CHOOSING VENUE CATEGORY WITH NEIGHBORHOOD CLUSTERING

Applied Data Science Capstone Final Assignment

André Costa Nogueira dos Santos

20/06/2019

Table of Contents

Introduction

Being an entrepreneur and starting your own business is a goal shared by many people. However in order to be successful hard work is often times not enough. Success is also dependent on other factors like luck and perhaps most important of all: finding the right opportunity.

When talking about opening a public venue the right opportunity is defined by the rules of supply and demand in a given market. There are a couple of ways to approach this problem:

1. Start by choosing a market location, analyze its supply and demand and then decide which type of venue is needed;
2. Start by selecting a business idea and afterwards search for a market location where demand for that type of business is higher than supply;

In this report I am working with a case of the latter.

Problem

My friend Louis has just inherited a building in *Popincourt*, a neighborhood of Paris, and he wants to take advantage of this opportunity to start his own business. He knows he wants to open a nightlife type of venue but he wants to know specifically which type of venue he should open. Venue Data can be used to cluster neighborhoods and discover which type of venues thrive in that cluster. Also by comparing the neighborhood in question with others from same cluster it is possible to find out which type of venues are in surplus and which are lacking.

Data acquisition and cleaning

Data Sources

Information about neighborhoods of Paris (or *Arrondissements* as they are called there) can be found at Wikipedia. Information found there is very complete, having not only names but also Area, Population and Density. To get GPS coordinates for each neighborhood I used GEOpy python library. Furthermore all coordinates were checked and confirmed using Latitude's website.

To retrieve venue information for each neighborhood I used Foursquare API.

Data Cleaning

Data scraped from Wikipedia contained some information that was not necessary for this analysis and was removed.

First, columns for population, date of peak population and Mayor's name were eliminated as they are not relevant for neighborhood clustering and/or can be calculated using kept data.

Second, area values were given in both square kilometers and square miles. Since we will use radius value in meters in Foursquare, data in miles was excluded. Also km^2 identification was dropped and values were converted from string to numeric.

Finally for neighborhoods *Reuilly* and *Passy* there were two values for both Area and Density. This is due to the fact that Paris biggest parks belong to these neighborhoods which considerably augment their area and consequently reduce Density since there is no population living in the parks. These parks although falling under jurisdiction of those neighborhoods, are located on the periphery and are not relevant for this analysis. Only values without considering both parks were kept.

After cleaning the data, a column was added at the end of the table to insert radius values to be used for foursquare calls. As a simplification it was considered that each neighborhood was a circle and a radius in meters was obtained based on its area.

$$Radius = \sqrt{\frac{Area}{\pi}}$$

Unfortunately venue category information in Foursquare is not 100% accurate. There are 3 situations that required cleaning in order to get a usable list of venue categories:

1. Venues which are incorrectly placed as Nightlife Venues: although Foursquare has a specific category for Restaurants and Hotels it also allows those types of venues to be assigned as nightlife venues. This is probably due to the fact that those can be open all night, however for the purpose of this analysis they are not of interest
2. Venue category which are vague: although Foursquare has very specific venue categories like "Karaoke Bar" for instance, some places only have a very vague category like "Bar".
3. Similar venue categories with different names: there are some categories with different names but which concept is similar. For instance "Beer Garden" and "Beer Bar", although there might be some differences between them (one has a garden while the other might not) for the purpose of this study they can be considered the same – venues targeting beer lovers.

Table 1 shows which categories were dropped, which were converted and which were kept.

Categories	Action	New Category
BBQ Joint; Breakfast Spot; Office; Steakhouse; Hotel; Boat or Ferry; Pizza Place; Bar;	Dropped	-
Beer Garden; Brewery; Beer Store;	Converted	Beer Bar
Wine Shop	Converted	Wine Bar
Gastropub	Converted	Pub
Wine Bar; Pub; Beer Bar; Lounge; Cocktail Bar; Hotel Bar; Music Venue; Bistro; Nightclub; Concert Hall; Karaoke Bar; Dive Bar; Sports Bar; Piano Bar; Rock Club; Smoke Shop; Roof Deck; Speakeasy; Hookah Bar; Gay Bar; Jazz Club; Champagne Bar; Beach Bar; Whisky Bar;	Kept	-

Table 1 - Category selection during data cleaning;

Unfortunately there were several results that were not usable. Initially I had 1559 samples (~78 venues per neighborhood) and after cleaning we were left with only 780 samples (~39 venues per neighborhood). Of the 779 samples that were dropped, 72% had vague category like "Bar" and only 28% were incorrect venues.

Cluster Analysis

Cluster analysis is a task that consist on dividing samples into groups (clusters) in a way that samples from same cluster are more similar to each other in comparison with samples from other groups.

In this project I wanted to separate Paris's neighborhoods, our samples, into different clusters based on the type of nightlife venues which are more common in those neighborhoods. I used **k-means clustering** which is a method of vector quantization.

Selecting number of clusters – Elbow method

Before performing clustering analysis using k-means it is necessary to define the optimal number of clusters. One method for doing so is by plotting the elbow curve (figure 1). In this method k-means clustering is run on the dataset for a range of values of k and for each k it returns the sum of squared errors (SSE). When the SSE is plotted against k in a line chart it looks like a flexed arm, hence the name "elbow curve". The elbow region represents when gradient rapidly diminishes which means that increasing number of cluster becomes less beneficial after the elbow.

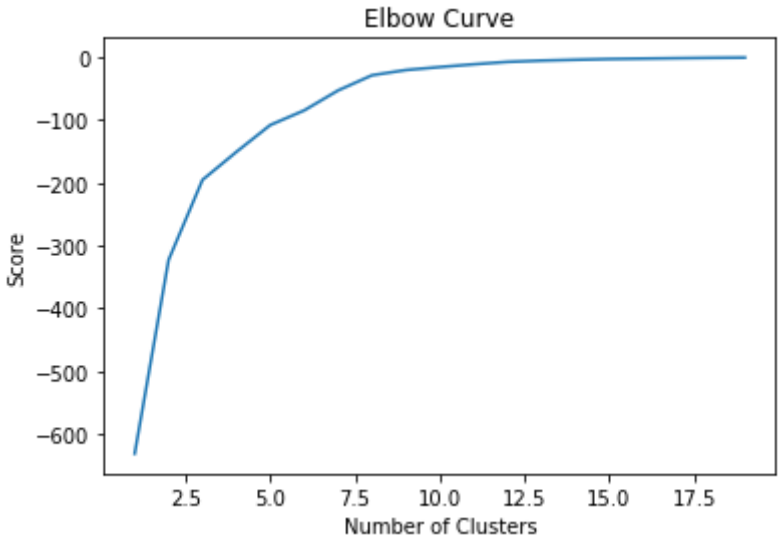

Figure 1 - Elbow curve for k-means clustering analysis;

In our case I used range from 1 to 20 which is the complete range of possible ks since I had 20 neighborhoods total. In figure 1 I plotted SSE against k and observed that the elbow ends around a value of k=8 and after that we have diminishing returns by increasing k. So I consider optimal cluster value to be 8.

K-Means Analysis

After determining the optimal number for k I ran k-means clustering on neighborhood data. In figure 2 I display a map of Paris with a circle on each neighborhood and a different color representing each cluster.

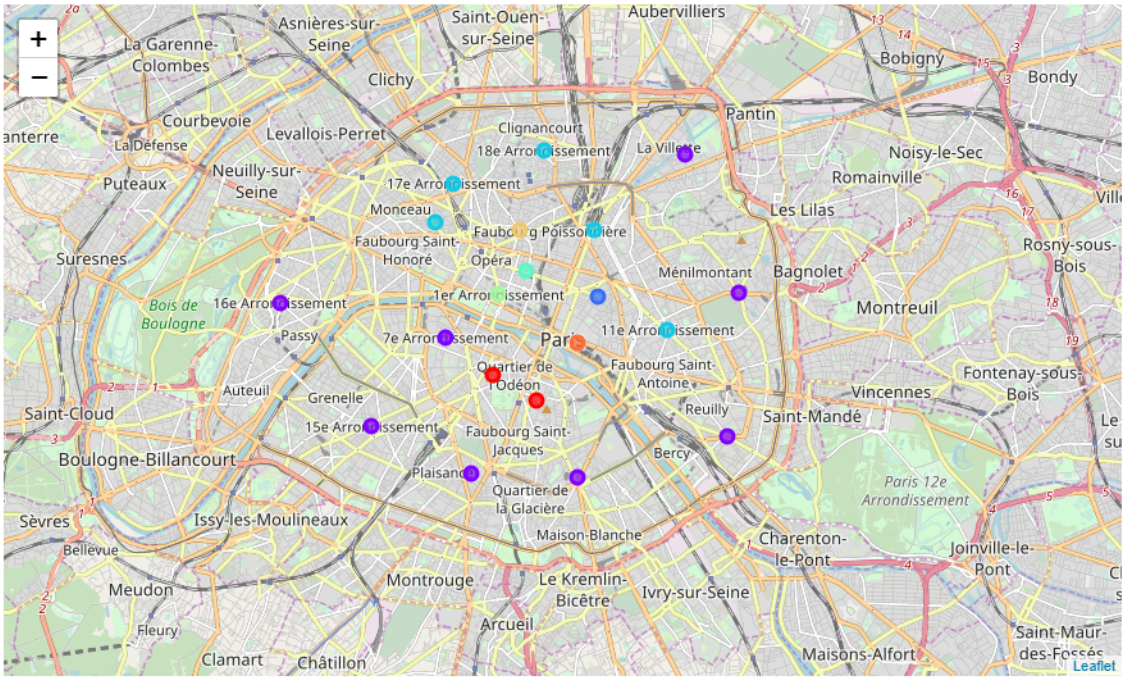

Figure 2 - Map of Paris with circles on each neighborhood identifying its cluster using a different color;

It is noticeable that neighborhoods on the periphery all belong in the same cluster and that different and smaller clusters appear as we get near the center of Paris. This was expected as usually center of cities tend to focus more on tourists while neighborhoods on the outside tend to be more for local population.

Target Cluster In-Depth Analysis

After I had segmented all neighborhoods into clusters I turned my attention to the problem that was presented to me: which type of nightlife venue should Louis open in his place at *Popincourt* neighborhood?

Popincourt belongs to Cluster nr. 3 together with neighborhoods *Batignolles-Monceau*, *Butte-Montmartre*, *Entrepôt* and *Élysée*. Let's take a look at the 5 most common venues in each of this neighborhoods (table 2).

Neigborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Batignolles-Monceau	Wine Bar	Cocktail Bar	Pub	Lounge	Beer Bar
Butte-Montmartre	Wine Bar	Cocktail Bar	Pub	Beer Bar	Lounge
Entrepôt	Wine Bar	Cocktail Bar	Pub	Beer Bar	Lounge
Popincourt	Wine Bar	Cocktail Bar	Beer Bar	Pub	Nightclub
Élysée	Wine Bar	Pub	Nightclub	Cocktail Bar	Bistro

Table 2 - Most common venues of each neighborhood from Cluster 3;

From the table it is immediately noticeable the similarities between each neighborhood. There are 3 categories (Wine Bar, Cocktail Bar and Pub) that are present in the top 5 venues of all neighborhoods and another category (Beer Bar) that is in the top 5 of all but one neighborhood. I took a further look at the number of each type of venue per km^2 in all neighborhoods (table 3).

Neigborhood	Nr. Of Places / km²			
	Wine Bar	Cocktail Bar	Pub	Beer Bar
Batignolles-Monceau	2,29	1,59	1,59	0,53
Butte-Montmartre	1,50	1,50	1,33	1,00
Entrepôt	2,89	2,42	1,73	1,04
Popincourt	3,67	3,00	1,36	1,36
Élysée	3,88	1,55	2,83	0,51

Table 3 - Number of places per unit of area in each neighborhood;

And finally I compared these values for Louis neighborhood, *Popincourt*, with the average from all other neighborhoods from same cluster (table 4).

	Wine Bar	Cocktail Bar	Pub	Beer Bar
Cluster Average	3,00	2,01	1,76	0,89
Popincourt	3,67	3,00	1,36	1,36
Difference	+22,53%	+49,2%	-22,92%	+53,42%

From the data I found that *Popincourt* has above average number of venues for each top category except Pubs, which has less 23% than average of other neighborhoods in same cluster.

Conclusions

In this project I analyzed Paris's neighborhoods based on type of venues present in each one, more specifically nightlife venues, with the purpose of helping decide which type of venue has most potential to be successful in the neighborhood of *Popincourt*. I clustered the neighborhoods using k-means method and concluded that *Popincourt* belongs in same cluster as *Batignolles-Monceau*, *Butte-Montmartre*, *Entrepôt* and *Élysée*. In this cluster most common type of nightlife venues are Wine Bars, Cocktail Bars, Pubs and Beer Bars. Considering that a large presence of a given type of venue is indicative of the appealing factor that that venue has on the people of a certain neighborhood we can consider this four categories as top choices for the new venue.

Furthermore I compared number of venues in *Popincourt* with average in the other neighborhoods from same cluster and realized that *Popincourt* has more venues per km² of all categories except Pubs. This additional information tells us that Pubs is the category least likely to be saturated in this neighborhood and that would be my final suggestion.

Future Directions

Small amount of usable Data is the biggest problem with this project. Almost half of all samples of nightlife venues had to be dropped as they were referred to only as "Bar" which is too vague to help clustering neighborhoods and deciding which type of venue is most likely to be successful. Finding a way to get specific venue information for those lost samples could help improve precision of clustering method.