IBM Data Science Professional Certificate

Where Should I Live ?

Capstone Project – The Battle of Neighborhoods

Jonathan DS
06/04/2020

I. Introduction

Introduction where you discuss the business problem and who would be interested in this project.

According to INSEE, in France, in 2013, about one household in five changed housing at least once between 2009 and 2013. The rate of residential mobility varies according to whether one owns or rents a dwelling, but also depends on the socio-demographic and professional characteristics of the household.

However, regardless of the characteristics of the household that is relocating, a common question remains: which neighbourhoods are the best place for them to start their new adventure?

Generally, a household does not have only one criterion on which it bases its choice of his future neighbourhood: it has several, and they are often complex, sometimes incompatible. The criterion looks like: "I wish to live in a quiet place, with green spaces, 45 min from my work, not too far from a tennis club, with shops, with schools. And above all that fits my budget..."

We will try to meet these wishes.

II. Data Acquisition and cleaning

Data where you describe the data that will be used to solve the problem and the source of the data.

a. Data Sources

For this project, we will use three sources of data

- Foursquare, which will give us access to all the facilities near the neighbourhoods. However, we will not extract the categories, but the families to which they belong. Thus, for instance, Casinos and Museums will be extracted as the same family “Arts & Entertainment”, while Asian Restaurant and Afghan Restaurant will belong to the “Food” family. One can find the Foorsquare categories tree there.

  Thus, we finally have nine families: parks & outdoors, food, nightlife spots, shops, arts & entertainment, building, travel, education , event.

- The second source of data is INSEE, the French National Institute of Statistics and Economic Studies (Institut National des Statistiques et des Etudes Economiques), which will provide to

us socio-demographic data corresponding to each neighbourhood, in particular the distribution of ages and professional categories.

The nomenclature of occupations and socio-professional categories classifies the population according to a synthesis of occupation, hierarchical position and status. Thus, for each district, we have the number of farmers, craftsmen, tradesmen, business managers, executives and higher intellectual professions, intermediate professions, employees, workers, and retired people. You can find the data there.

- Finally, the IGN, French National Institute of Geographic information (Institut National de l'Information Géographique et Forestière), provides a breakdown of the French territory into districts, with geometric contours. "Contours IRIS" is in the shapefile format.

  You can find the data there.

b. Data cleaning and feature selection

Our analysis will be restricted by two hypotheses:

- The first one is the geographic limitation. We will only consider four cities: Avignon, Arles, Aix-en-Provence and Marseille. These cities are in the South of France. Together they represent 508 neighbourhoods.
- We also simplify the complexity of the household choice, by assuming they are choosing their next neighbourhood based on three criteria: the age distribution, the professional distribution, and finally the facilities' categories distribution.

Insee provides detailed age groups: 0-2 years, 3-5 years, 6-10 years, 11-17 years, 18-24 years, 25-39 years, 40-54 years, 55-64 years, 65-79 years, over 80 years. We decided to aggregate these categories into four groups:

- Young (0-2 years, 3-5 years, 6-10 years, 11-17 years)
- Student (18-24 years)
- Active (25-39 years, 40-54 years, 55-64 years)
- Retired (65-79 years, over 80 years)

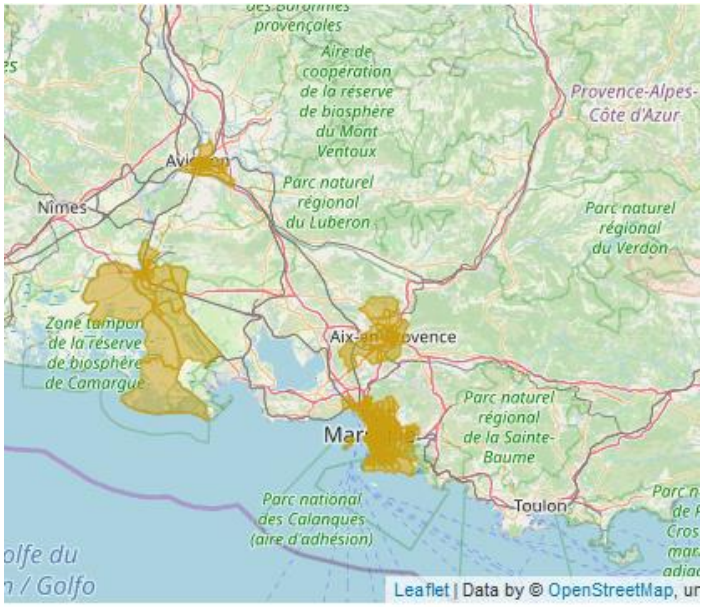


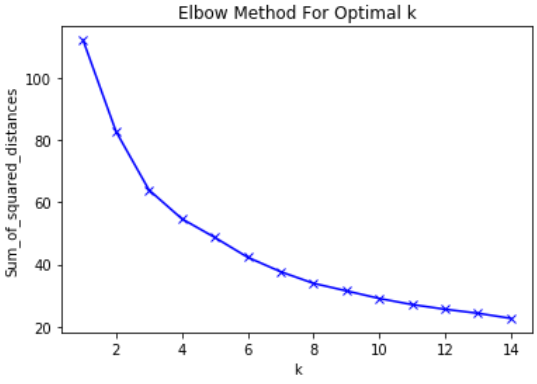
Figure 1 - Map of the four cities

III. Methodology

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

In order to best propose a neighbourhood to a household, we will build clusters for the three indicators. These clusters will form groups that will translate into choices for the household.

The clustering method used is the K-means, which suggested four clusters for each of these indicators using the "elbow method".



a. Exploratory data analysis

Below are two figures relating to the distribution of each age category over all neighbourhoods, as well as the four age clusters. These clusters clearly indicate that:

- the first (cluster 0) is made up of a significant proportion of students, which could be called a "student neighbourhood"; and

- the second (cluster 1) consists of a higher than average proportion of retirees
- the third (cluster 2) is made up mainly of assets, which could be called "business district"; and
- finally, the fourth (cluster 3) is made up of a significant proportion of young people and working people. This neighbourhood could be called "family district".

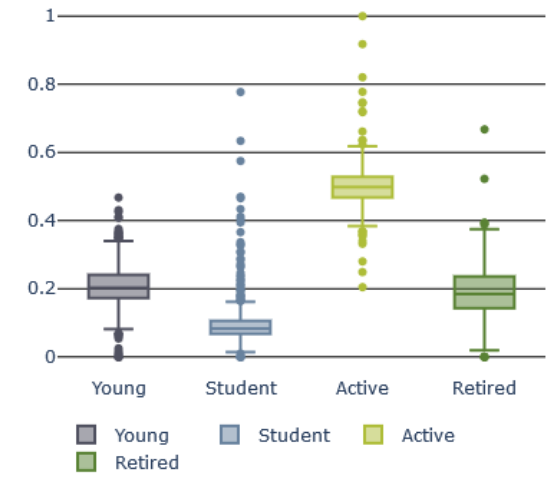


Figure 2 – Age categories distribution

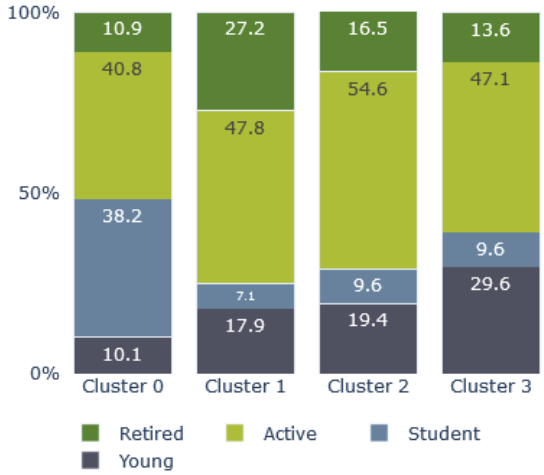
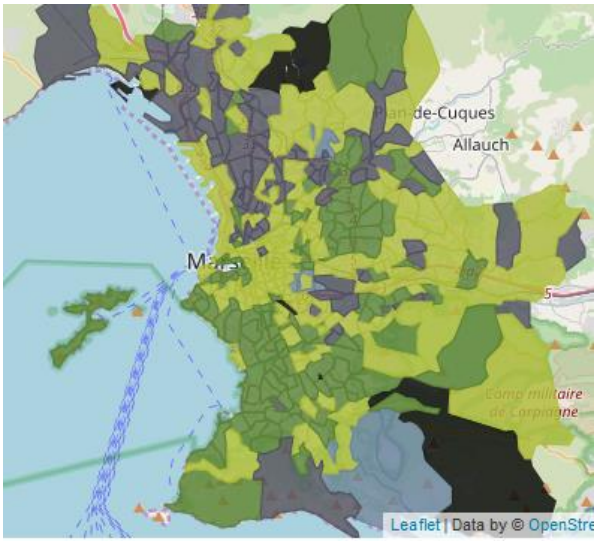


Figure 3 – Cluster age categories average

Here is a map of Marseille cut out by the four clusters.



The same analysis goes for the professional indicator. Below are two figures relating to the distribution of each professional category over all neighbourhoods, as well as the four professional clusters. These clusters clearly indicate that:

- the first (cluster 0) is made up of a significant proportion of executives, intellectual professions, as well as intermediate professions. These district could be related to the “top-level districts”.
- the second (cluster 1) consists of a higher than average proportion of retirees

- the third (cluster 2) is made up mainly of employees and workmen, which could be called "lower-class district"; and
- finally, the fourth (cluster 3) is made up of mixed professional categories, but lower proportion of executives. This neighbourhood could be called "medium-class district".

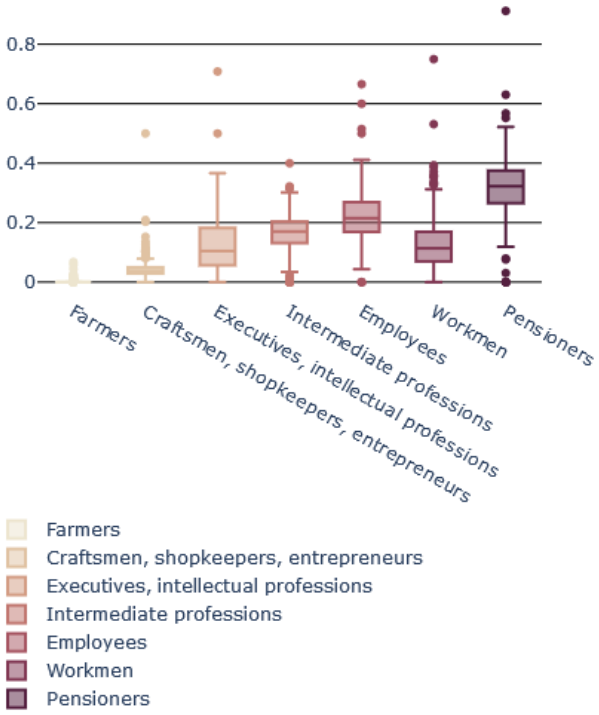


Figure 4 –Profession categories distribution

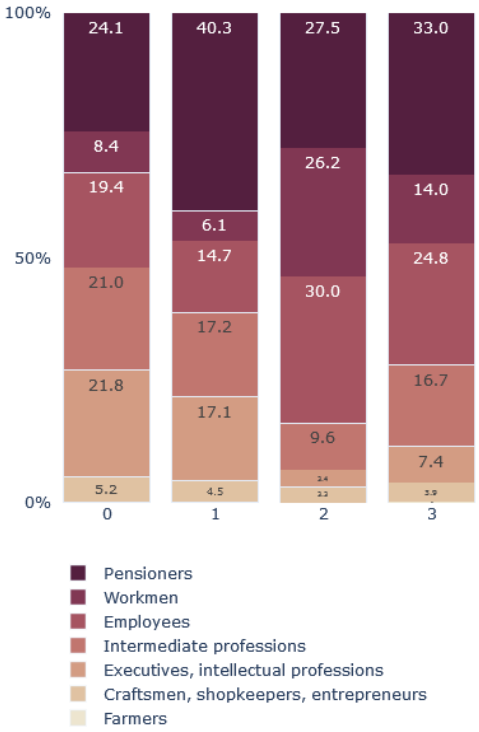
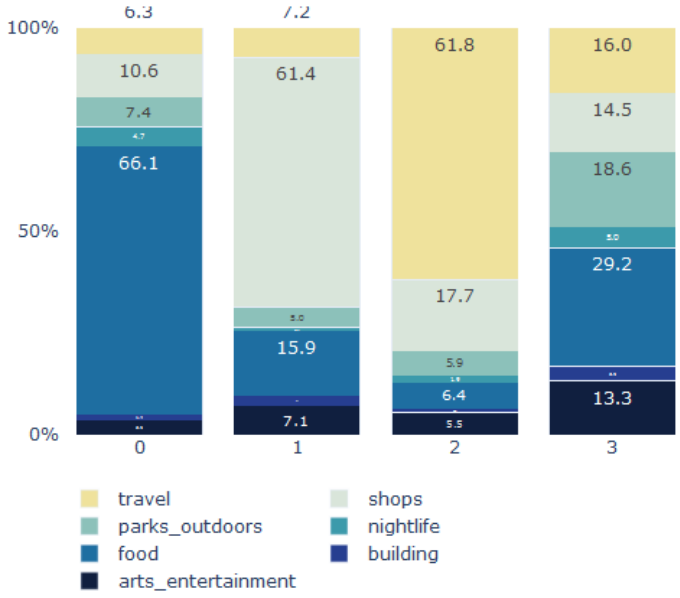


Figure 5 – Profession clusters

We finally cluster the facilities data from Foursquare. These clusters clearly indicate that:

- the first one (cluster 0) has a high proportion of restaurants and food related shops. We will refer to these districts as “restaurant districts”
- the second (cluster 1) consists of a higher than average proportion of shops: these are the “shopping district”
- the third (cluster 2) is made up mainly of travel facilities, like taxis stations or hotels. These are known as the "travel districts"; and
- finally, the fourth (cluster 3) is made up of mixed facilities categories. These neighbourhoods could be called "mixed districts"



IV. Discussion

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

- Clustering is based on proportions. These two examples would be in the same category

	Young	Student	Active	Retired	Total
X	100	50	200	100	450
X_{weight}	22.2%	11.1%	44.5%	22.2%	
Y	1000	500	2000	1000	4500
Y_{weight}	22.2%	11.1%	44.5%	22.2%	

- Choices are restricted to three indicators: the age distribution, the professional distribution and the facilities' categories distribution. However, households' choices are more complex.
- The age distribution could bias the analysis. Indeed, the presence of one retirement residence could sharply increase the proportion of retired people, mischievously transforming the neighbourhood to a "retirement district".

V. Conclusion

Conclusion section where you conclude the report.

Now that we have constructed clusters for each of our indicator, we know the similarities of each district of the cities according to each indicator. Hence, we can extract neighbourhoods that relate to shopping that also are "rich" quarters.

The following work should focus on taking into account the preferences of an household, build an index, and extract the neighbourhoods that are the most similar to its wishes. We could use for this the Euclidean distance, but also the Jaccard similarity.