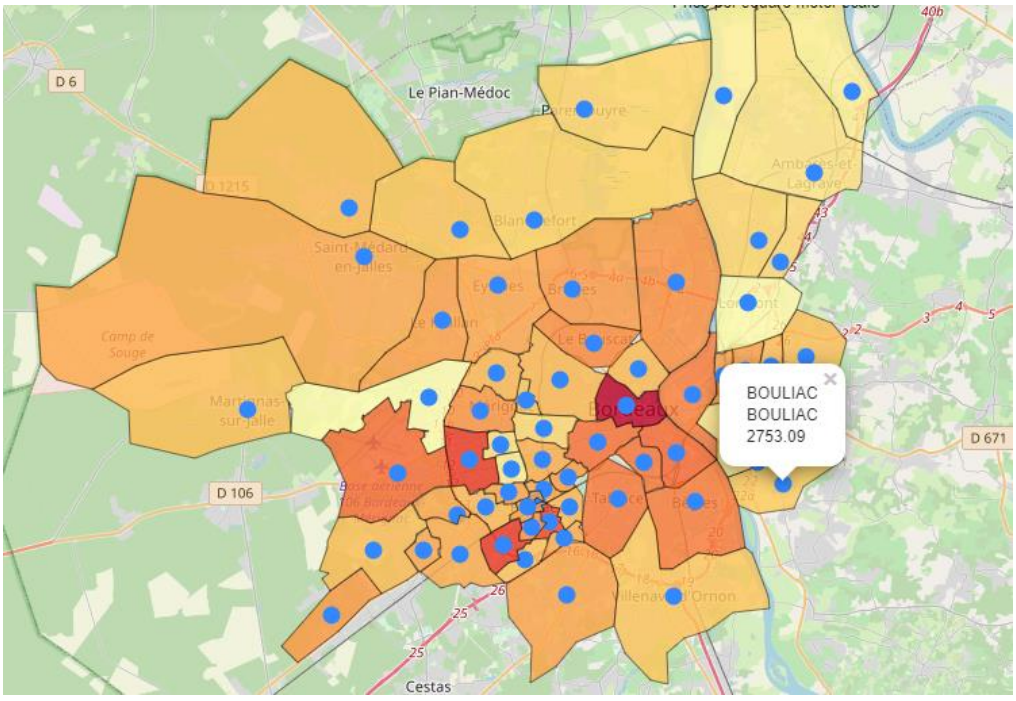


We replaced the result average price per square meter of LA PAILLERE-COMPOSTELLE and TOCTOUCAU as one result is not enough to confirm the potential price of the neighborhood. We used the average price of their town to calculate the new price for these neighborhoods.

To finally prepare the real estate data, we have operated two last modifications:

1. We have calculated the average price per square meter per transaction
2. We have grouped the data per neighborhood by calculated the median price per square meter per neighborhood and keeping only the most recurrent property type per neighborhood:

By using the geometry data obtained in the previous data set we have displayed the information of price per square meter and most common property type on a choropleth map:



Foursquare venues:

At first sight Foursquare returned 3 648 venues for the 62 Bordeaux metropole identified neighborhoods. But after a quick analysis, it seems that foursquare has returned some duplicated venues.

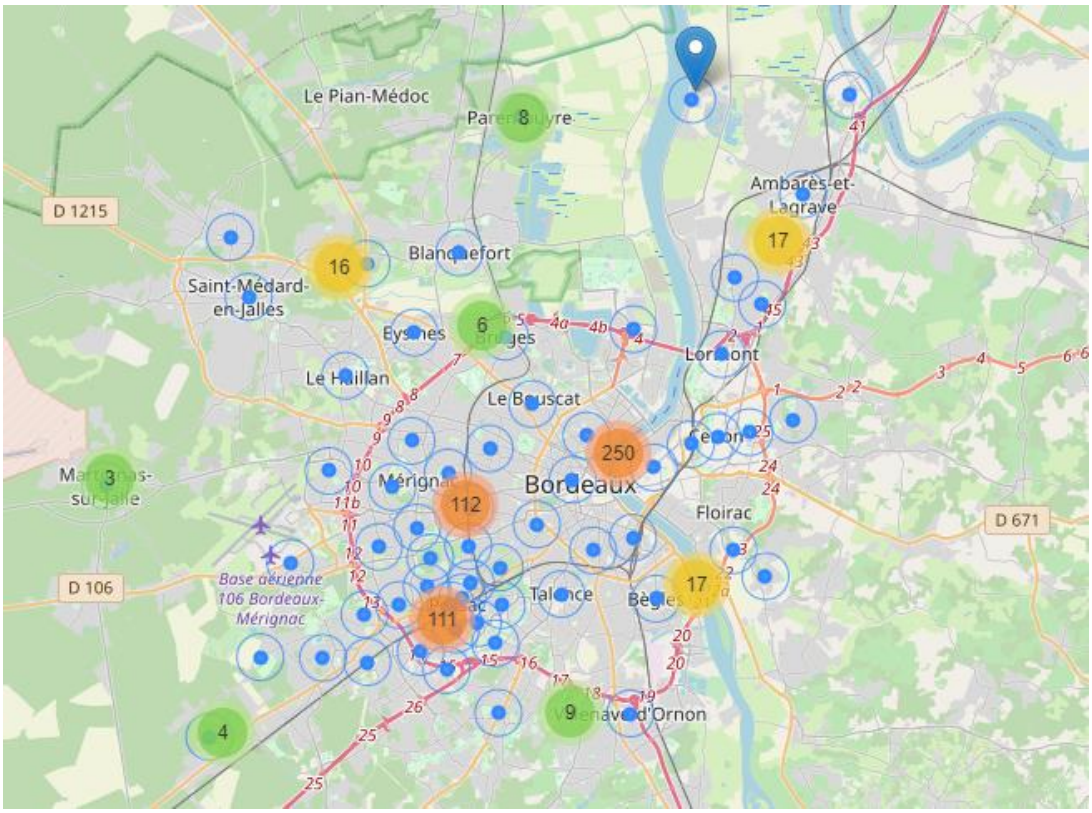
Our first task is to clean these duplicates by dropping them. By doing that, we have lost 80% of the returned venues to a total of 556 venues.

Some neighborhoods had less than 3 returned venues and even only one venue.

ID	venue_name
58 SAINT-LOUIS-DE-MONTFERRAND	1
13 MAGONTY	1
25 BEUTRE	2
52 SAINT-VINCENT-DE-PAUL	2
61 AMBES	2
45 TAILLAN-MEDOC	2
59 SAINT-AUBIN-DE-MEDOC	3
0 TOCTOUCAU	3
49 MARTIGNAS-SUR-JALLE	3

In order to understand why we had this few number of venues, we made a folium map displaying :

- Neighborhoods,
- A 750 meters circle around the neighborhood's coordinates
- We also add all the venues to marker clusters.



By looking at the map, it seems that suburban neighborhoods have less results than Bordeaux city center. By focusing on a single neighborhood, we can see that returned venues can be anywhere into the radius of 750 meters set into the API request. Furthermore,

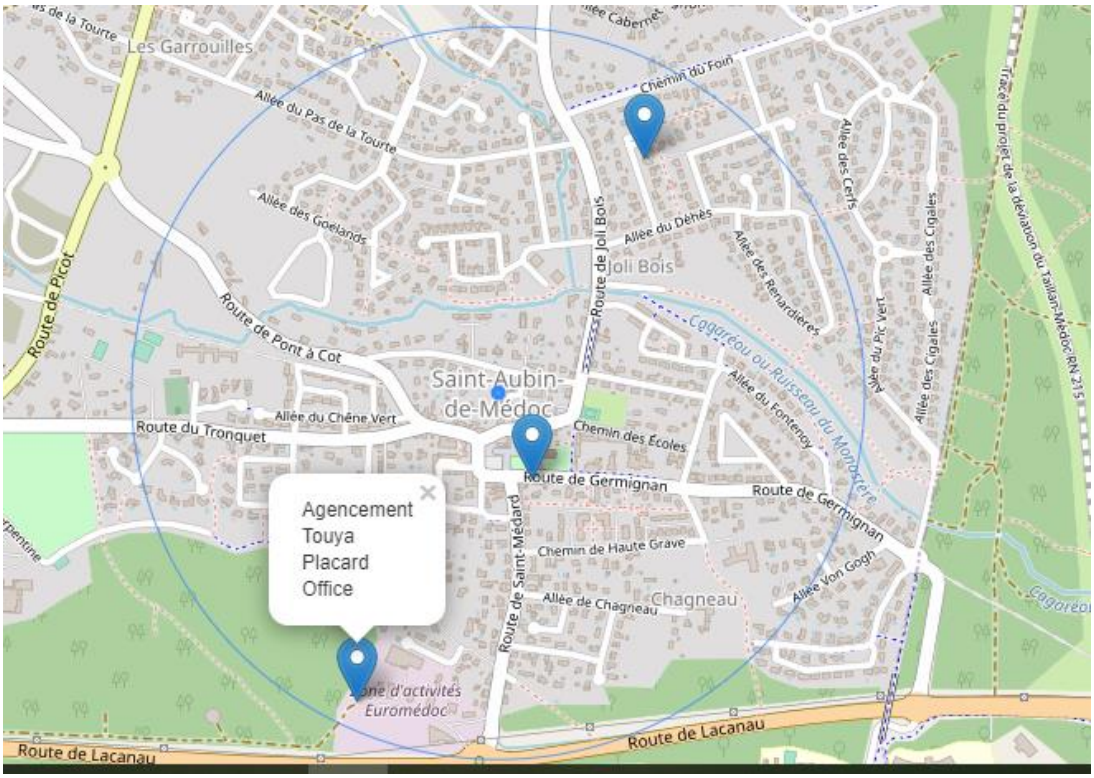


Figure 1Focus on Saint Aubin du Médoc

Figure 2-Zoom into Saint Aubin du Médoc cluster center

By doing this short analysis, we can understand that the problem does not come from the coordinates of neigborhoods neither from the radius but from a lack of venues into the database of Foursquare.

In order to prevent two solutions may be used :

- Change coordinates of neighborhood centers,
- Enlarge radius to capt more venues

In our case, many suburbs towns with less results cover a large area, in order to capture more results, we need to enlarge radius of research.

By doing it, we raised our first results to 4111 venues and after dropping duplicates to 820 venues (raise of 40%).

Even with this improvement some towns had not enough results to be clustered in an efficient way...

To improve it, we looked at the Foursquare classification tree and see two things:

- We obtained 173 different venues categories for a total of 800 venues
- These categories are level 2 or 3 from the classification tree.

The parent categories are less numerous (7 parent categories) and provide enough details to classify neighborhoods. For this reason, we decided to reclassify the neighborhoods into their parent categories:

```
Entrée [130]: 1 bordeaux_venues_clean['parent_category'].value_counts()

Out[130]: Food                        265
          Shop & Service              215
          Travel & Transport          130
          Outdoors & Recreation        98
          Arts & Entertainment         62
          Nightlife Spot               38
          Professional & Other Places  12
          Name: parent_category, dtype: int64
```

Modeling

We aim to cluster our neighborhoods; several models exist like:

Method name	Parameters	Scalability	Usecase
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry
Ward hierarchical clustering	number of clusters or distance threshold	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes
OPTICS	minimum cluster membership	Very large `n_samples`, large `n_clusters`	Non-flat geometry, uneven cluster sizes, variable cluster density
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.

In our case, we cannot use the density-based algorithm as we are not dealing with spatial cluster. Furthermore, we have test K means and agglomerative clustering that provided similar results. But agglomerative clustering seems to have a better 'noise' analysis. We have conducted a short visual analysis based on our knowledge of the metropole and by comparing the delta between the two algorithms. For this reason, we will focus our results analysis on the hierarchical algorithm.

However, to select the perfect number of clusters, we have conducted two tests: The Silhouette and the elbow methods. Both proposed to dispatch our neighborhoods into respectively 5 and 4 clusters:

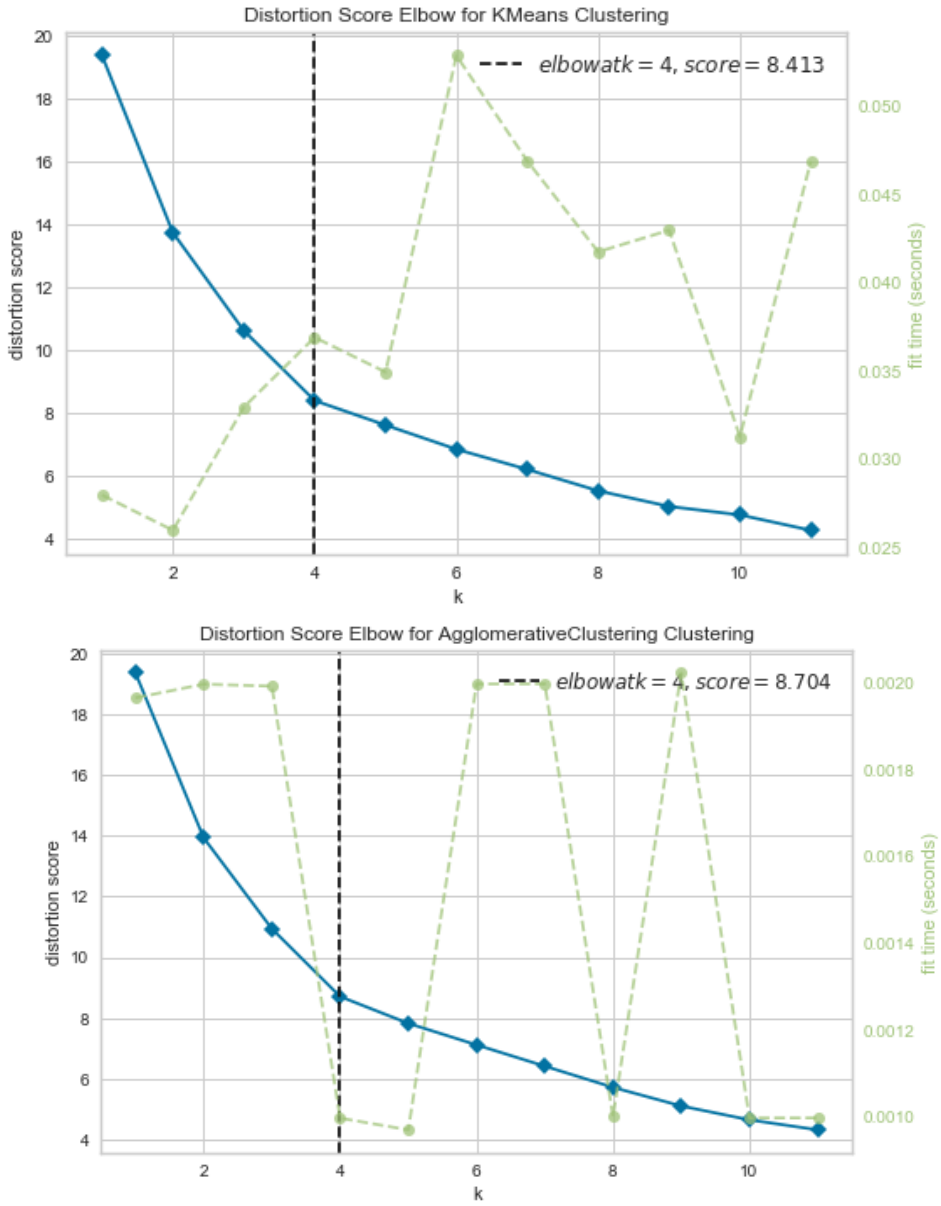


Figure 3- The Elbow Method comparison between Kmeans and agglomerative clustering

For both clustering models, the elbow method preconizes to split our dataset into 4 clusters.

To confirm this, we will also conduct another analysis of optimal clusters number by using the silhouette analysis. Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1][7].

[7] https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

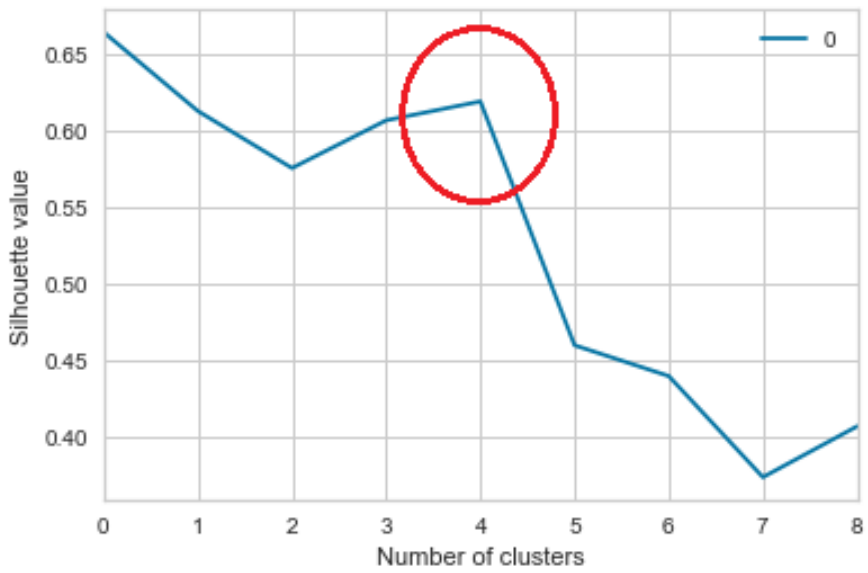


Figure 4- The Silhouette method K means algorithm

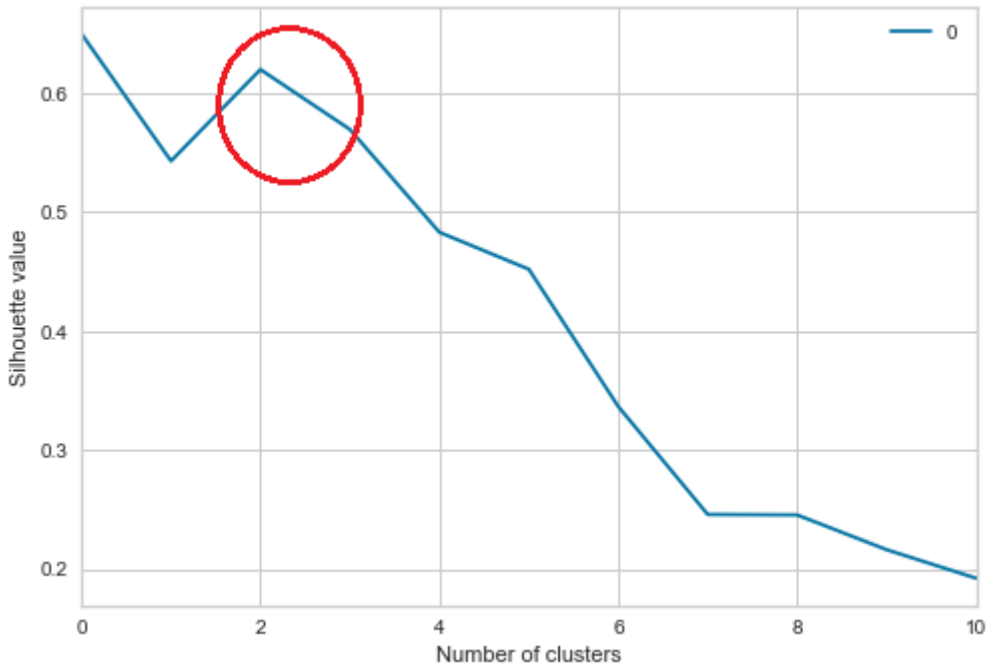


Figure 5-Silhouette method agglomerative clustering

The silhouette score preconizes a 2 clusters approach in the agglomerative method, but the 4 clusters approach can be used in our case. Indeed, silhouette coefficients near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or remarkably close

to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. For k=4, we are still at a score of 0.5 that means, that the clusters are still far away to one another. As it does not make sense to separate neighborhoods into 2 clusters, we will use a k of 4 for clustering our neighborhoods.

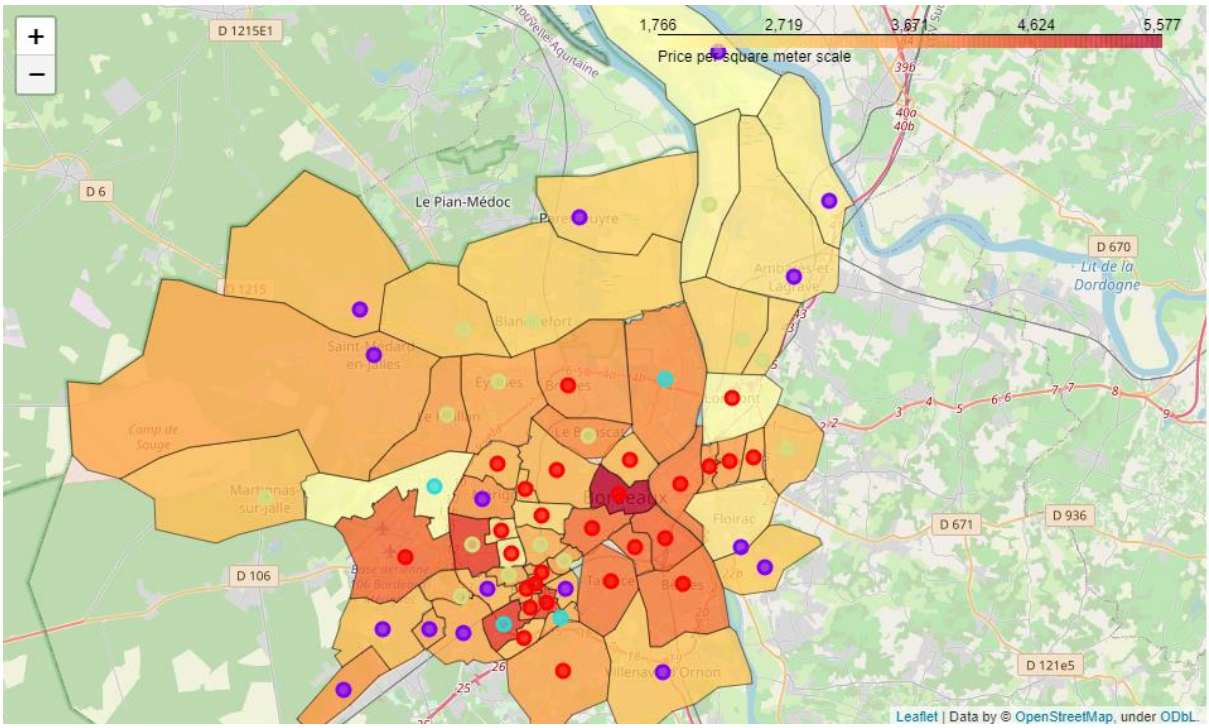


Figure 6 Clusters clhoropeth map

Results

Evaluate

The results show 4 types of neighborhoods :

- **Family friendly**: For the people who wants a house and some shops arounds
- **Dynamic cities centers**: Best suits for people who like to live in apartments close to all commodities
- **Dynamic suburbs**: Suburban towns quite active with lots of shops and activities
- **Business centers**: These neighborhoods are business centers, not the best place to live in but perfect to install professional premises.

We decide to not consider price per square meter as an insight for clustering. We display this information on the map to help stakeholders taking a decision.

By including into the clustering algorithm information on real estate market, we have considered the lifestyle of the inhabitants and our clusters are clearly marked by the type of properties sold in the neighborhood. To better understand this part, it is quite important do dig further into the data. In fact, some neighborhoods have been classified as business centers by our algorithm since the amount of real estate's business transactions are more important that other transactions inside these neighborhoods for the last past five years. However, it does not consider the number of business per inhabitants that may be a better indicator of neighborhood constitution.

Finally, due to the limit number of results returned by foursquare, we decided to implement a higher level of cauterization directly into the parent categories. Some neighborhoods with only two venues returned may not be accurately clustered.

Discussion

Our project had several limitations :

The first one, is that we did not found a full and complete source to locate all neighborhoods. Even if the main towns are split in neighborhoods some minor towns would have bring a better analysis if split.

The second issue was the lack of data returned by Foursquare that has influenced the pertinence of the analysis, indeed, some of the neighborhoods had only one result, for this reason, the analysis is not always accurate.

Finally, by adding information on real estate market, we have brought important information on how evolve the local market but it does not take into account the real estate market in its globality, indeed, the information received comes only from the sales done during the past years but some shops, house, apartments that have not been sold are not in the data.

The price per square meters may evolve rapidly through time. By taking data from the past 5 years, we may have under estimate the price per square meters of some neighborhoods. In the other hand, by reducing the duration of collected data, we may not have obtained enough values to calculate a correct average price per square meter. For further analysis, it will be important to measure the accuracy of the real estate data by balancing duration and number of sales.

Furthermore, we decide to make a high level analysis that brings information on how is structured the metropole, of course it may be necessary to dig further in details in each neighborhoods to choose the perfect neighborhoods.

Conclusion

When looking at the choropleth map, we can rapidly see the important information we wanted to display : price per square meter, most sold type or property of the neighborhood and finally the "type" of neighborhood thanks to the venues returned by foursquare.

This works brings a lot of useful information of the structure of the real estate market but needs to be enrich with more accurate data on the local venues. We may decide to raise the radius of Foursquare explore API or add new neighborhoods coordinates.

We highly recommend improving the cluster algorithm by enriching it with all the neighborhoods of the Metropole and with more venues from Foursquare or another exploratory API (google for example).

This work can also be taken into a further step by adding information on population density that will ease the stakeholder's decision.

As we told before, by adding information on numbers of businesses installed and number of inhabitants of the neighborhood we may have more insight to determine the type of neighborhood.

Table des matières