

Data

Capstone Project

Science

Introduction.

Case study

High-tech company
Set up a headquarter shift
In Paris from Petit-Montrouge neighborhood

Since work done from office
→ Mainly consider workers interest

Same characteristics neighborhood based on :

Venues in the area
Population attributes

Project description

Project : finding similarities in Paris neighborhood to inform the company decision to acquire new headquarters

Idea : clustering twice
On venues
On population attributes

Data required :

Paris neighborhoods list and geographical information
Paris neighborhoods most frequent venues
Paris neighborhoods population data

Data.



geographical information

As geographical_df

Source

[open data paris - geographical information](#)

Attributes

Id_seq (link to other data)
Code INSEE (idem)
Borough
Neighborhood
Latitude
Longitude



Venues data

As paris_venues

Source

Function getNearbyVenues from
foursquare API
[Foursquare - developers apps](#)

Attributes

Neighborhood
1st Most Common Venue
...
nth Most Common Venue



Population attributes

As population_df

Source

[open data aput - recensement](#)

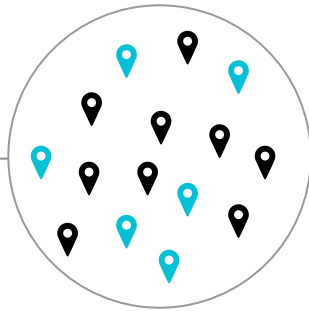
Attributes

Code INSEE commune
Density
Population
Population under 40

Idea diagram.

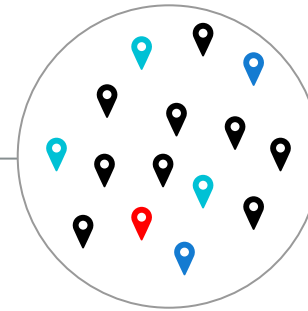
First clustering

Discriminate neighborhoods based on most common venues



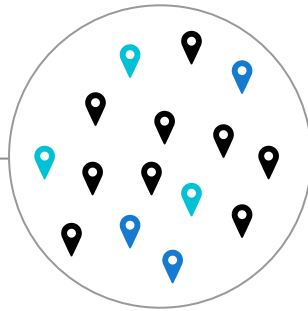
Final decision

Choose neighborhood based on current possibilities (and again worker preferences if possible)



Second clustering

Discriminate neighborhoods based on population information



Methodology.

1

Geographical Dataframe

Importing data

Selecting meaningful attributes

Preparing geographical data (cf to numeric)

Removing abnormal row

2

Finding venues from Foursquare

Import necessary libraries
Import configured function

Set parameters (radius and limit) to get the best from Foursquare

Check result

3

First clustering

Pre-process data (cf onehot encoding)

Process clustering

Set parameters (kclusters) to partition Paris neighborhoods rather equally

4

Geographical display

Find Paris location

Create map

Add neighborhoods and denote clusters using colours

Check result (clusters repartition)

5

Second clustering

Importing data

Selecting meaningful attributes

Uniformize attributes

Process clustering

Extract project's analytic result

Conclusion.

Neighborhood

Petit-Montrouge

Plaisance

Necker

Grenelle

Plaine de Monceaux

Saint-Georges

Chaussée-d'Antin

Result

Find the attached the list of neighborhoods corresponding to 'Petit-Montrouge' in terms of common venues in the area and of population related information.

Discussion

The first clustering is based on most common venues, if on particular venue is in fact especially important, we missed this particular information.

The second clustering take into consideration population under 40, it is an arbitrary choice.

Important to note, we processed both clustering multiple time to make sure that data retrieved offers the company the final choice.

Conclusion

Company now have to chose one of the highlighted neighborhoods.

Data science can inform the choice by giving new insights, but can't take the decision.
This project teach us to stay humble.