Recommender System for the City of Rennes Neighborhood's Clusters.

MAHE Pierre-Antoine

1 Introduction

Choosing a new place to live can be a daunting task, especially when arriving in a new city. Even for a real-estate agent with previous knowledge of the city, meeting the client expectations can be challenging. Even more so as today's cities are bigger than ever and always changing.

This work objective is to help real-estate agent in that task and we will focus on th city of Rennes, Brittany, France.

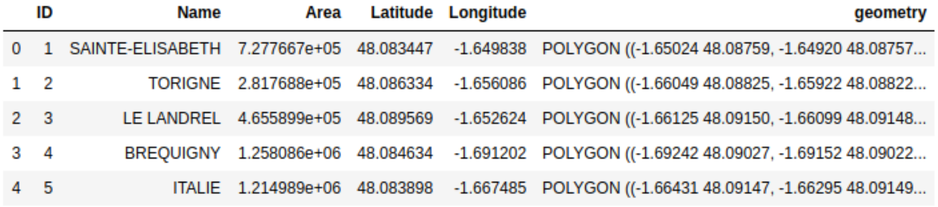
We will first classify Rennes neighborhoods based on their venues returned by Foursquare and then, build a Content-Based Recommender System to find the most appropriate living neighborhood for defined 'user-profiles'.

2 Data

Neighborhoods : We will be using two sources of data. First, the neighborhoods location and geometry extracted from the French Administration Open Data Website (https://www.opendata.gouv.fr) and published under the Open Database license v1.0.

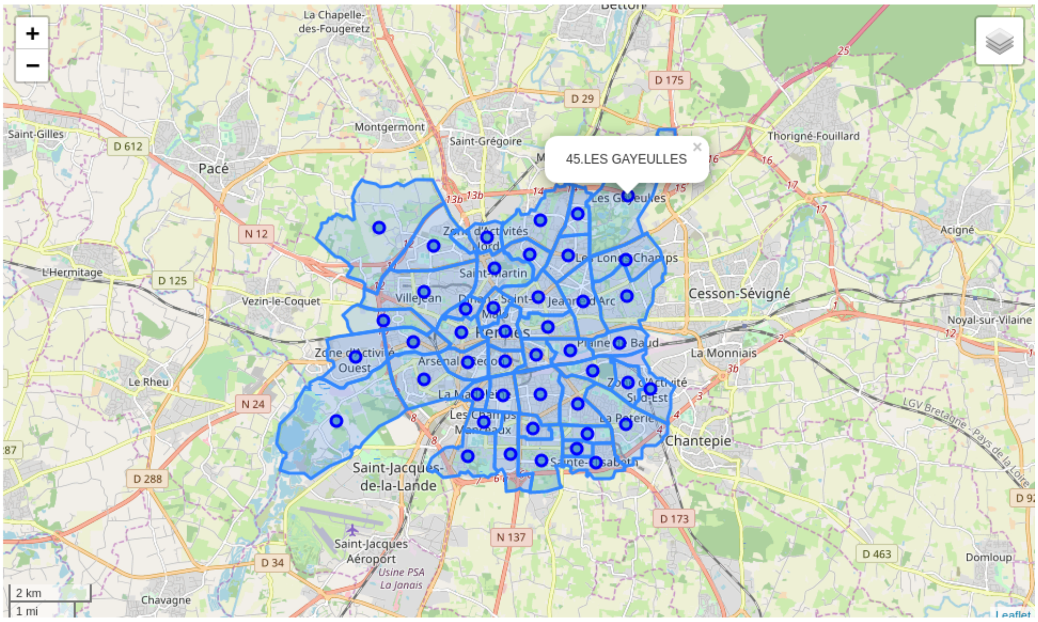
The dataset we used is named 'perimetre-des-45-sous-quartiers-de-la-ville-de-rennes' and extracted as a geojson file and contains all the data we need to identify and locate each neighborhood.

With a little wrangling, we can get the following geodataframe:



	ID	Name	Area	Latitude	Longitude	geometry
0	1	SAINTE-ELISABETH	7.277667e+05	48.083447	-1.649838	POLYGON ((-1.65024 48.08759, -1.64920 48.08757...
1	2	TORIGNE	2.817688e+05	48.086334	-1.656086	POLYGON ((-1.66049 48.08825, -1.65922 48.08822...
2	3	LE LANDREL	4.655899e+05	48.089569	-1.652624	POLYGON ((-1.66125 48.09150, -1.66099 48.09148...
3	4	BREQUIGNY	1.258086e+06	48.084634	-1.691202	POLYGON ((-1.69242 48.09027, -1.69152 48.09022...
4	5	ITALIE	1.214989e+06	48.083898	-1.667485	POLYGON ((-1.66431 48.09147, -1.66295 48.09149...

Which can easily be plotted using Folium



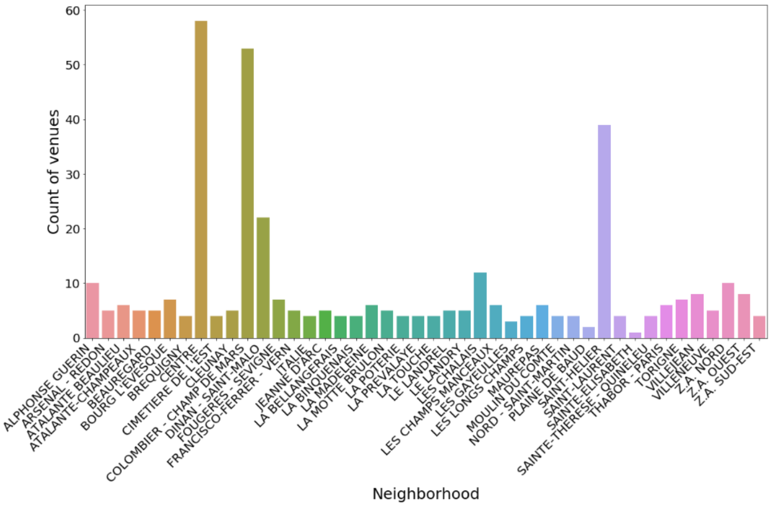
Venues: We will be using the Foursquare API to return the list of venues in each neighborhood, but we will onl take into account their top-level category defined by Foursquare. There is 10 top-level categories:

- Arts & Entertainment
- College & University
- Event
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional & Other Places
- Residence
- Shop & Service
- Travel & Transport

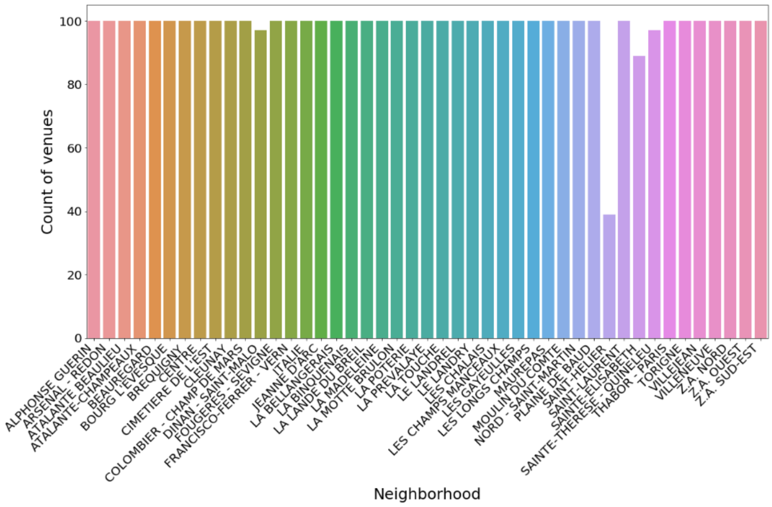
3 Methodology

3.1 Foursquare Data

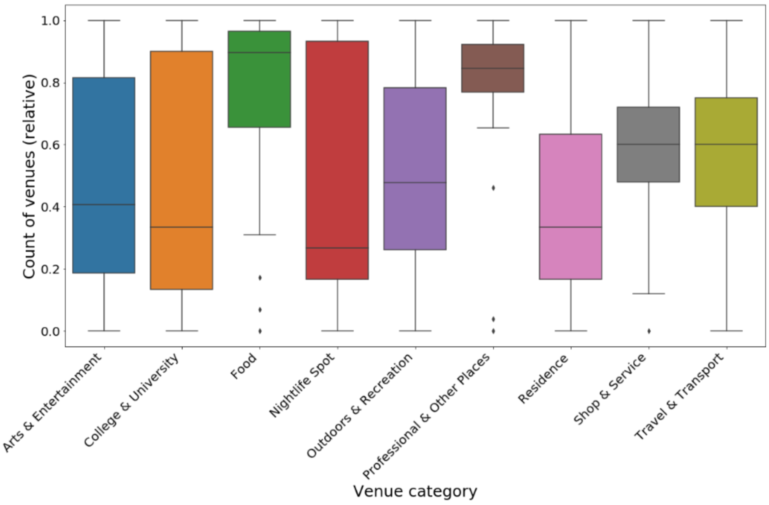
Explore Endpoint: We first use the explore endpoints to get data from Foursquare but since it is limited to recommended venues, it was very limited with less than 10 venues for most neighborhoods despite a limit set to 100.



Search Endpoint: The Search Endpoint returns any venues in the specified radius and limit parameter. The number of returned venues is then much higher with close to 100 venues per neighborhood. However, returned venue can be pretty much anything (including hospitals, bus stops ...) which makes it really hard to analyze.

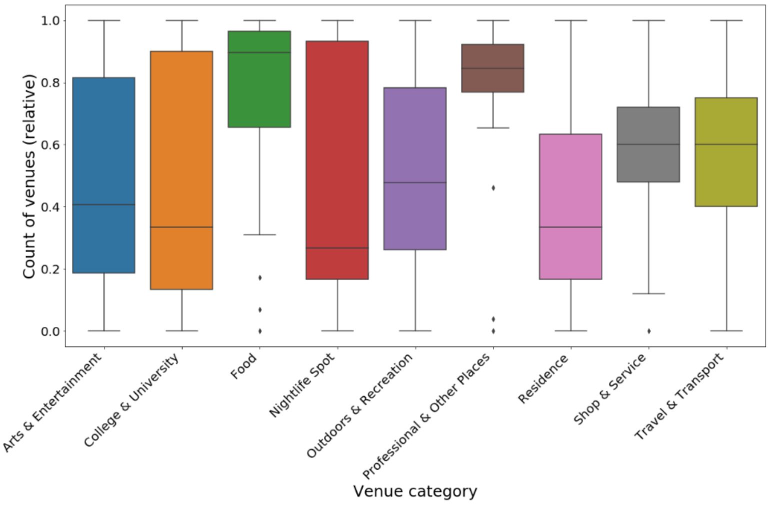


Category and Search Endpoint: The solution was to use the Category Endpoint to extract the 10 top-level categories defined by Foursquare, then calling the Search Endpoint and classify each venue in one of those categories.



3.2 Exploratory Analysis and Data Wrangling

The 'Event' category is almost non-existent and therefore will be dropped. We also used a MinMaxScaler to normalize the number of venues in the neighborhoods. Making it easier to compare them.

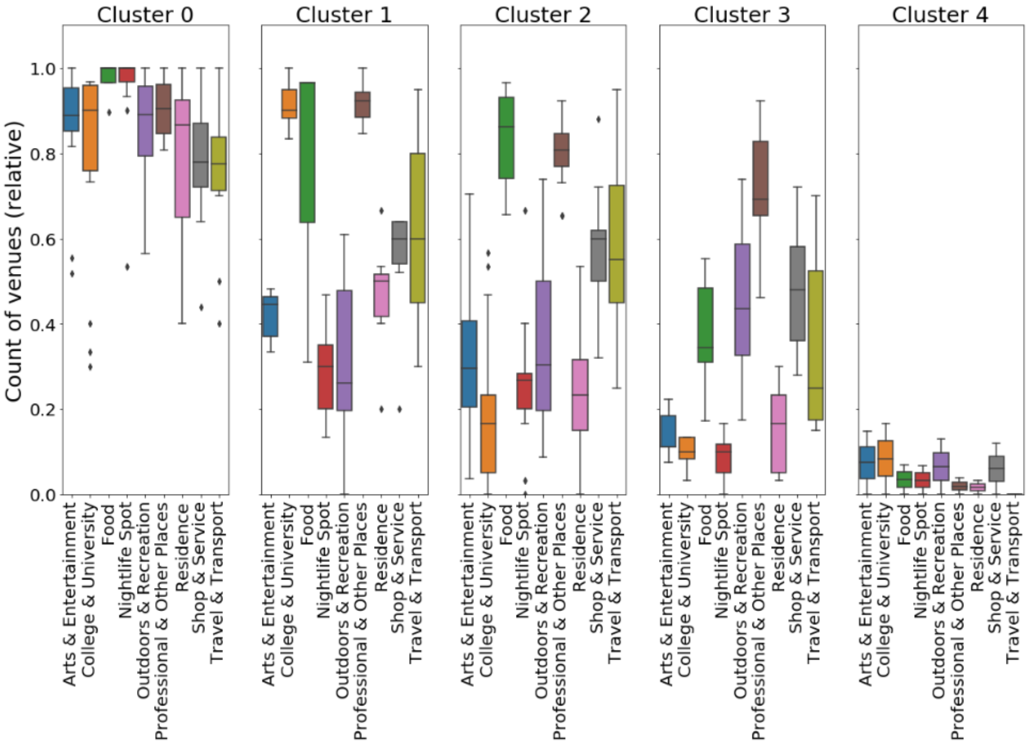


3.3 Clustering

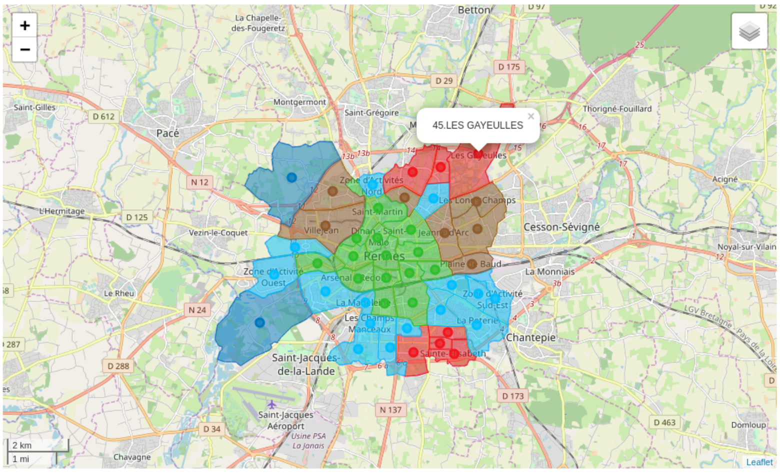
Number of Clusters: We used the k-means clustering algorithm and tried to get result with different number of clusters :

- 2 clusters shows only the separation between the city center and less-dense areas
- 3 clusters add a new separation in suburban areas but does not split the city center cluster
- 4 clusters is better but there is still an important variation within some clusters
- 5 clusters is where we stopped
- 6 clusters start to make it difficult to analyze results

For the final analysis with 5 clusters, the boxplot visualization is the following :

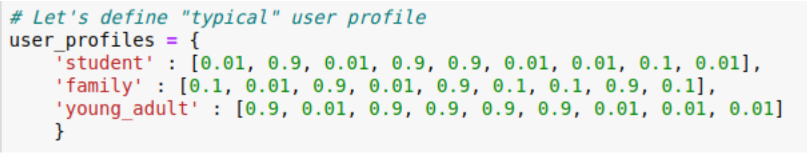


And we can see the cluster separation on the map of Rennes :



3.4 Recommender System

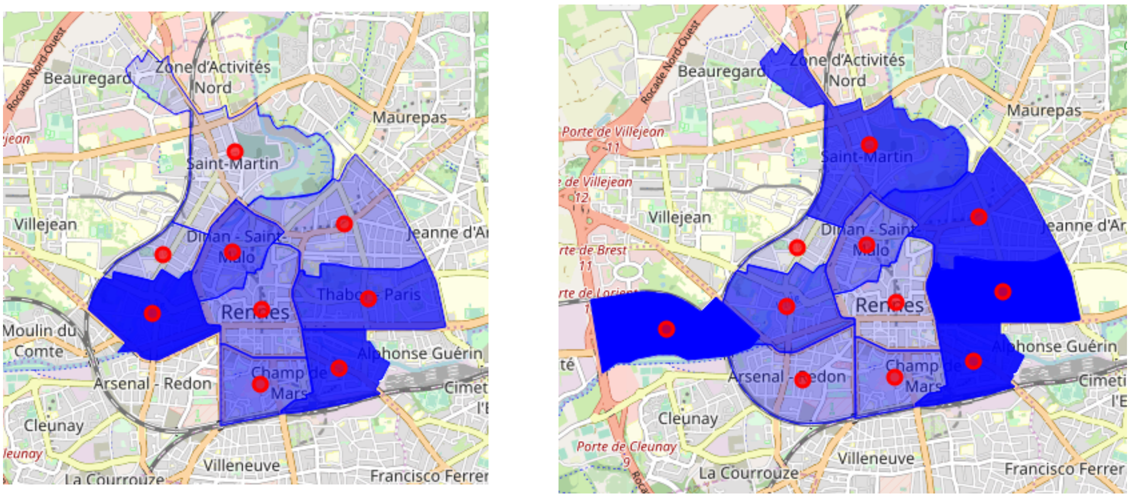
With the MinMaxScaler work on venues categories for each neighborhood, we already have what is needed to build a Content-Based Recommender System. We only need to define 'typical' user-profiles :


```
# Let's define "typical" user profile
user_profiles = {
    'student' : [0.01, 0.9, 0.01, 0.9, 0.9, 0.01, 0.01, 0.1, 0.01],
    'family' : [0.1, 0.01, 0.9, 0.01, 0.9, 0.1, 0.1, 0.9, 0.1],
    'young_adult' : [0.9, 0.01, 0.9, 0.9, 0.9, 0.9, 0.01, 0.01, 0.01]
    }
```

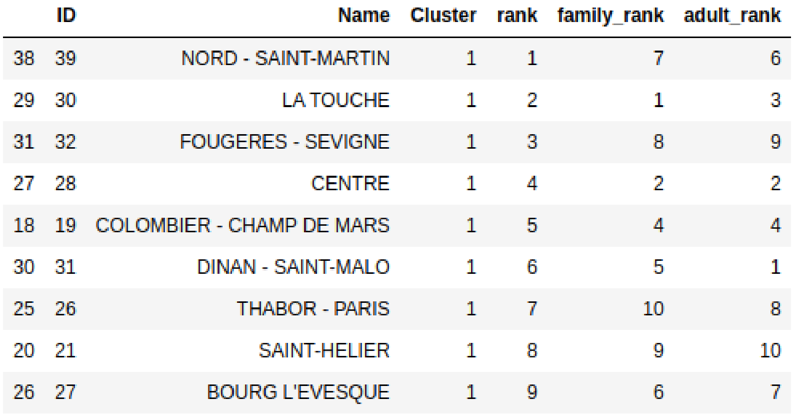
We can then multiply those profiles with the MinMaxed dataframe of neighbhorhood to get their rating for each user.

This allow us to build maps based on the ratings. The darker a neighborhoods, the more it is recommended for a user. To make the differences between users more visible, we will only plot the neighborhoods with a rating superior to 0.8 (80%). And even then, the difference are not always obvious.

Below we can see the map for the student profile (left) and the family profile (right). As mentioned previously, all recommended neighborhoods are within the city center but there are a few differences.



The differences between profiles are more visible with the direct number (NB: the rank column is actually the student_rank ranking):



	ID	Name	Cluster	rank	family_rank	adult_rank
38	39	NORD - SAINT-MARTIN	1	1	7	6
29	30	LA TOUCHE	1	2	1	3
31	32	FOUGERES - SEVIGNE	1	3	8	9
27	28	CENTRE	1	4	2	2
18	19	COLOMBIER - CHAMP DE MARS	1	5	4	4
30	31	DINAN - SAINT-MALO	1	6	5	1
25	26	THABOR - PARIS	1	7	10	8
20	21	SAINT-HELIER	1	8	9	10
26	27	BOURG L'EVESQUE	1	9	6	7

Where we see easily that even if the recommended neighborhoods are the same, they are not ranked in the same way.

4 Results

As shown in the table, as well as in previous maps, the recommendation system is working but is always recommending the city center neighborhoods.

This is mostly due to the concentration of venues in those neighborhoods. As shown in the Clusters categories barchart, Cluster 0 contains a lot of venues of each categories, i.e. is a good choice for every profile.

But within the city center, not all neighborhoods are equivalent, and their ranking can be significantly different from one user to another. This allow us to say that the model is working.

5 Discussion

The clustering and recommendation are performing as expected and provides valuable insights to solve the business problem.

The downside of our analysis is the limitation of Foursquare data for different reasons:

- Data is biased towards the Food and Recreational categories
- The Explore Endpoint is too limited (low number of returned values)
- The Search Endpoint is too wide (everything is returned, no matter how good or relevant)

The best way to improve our analysis will be to gain access to other datasets, mainly:

- real-estate past transaction for the city to extract data not available on Foursquare (m^2 price, available private parking spots, ...)
- venues data including rating (premium endpoint in Foursquare) to clean the output of the Search Endpoint

6 Conclusion

Foursquare data, although limited, helped us prove the feasibility of a living place recommender system. With more precise and relevant data, this model could be vastly improved.