Clustering Paris (France) Districts

By C.E Ramamonjisoa | Capstone for Data Scientist Certification

INTRODUCTION AND CONTEXT

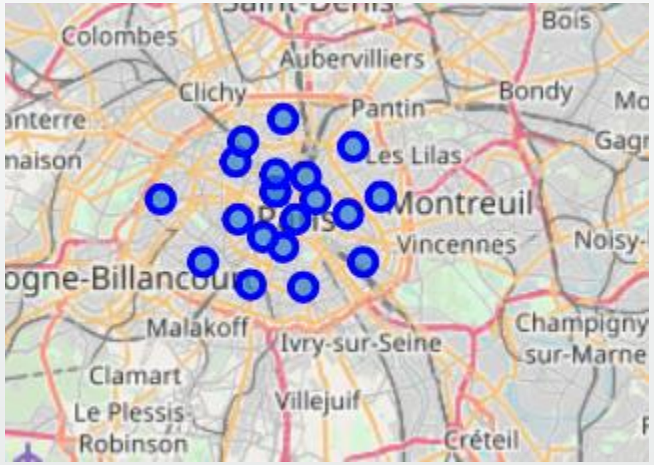
- Paris, the capital city of France : large metropolis with more than 2.2 million inhabitants with a rich history and a cosmopolitan and multicultural population.
- Divided into 20 districts from the first to the 20th. The district is an administrative division, headed by an elected mayor.
- Dense installation of venues and interesting places (restaurant, hotels, café, parks, museums, ...).
- Dispersed population distribution in districts

BUSINESS PROBLEM

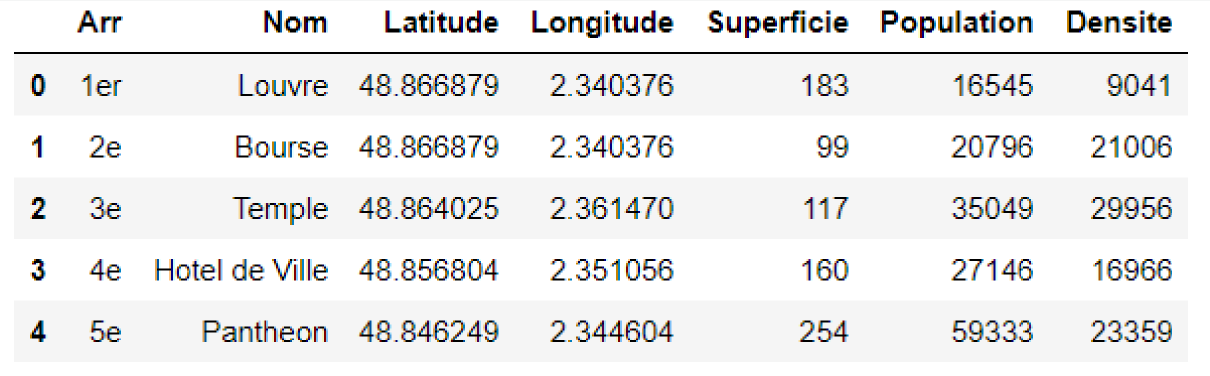
- Segment the 20 districts of Paris to group those who presents some similarities and characteristics in terms of venues categories around each district.
- Consider the number of inhabitants in the analysis.
- The objective is to have a tool to guide any users for decision making to respond to the following questions:
  - If I want to open a new restaurant in Paris, depending on the type of my restaurant, in which district would I create it according the existing restaurant in the area ?
  - If I want to rent a house, in which district can I look first to fullfill my needs in terms of local amenities and quality of life?

DATA INVENTORY, DESCRIPTION AND SOURCES

- Data collection (Wikipedia, …)
  - ❖ The geospatial coordinates of Paris (France)
  - ❖ The order and the name of the 20 Paris Districts.
  - ❖ The coordinates of a location in each district: we can take here the well-known coordinates of the Hall of the City in each district.
  - ❖ The number of inhabitants in each district.
  - ❖ The area of each district



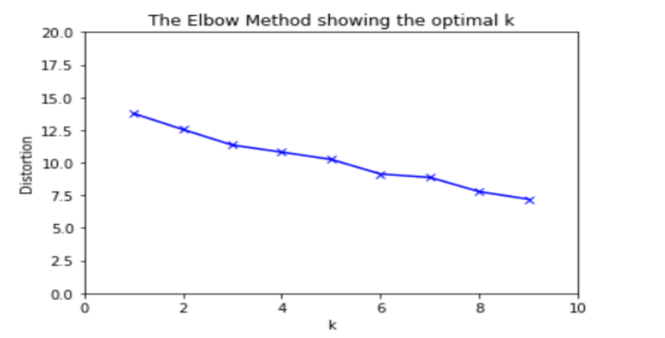
- Venues by categories (Foursquare)
  - ❖ Number of venues per district = 150
  - ❖ Area of collect = 1,500 meters around the location.



	Arr	Nom	Latitude	Longitude	Superficie	Population	Densite
0	1er	Louvre	48.866879	2.340376	183	16545	9041
1	2e	Bourse	48.866879	2.340376	99	20796	21006
2	3e	Temple	48.864025	2.361470	117	35049	29956
3	4e	Hotel de Ville	48.856804	2.351056	160	27146	16966
4	5e	Pantheon	48.846249	2.344604	254	59333	23359

METHODOLOGY AND APPROACH

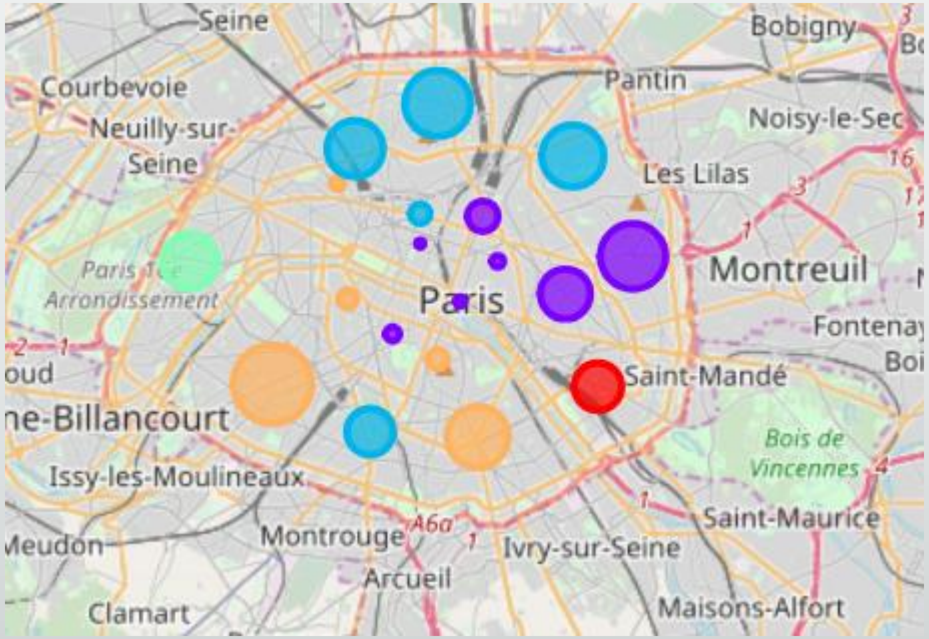
- ✓ Load data into pandas data frames from csv files.
- ✓ Use Foursquare API for places geo-localisation.
- ✓ Use Google Geocoder to get the coordinated of some places in Paris by districts.
- ✓ Use Folium libraries to visualize the places in a map.
- ✓ Search the first 150 venues by district at 1500 meters around with their geospatial coordinates¶
- ✓ Get the categories of each venue (1,513 for the 20 districts with the previous restriction).
- ✓ Group venues by categories (205 unique categories).



- ✓ Select the 10 most common venue for each district.
- ✓ Use the machine learning algorithm K-Means to segment Paris districts to 5 clusters by categories and the population.
- ✓ Use the elbow method for K-Means with best value of K.
- ✓ Visualize the results in a map with specific color for each cluster.
- ✓ Characterize each cluster from the most common venue and the number of inhabitants.
- ✓ Revert back to the initial business problem and discuss.

ANALYSIS RESULTS

- 5 groups (clusters) of districts having similar characteristics in terms of existing venues and places.
- Each district with their respective number of inhabitants presented in a map.
- Worksheet resulting from our analysis can help us to respond to our initial questions.



CLUSTER #	Number Of Districts	DISTRICTS	CHARACTERISTICS (Segmentation)
1	1	12th – Reuilly	Residential district with many parks/gardens and commodities (Hotel, Restaurants, …)
2	8	1st - Louvre 2nd - Bourse 3rd – Temple 4th – Hotel de Ville 6th – Luxembourg 10th - Entrepot 11th - Popincourt 20th – Mesnilmontant	Particularly provided in terms of food (Restaurants, Café, Bar, Bakery, Bistro with the lowest population in Paris.
3	5	9th – Opera 14th - Observatoire 17th - Batignoles-Monceau 18th – Buttes-Montmartre 19th – Buttes-Chaumont	Most popular districts in Paris with a lot of hotels and bars (wine) but not so much restaurants.
4	1	16th – Passy	Bourgeois population with high density with commodities like café and bakery.
5	5	5th – Pathéon 7th – Palais Bourbon 8th - Elysée 13th - Gobelins 15th - Vaugirard	Multicultural and very popular districts with all commodities (Hotel, Bar, Restaurant, café, …). Include universities and touristic places.
	20		

DISCUSSION

- “French Restaurant” category ignored in the analysis : not discriminatory.
- Homogenous clusters in terms of categories of venues but also in terms of number of inhabitants.
- Can identify the best cluster fitting with the initial requirements (business problem)
- Example : if we want to open a new restaurant targeting the student community, we should select one of districts in the cluster 5.

CONCLUSION AND FUTURE DIRECTION

- Report demonstrating the strength and the efficiency of data analysis coupled with the use of machine learning algorithm to solve a concrete business problem .
- Methodology based on collecting the venues around a location in each administrative division in the area of study.
- Data analysis and machine learning algorithm K-Means (reputed to be efficient in segmenting and clustering).
- Visualization of the results in a map.
- Future direction :
  - ✓ Strengthening the results in capturing some more data like economic situation of the inhabitants of each district or the existing means of transport in the district etc …
  - ✓ Cross-check the results using other machine learning models and algorithms.

THANK YOU

Q/A