CLUSTERING THE CITIES OF GRAND PARIS

Coursera Capstone Project

Frédéric Benoit – September 2021

BACKGROUND

- With the covid-19 crisis, many people want to leave Paris and its high housing prices for the suburbs (« Grand Paris ») where the house surfaces are greater.

- The target of this study is to cluster the cities of Grand Paris to select the most suited for the needs and centers of interests of peopl willing to relocate out of Paris.

- The results of the study are aimed at individuals willing to move out of Paris, to real estate agents or investors.

DATA

- List of cities of Grand Paris available as open data in Excel format:

https://www.data.gouv.fr/fr/datasets/communes-de-la-metropole-du-grand-paris-par-ept/

- Geospatial data of cities from package geopy/geocoder.

- List of venues available in each city with the use of the Foursquare API.

METHODOLOGY (1/4)

- Data Collection and Cleaning:
  - The list of cities of Grand Paris is downloaded as open data from: https://www.data.gouv.fr/fr/datasets/communes-de-la-metropole-du-grand-paris-par-ept/
  - For each city, the geospatial data of is obtained from the package geopy/geocoder.
  - Based on the geospatial data of each city, the list of venues available is obtained with the use of the Foursquare API.

METHODOLOGY (2/4)

- Group the venues in super-categories:
  - A function is used to group the venues in super-categories. Ex: all restaurants, food markets are grouped in a super-category « Food ».
  - There are 8 super-categories:
    - Art & Entertainment
    - Food
    - Nightlife
    - Residence
    - Shop & Service
    - Travel and Transport
    - Outdoors & Recreation
    - Professional & Other Places

METHODOLOGY (3/4)

- Clustering
  - I cluster the cities based on the 8 super-categories defined.
  - k-means method with 4 clusters.

METHODOLOGY (4/4)

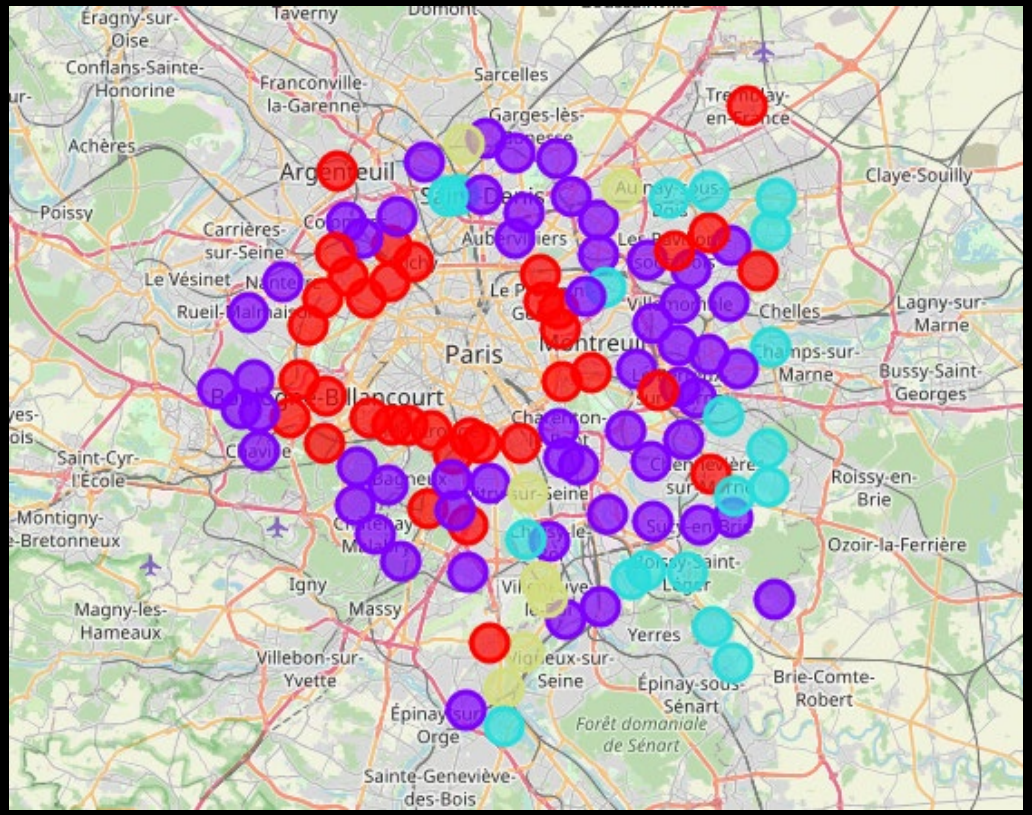
- The characteristics of the 4 clusters are the following:

Cluster Labels	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	0.055275	0.538116	0.032077	0.112140	0.005029	0.000393	0.156135	0.100835
1	0.048000	0.336544	0.014140	0.174600	0.002478	0.000565	0.272524	0.151149
2	0.026816	0.220651	0.002156	0.101520	0.016951	0.000000	0.537240	0.094666
3	0.021296	0.186731	0.000000	0.027437	0.008547	0.000000	0.274882	0.481107

- This leads us to define target-populations for each cluster:
  - Cluster 0 -> higher density for food, nightlife, Arts & Entertainment and is 2nd for the professional category -> recommended for students, young workers.
  - Cluster 1 -> more residence and outdoor and recreational areas ->recommended for practitioners of outdoor activities and possibly retired people.
  - Cluster 2 -> higher density of professional places, shops and services -> recommended to install companies, businesses.
  - Cluster 3 -> best for Travel and Transport -> recommended for frequent travelers.

RESULTS

- Map of the different clusters:



- Cluster 0
- Cluster 1
- Cluster 2
- Cluster 3

CONCLUSIONS

- This study shows that it is possible to cluster the cities to find the most attractive areas depending on the peoples' needs and centers of interest.

- Limitations:

  - Venues data not comprehensive enough (for example not showing schools in this area).

  - Clustering based on relative distribution of venues can be misleading, a clustering based on venues density would give a more relevant result.

FUTURE DIRECTIONS

- This study could be the starting point for an application in which a customer enters its needs and centers of interest and obtains the best groups of cities to live in.

- For this, it would be necessary to:

  - Make a convenient user interface.

  - Enlarge the number of input criteria: increase the venues dataset (ex: with schools) and add external data (ex: housing prices).

  - Cluster based venues density and not venues distribution.