Where Should I Live?

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

CAPSTONE PROJECT – THE BATTLE OF NEIGHBOURHOODS

Business Problem

- According to INSEE, in France, in 2013, about one household in five changed housing at least once between 2009 and 2013.
- A common question remains: which neighbourhoods are the best place for them to start their new adventure?
- A household has several criterions on which it bases its choice, and they are often complex, sometimes incompatible

Business Solution

- Provide households with a decision-making tool

- Guide households in their choice of their future neighbourhood

HOW

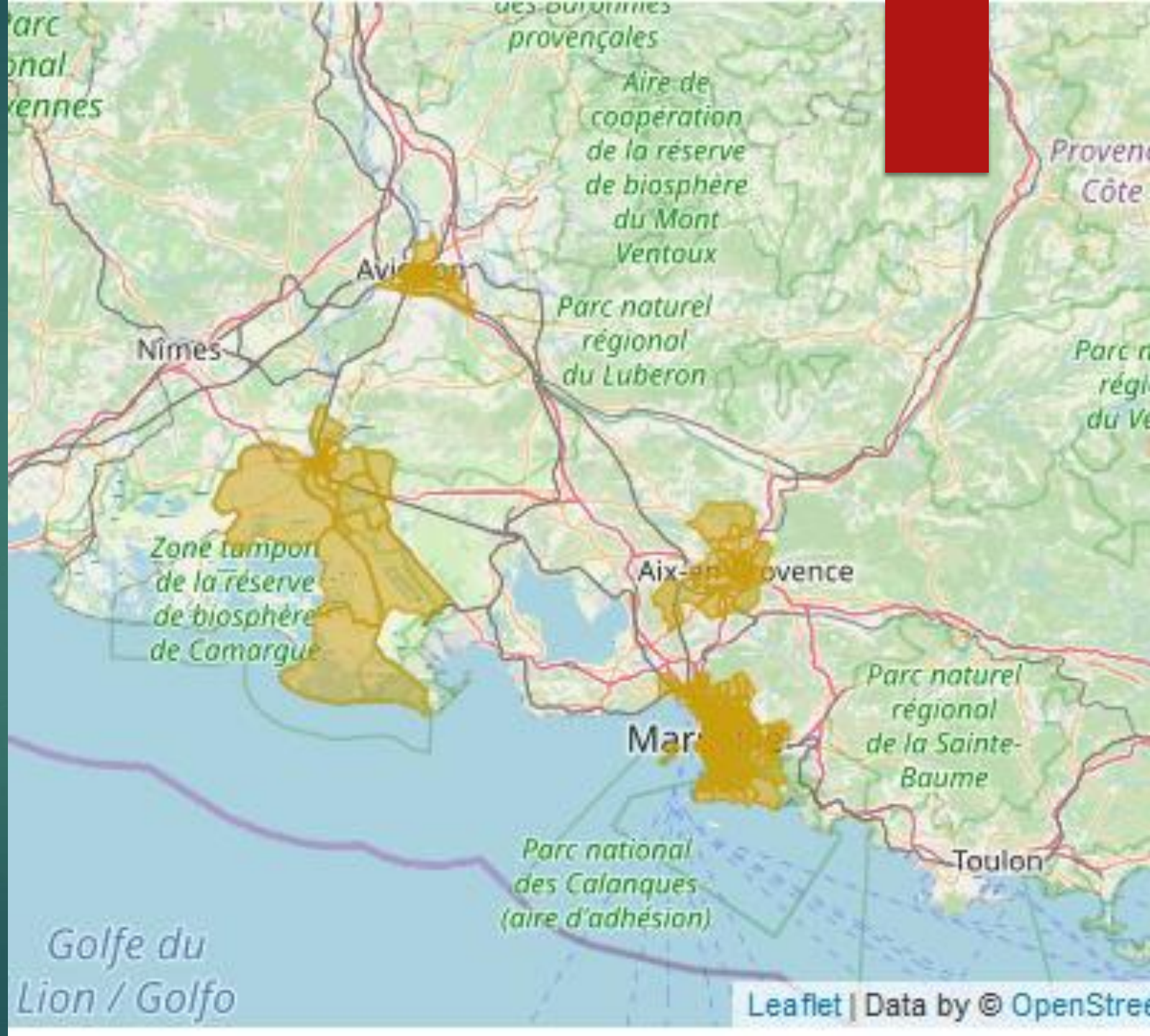
- Comparing and reconciling the characteristics of each city's neighbourhood

Data acquisition and cleaning

- Three sources of data
  - Foursquare, gives us access to all the facilities near the neighbourhood. We will just extract the families of the categories. One can find the Foorsquare families' tree there.
  - INSEE, the French National Institute of Statistics and Economic Studies which will provide to us distribution of ages and professional categories for each neighbourhood. You can find the data there.
  - IGN, French National Institute of Geographic information, which provides a breakdown of the French territory into districts, with geometric contours. You can find the data there.

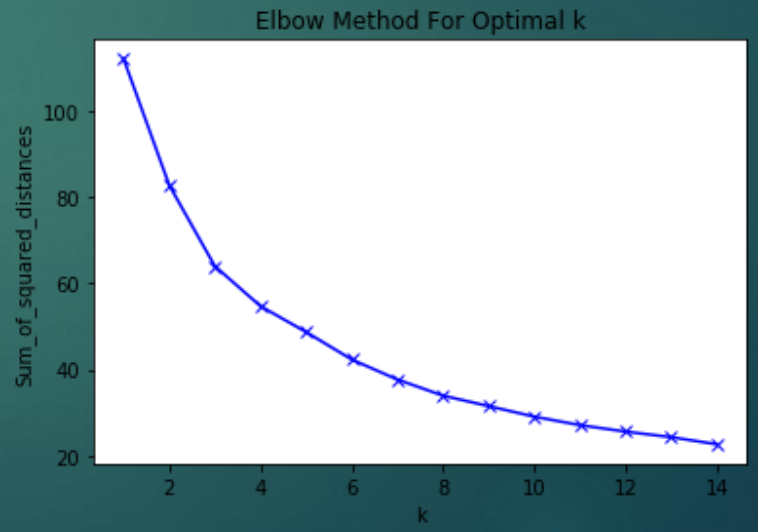
Geographic limitation

- We have chosen to limit our analysis to four cities of the South of France: Aix-en-Provence, Arles, Avignon, Marseille



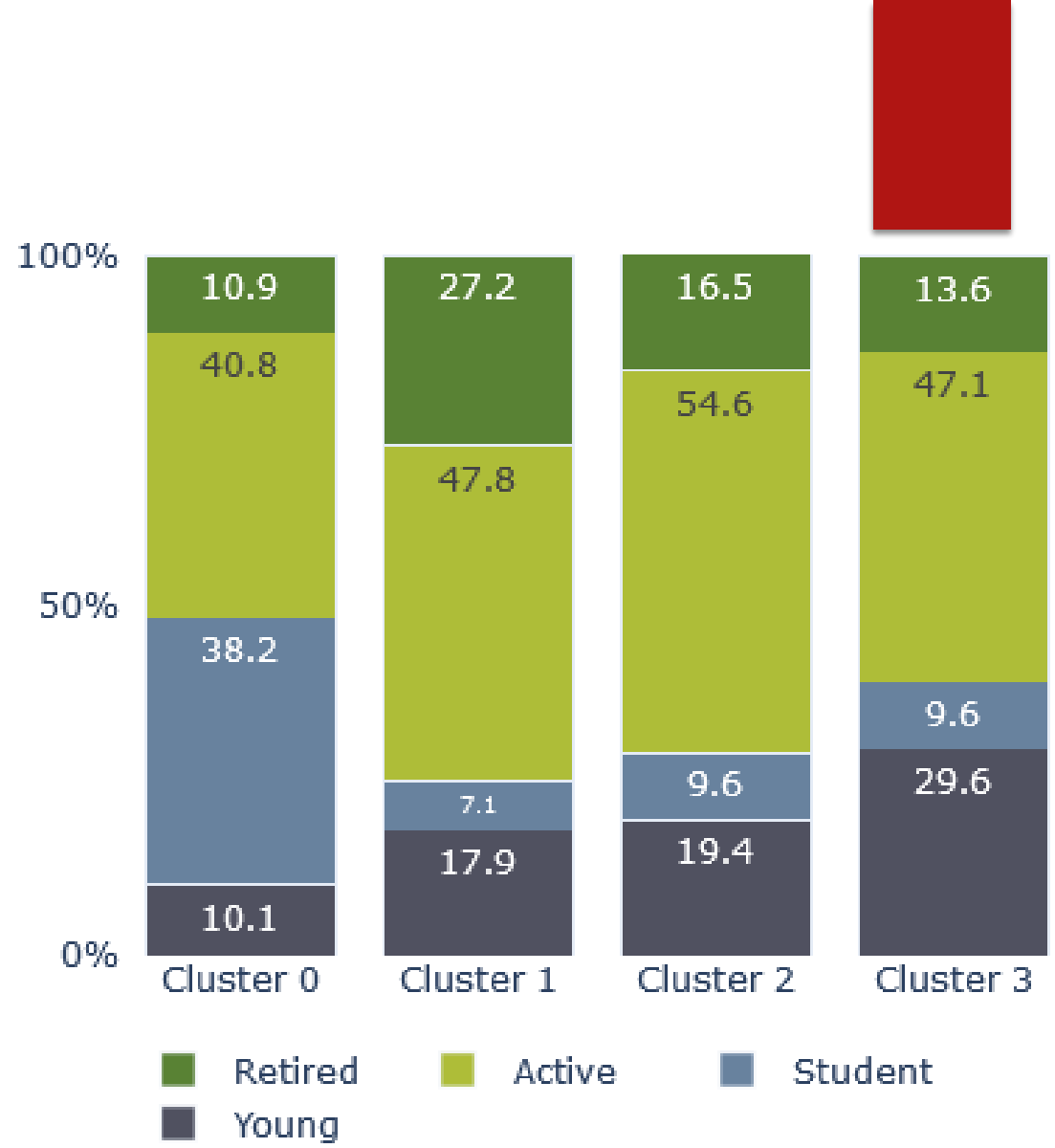
Methodology

- Clustering the indicators
  - Clustering algorithm
  - Optimization: Elbow method
  - Suggested 4 clusters for each indicator



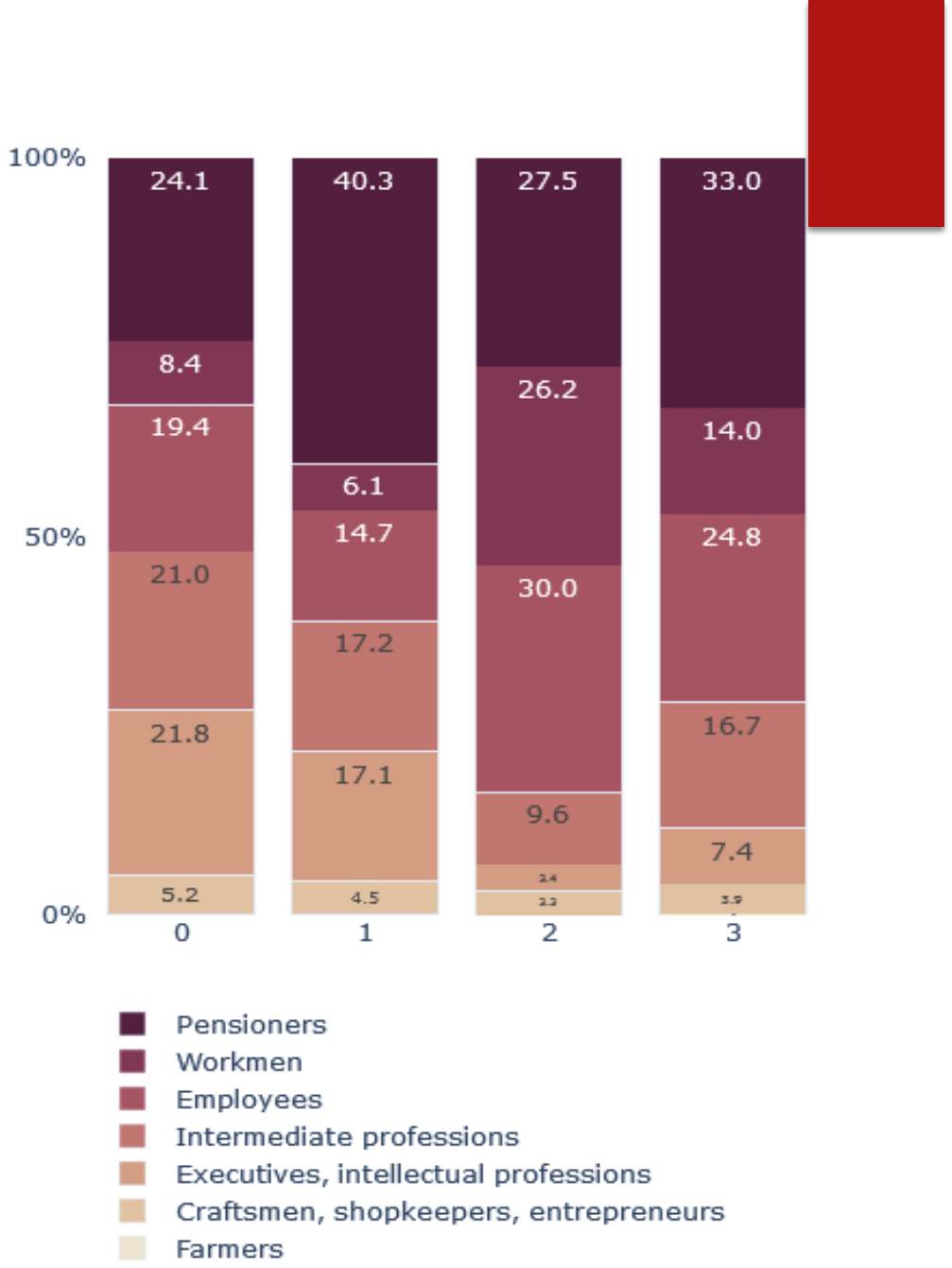
Clustering age groups

- 4 clusters
- Cluster 0: student districts
- Cluster 1: retirement districts
- Cluster 2: business districts
- Custer 3: family districts



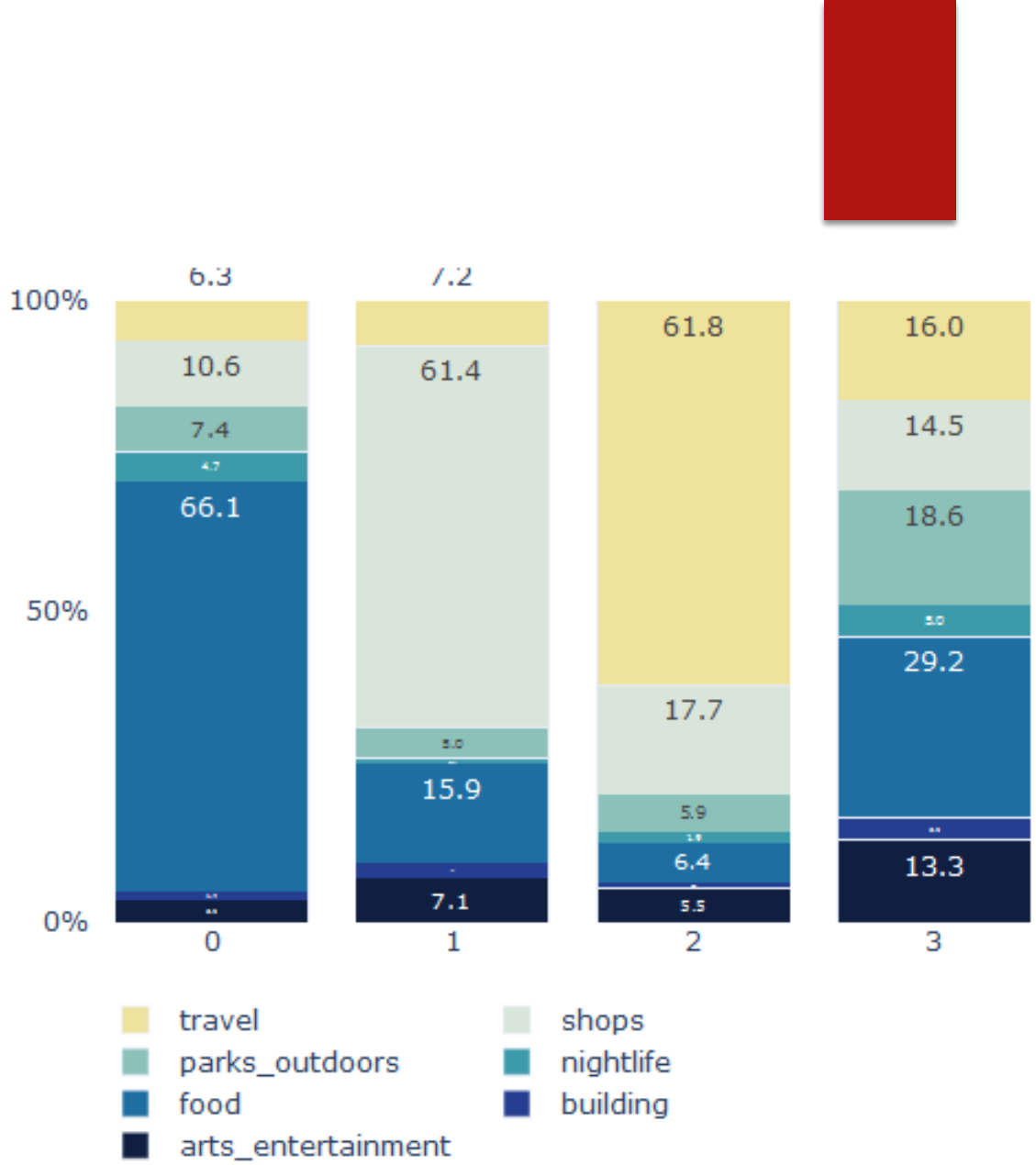
Clustering Professional groups

- 4 clusters
- Cluster 0: Upper-middle class districts
- Cluster 1: retirement districts
- Cluster 2: low-middle class districts
- Custer 3: mixed professional districts



Clustering Professional groups

- 4 clusters
- Cluster 0: restaurant district
- Cluster 1: shopping districts
- Cluster 2: travel districts (hotels, …)
- Custer 3: complete districts



Discussion (1)

- 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
$\mathbf{X_{weight}}$	22.2%	11.1%	44.5%	22.2%	
Y	1000	500	2000	1000	4500
$\mathbf{Y_{weight}}$	22.2%	11.1%	44.5%	22.2%	

Discussion (2)

- Choices are limited 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".
- This method can lead to an incomplete picture

Results & Conclusions

- To each neighbourhood is assigned a cluster based on prefessional activity, age distribution, and facilities presence.

- Next step:
  - Consolidate the three indicators
  - Get a houselhod wish based on our criteria
  - Extract most similar neighbourhoods, with the Jaccard similarity