Analysis & exploration of car parks according to proximity of amenities in Paris city

Marouane ALAMI - 2019

Business Intelligence & Data Sciences consultant

Table des matières

A. Introduction

1. Description & Discussion of the background

Paris is the capital and most populous city of France, with an area of 105 square kilometers (41 square miles) and an official estimated population of 2,140,526 residents as of 1 January 2019.[1] Since the 17th century, Paris has been one of Europe's major centers of finance, diplomacy, commerce, fashion, science, and the arts.

Whether you are Parisian or coming for tourism in Paris, or for a particular event taking place in the capital, if you are planning on getting there by car, we do not doubt that you have tried to understand **circulating and parking in Paris**.

The mayor has decided an application (named PARISPARCS in our example) to citizens to list the available parking spaces. A company specializing in the field has been mandated to carry out this project. the project team needs to know which car parks have close proximity to all the conveniences, in order to propose it to the application's users, also allow them the possibility **to park their cars as quickly as possible and to find themselves in a place full of amenities that meet theirs needs**

2. Data Description

To consider the problem we can list the data as below:

- Use the cars parks data set found in the https://opendata.paris.fr as different formats, the **.json** file was downloaded and used for feeding the prime Dataframe.
- Foursquare API is used to get the most common venues of a given Park

The direct link to download the data set is : https://opendata.paris.fr/explore/dataset/parcs-de-stationnement-concedes-de-la-ville-de-paris/information/

The Parks data set is structured with 174 parks and more than 30 fields.

The useful fields in our case are :

- ID: id of the park
- NOM_PARC: park's name
- latitude: latitude position of the park
- longitude: longitude position of the park

For some Foursquare restrictions (number of access), the study will be focused on just 30 parks.

This is a head of the used dataframe:

	ID	NOM_PARC	latitude	longitude
0	94	CAMBRONNE (SURFACE)	48.847687	2.303116
1	190	CAMBRONNE (SURFACE)	48.847333	2.302846
2	111	DIDOT	48.833902	2.321473
3	149	MANDEL 2 (SURFACE)	48.863838	2.277569
4	176	PORTE D ORLEANS	48.820383	2.325949

3. Methodology

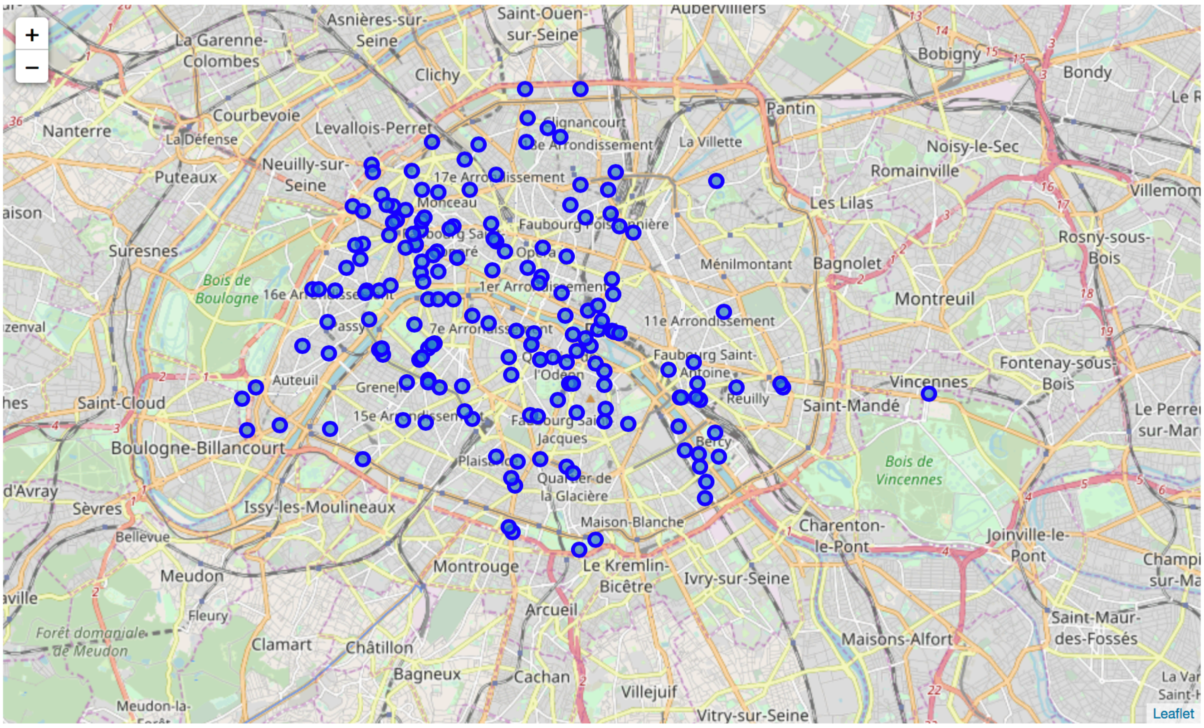
Prime Data

As presented in the previous section (Data description), our main dataset will be based on 4 fields:

- id
- parc name
- latitude & longitude let us check the avenues on the Foursquare API and to show the parcs on the map.

Parcs map

This is a Paris map representation for 174 parcs :



Retrieve all nearby venues for all parcs

We need to get all venues near each parc and feed our data by avenue's categories and segment them. We used the Foursquare API to explore the parc's venues. we designed the limit as **100** venue and the radius **500** meter for each parc from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Forsquare API.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	94	48.847687	2.303116	La Villa Corse	48.848220	2.300556	Corsican Restaurant
1	94	48.847687	2.303116	ibis Styles Paris Eiffel Cambronne	48.848113	2.300731	Hotel
2	94	48.847687	2.303116	El Farès	48.848120	2.300811	Lebanese Restaurant
3	94	48.847687	2.303116	La Bodega	48.845364	2.302019	Spanish Restaurant
4	94	48.847687	2.303116	La Cantina	48.847348	2.305497	Italian Restaurant

For this result, There are **324** unique categories

This categories volume is higher, so we reduced it in our study to **10 categories** by matching and regrouping the wub categories into principal ones.

This is the final categories list:

- RESTAURANT
- HOTEL
- CAFÉ
- BAR
- SHOP
- ENTERTAINMENT
- STORE
- GALERIE
- TRANSTPORT
- ADMINISTRATION

Structuring the data

The table below shows the categories occurrences for each parc (this is just a part of the data)

	Neighborhood	ADMINISTRATION	BAR	CAFÉ	ENTERTAINMENT	GALERIE	HOTEL	RESTAURANT	SHOP	STORE	TRANSTPORT
0	94	0	0	0	0	0	0	1	0	0	0
1	94	0	0	0	0	0	1	0	0	0	0
2	94	0	0	0	0	0	0	1	0	0	0
3	94	0	0	0	0	0	0	1	0	0	0
4	94	0	0	0	0	0	0	1	0	0	0

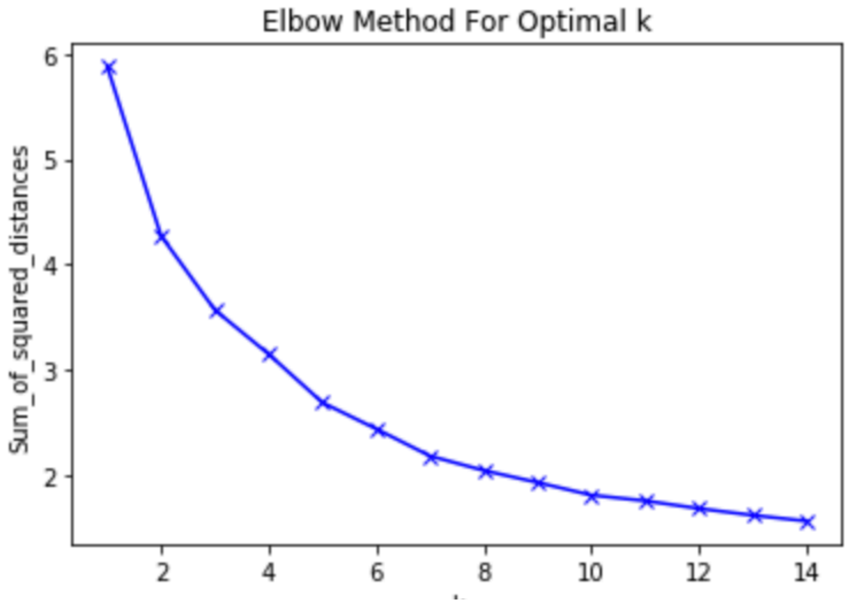
Based on the result table, we concluded the top 10 common venues for each parc

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10t Common
0	1	RESTAURANT	HOTEL	ENTERTAINMENT	SHOP	BAR	TRANSTPORT	STORE	GALERIE	CAFÉ	ADMINISTF
1	2	RESTAURANT	SHOP	HOTEL	STORE	ENTERTAINMENT	BAR	CAFÉ	TRANSTPORT	GALERIE	ADMINISTF
2	3	RESTAURANT	SHOP	ENTERTAINMENT	BAR	STORE	HOTEL	CAFÉ	GALERIE	TRANSTPORT	ADMINISTF
3	4	RESTAURANT	SHOP	ENTERTAINMENT	BAR	STORE	GALERIE	HOTEL	CAFÉ	TRANSTPORT	ADMINISTF
4	6	RESTAURANT	HOTEL	STORE	SHOP	ENTERTAINMENT	BAR	CAFÉ	TRANSTPORT	GALERIE	ADMINISTF

Building model for clustering the parcs based on their nearby categories

We have some common venue categories in boroughs. In this reason I used unsupervised learning **K-means** algorithm to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning.

First, we will run K-Means to cluster the boroughs into 5 clusters because when I analyze the K-Means with Elbow method it ensured me the **5** degree for optimum k of the K-Means.

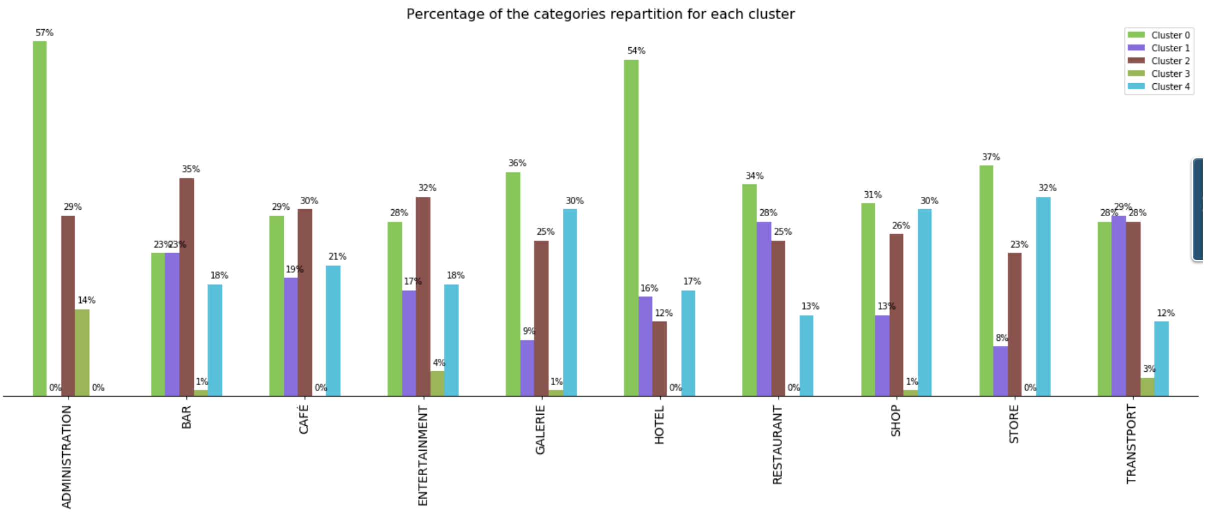


As we remark, the 5 degree is the optimal number of clusters

After fitting the model on our data, we tried to explore the categories for each cluster

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ADMINISTRATION	0.57	0.00	0.29	0.14	0.00
BAR	0.23	0.23	0.35	0.01	0.18
CAFÉ	0.29	0.19	0.30	0.00	0.21
ENTERTAINMENT	0.28	0.17	0.32	0.04	0.18
GALERIE	0.36	0.09	0.25	0.01	0.30
HOTEL	0.54	0.16	0.12	0.00	0.17
RESTAURANT	0.34	0.28	0.25	0.00	0.13
SHOP	0.31	0.13	0.26	0.01	0.30
STORE	0.37	0.08	0.23	0.00	0.32
TRANSTPORT	0.28	0.29	0.28	0.03	0.12

This is a data vitalization for the cluster categories repartitions. This is useful for deciding on which parc to choose based on the user need:



When we examine above graph we can label each cluster is known by its main category:

Cluster	Main known categories
Cluster 0	Administration & Hotels
Cluster 1	Transport
Cluster 2	Bar & Entertainment
Cluster 3	Administration
Cluster 4	Stores & Galeries

4. Results

Prime Data

Once, we got clusters, we merged them with the main data to get for each parc the following useful information:

- Parc ID
- Parc Name
- Latitude
- Longitude
- Membership cluster
- Most Common venues

	ID	NOM_PARC	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	(
0	94	CAMBRONNE (SURFACE)	48.847687	2.303116	0	RESTAURANT	HOTEL	SHOP	ENTERTAINMENT	STORE	CAFÉ	BAR	
1	190	CAMBRONNE (SURFACE)	48.847333	2.302846	0	RESTAURANT	HOTEL	SHOP	ENTERTAINMENT	STORE	CAFÉ	BAR	
2	111	DIDOT	48.833902	2.321473	0	RESTAURANT	HOTEL	BAR	SHOP	ENTERTAINMENT	STORE	CAFÉ	
3	149	MANDEL 2 (SURFACE)	48.863838	2.277569	1	RESTAURANT	SHOP	CAFÉ	TRANSTPORT	STORE	HOTEL	GALERIE	E
4	176	PORTE D ORLEANS	48.820383	2.325949	0	RESTAURANT	HOTEL	ENTERTAINMENT	SHOP	BAR	STORE	CAFÉ	

In the section bellow, we will represent the parc on the map in function with their membership clusters.

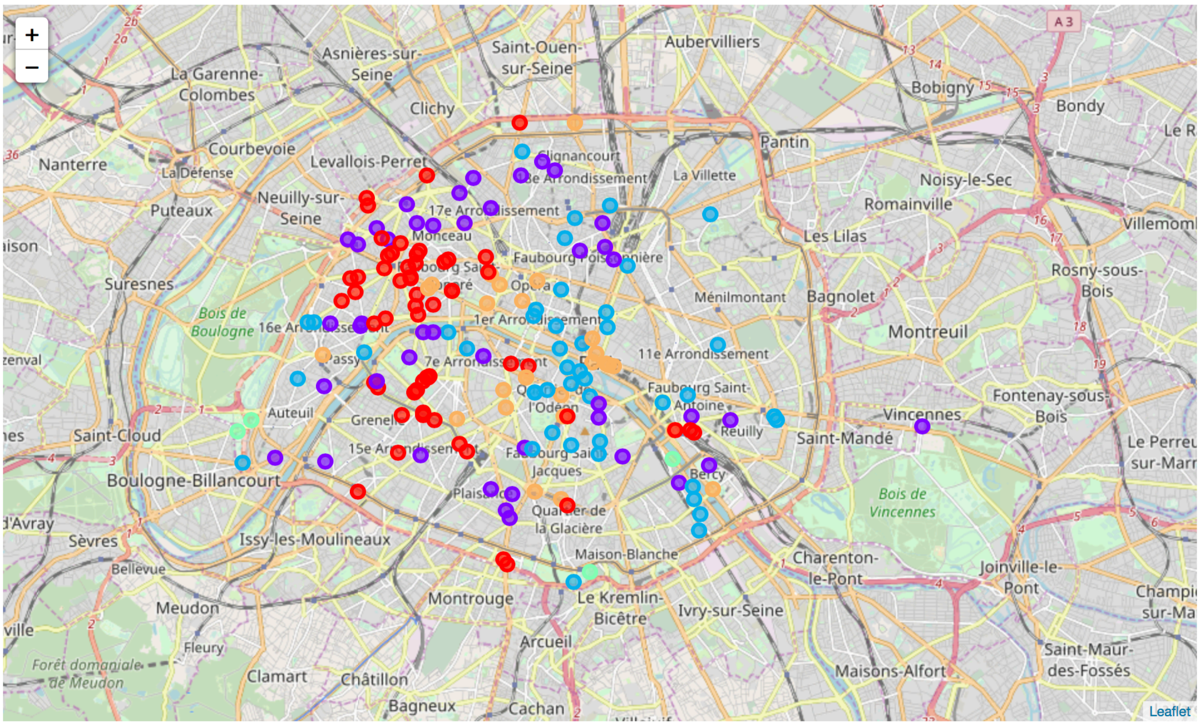
Clusters map representation

Map clusters is graphically represented by 5 colors (5 clusters)

Cluster	Color	Map representation
Cluster 0	Red	
Cluster 1	Purple	

Cluster 2	Blue	
Cluster 3	Green	
Cluster 4	Orange	

Bellow, the map representation:



5. Discussion

In each use case study, preparing data is the main step, in this example, I got the data from the Paris provided data site, and according to the analysis need, I enriches with needed information from the Foursquare API to get the nearby venues for each parc.

In a second stage, I made mapping between the main categories and subcategories to reduces the number of the unique ones.

I used the Kmeans algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 5. After fitting the model with 5 clusters, the results are pertinent, according to the geographical streets in Paris.

I also performed data analysis through this information by adding main data to the COGNITIVECLASS and GITHUB. In future studies, these data can also be accessed dynamically from specific platforms or packages.

The future of this study, we can also explore more points in the real time (inside the PARISPARCS application) by fitting the model once a new entry is detected (new restaurant, trends, new place extension…)

6. Conclusion

As a result, people using “PARISPARCS”, they will have ability to have for each parc in Paris, the number of free available places, also the most information about nearby amenities. The data may be updated in the platform as needed to fit the model new amenities values and generate the new real clusters.

For managers can manage the city more regularly by using similar data analysis types or platforms.

To the future,

Marouane ALAMI

7. References

- Eblow method:

  https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6

- The parcs Data set
  https://opendata.paris.fr/explore/dataset/parcs-de-stationnement-concedes-de-la-ville-de-paris/information/

- Development Environment (Cognitive classes)
  https://labs.cognitiveclass.ai

- Github notebook
  https://gist.github.com/alamimarouane/40d711d527dba15241f05d07b1b6b52d