Final Report

Clustering of Paris Arrondissements for e-scooter sharing services optimization

Helder Peixoto

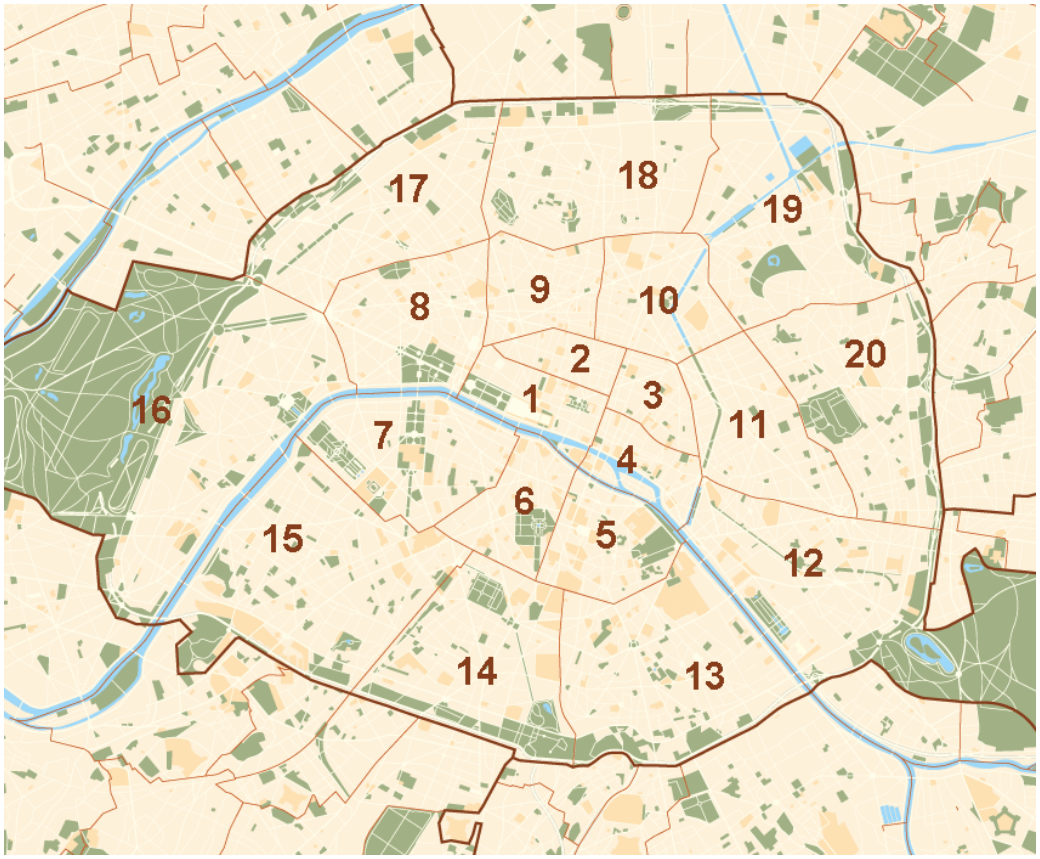
Introduction/Business Problem

A scooter-sharing system is a service in which scooters are made available to use for short-term rentals. The term describes the sharing of mostly electric motor scooters (also referred to as electric mopeds) as well as electric kick scooters. The sharing of scooters is similar to carsharing or bicycle-sharing systems; with some scooter-sharing companies offering more than one type of vehicle via their service. Recently, companies such as Bird, Lime and VOI began services in many cities across US and Europe.

Scooters are generally "dockless", meaning that they do not have a fixed home location. They are dropped off and picked up from arbitrary locations in the service area. Using a mobile app, the user have access to a map where they can check the real time location of available vehicles. This makes them a convenient mobility option in many situations such as the first-/last-mile mobility in urban areas. Moreover, it can also be a great option for tourists when visiting a city by being easily driven and handy.

In each city, there is a limited number of scooters available. There is a possibility that a user gets disencouraged to use the service because the nearest scooter is too far. To avoid that, we can more efficiently distribute the number of scooters per neighborhood based on the characteristics of the region. Such characteristics can be the most common categories of venues in the neighborhood, but also demographic data such as population density and population average age. If we also want to take into account the use of services by tourists we can use data such as the number of hotel rooms per neighborhood and also the number of landmarks/attractions/museums.

In this report the aim is to use machine learning techniques such as clustering, to identify the neighborhoods in a city in which there is a potential need of a bigger density of scooters available. The insights obtained from this work may be helpful for companies to better usage of their resources in field and potentially improve profits, but, at the same time, improving user's experience.



The report will be focused in the city of Paris, France. Accordingly to Business Insider, Paris was the 3rd most visited city in the world in 2018 with about 18 million foreign visitors and it is the 9th most populated city in Europe with about 2.2 million habitants in the central area. Central Paris is divided in 20 administrative districts called "Arrondissements" as you can see in the map below.

Paris is one the most important cities worldwide and for sure a big challenge for a company that wants to introduce its service. So, we hope to provide relevant information, allowing the decision making to become easier.

Data

There are several free datasets about the city of Paris. The main sources are we will use in this project to fetch demographic data and coordinates of administrative regions are Paris Data, Institut national de la statistique et des études économiques INSEE, Statista and UrbiStat.

The Arrondissements coordinates where obtained from this dataset at Open Data. This dataset will be important not only for graphical representation of each arrondissement in a map, but also to give the imput coordinates for the Foursquare API when searching for Venues. The area of each arrondissement was obtained from a Wikipedia page. The area dataset will allow us to calculate the population density.

Regarding demographic data, to get population per arrondissement we use a dataset from INSEE. This is the most up to date available data provided by the governmental institute. The age per arrondissement is available in this dataset from UrbiStat. Younger people may be more open up to the use of this type of transport.

As previously mentioned, in 2018 Paris was visited by 18 million people. We will try to take into account the most important locations for tourist. To do so we got this dataset from Statista which contains data about the number of hotel rooms per arrondissement. The hotel will be always the start/end point in the day of a visitor.

To obtain information about the venues existent in a given area we will use Foursquare. This API has a database of more than 105 million places. Many organizations are using it to geo-tag their photos with detailed info about a destination, while also serving up contextually relevant locations for those who are searching for a place to eat, drink or explore. This API provides the ability to perform location search, location sharing and details about a business. Foursquare users can also use photos, tips and reviews in many productive ways to add value to the results. The data collected will allow us to cluster similar arrondissements and obtain information about the most common type of venues in the area. For example, we can obtain the 10 most common categories of venues for every arrondissement. We will also use the Foursquare API to search for Landmarks/monuments/museums and quantify which locations are more likely to be the tourists destinations.

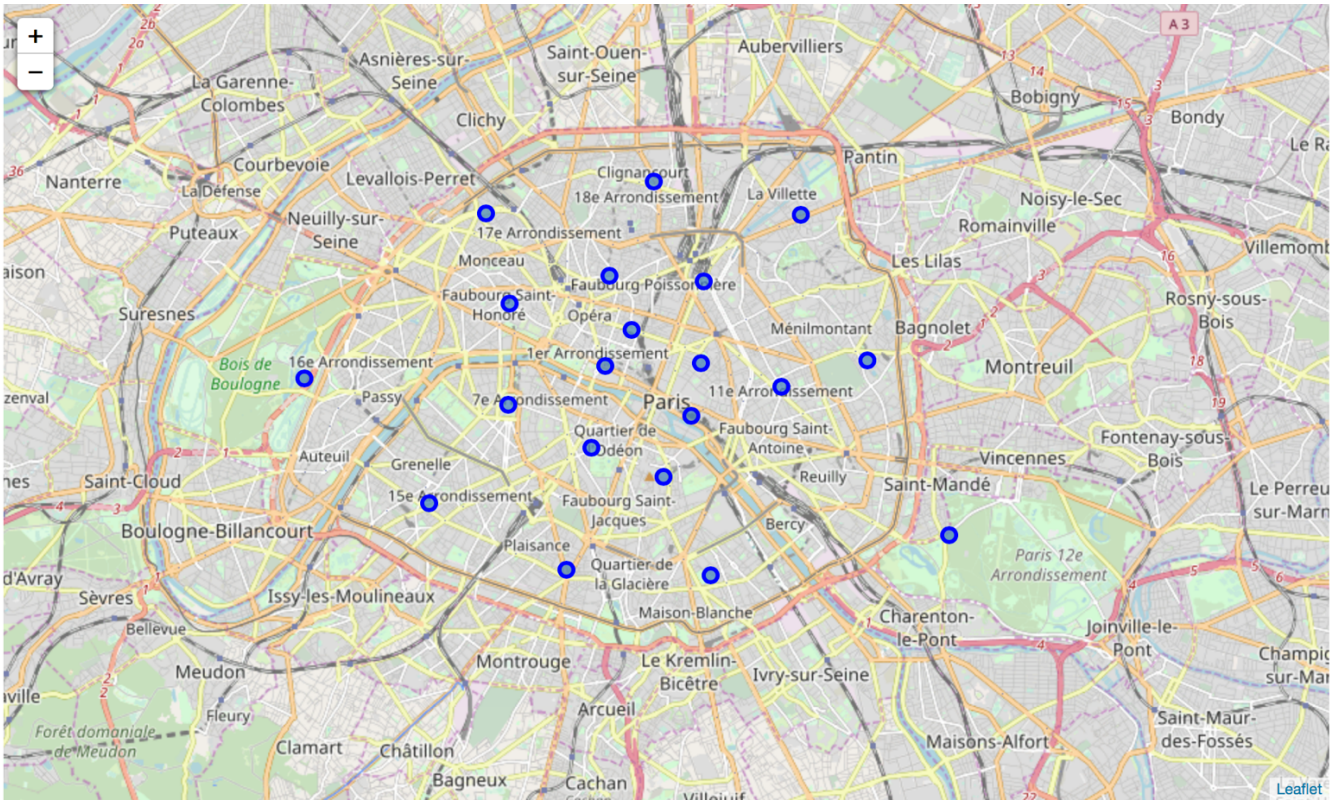
Methodology

The goal of this work is to determine which districts in the city of Paris require more density of scooters in order to provide better user experience and also better resources usage to the company.

In such a big city as Paris, many parameters may be used to characterize an area of the city, the most common categories of venues. If there are a lot of offices in the area, maybe that's more a business area than a residential area. If art galleries, museums, monuments are very common maybe that's the case of a more touristic area. For a city with more than 2 million people in the inner core, it may sound very complex to make, however using the Foursquare API function "search" it is possible.

The API retrieves a limited number of venues per query. So, to instead of only using the 20 arrondissements of paris to make 20 queries, we will use the subdivisions of each Arrondissement, called Quartier. There are 80 quartiers in Paris. Using this little modification, we can analyze the categories of more than 11000 venues instead of about 3000 if we only make queries for the 20 arrondissements.

To perform visual representation of geographical data we use Folium. In the image below we show the Arrondissements of Paris.



The next step after collecting the categories of venues in each Arrondissement, we will group them in terms of similarity. To do so, we will use a machine learning algorithm called K-means, which in present in the machine learning library SciKit Learn for Python.

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

- The centroids of the *K* clusters, which can be used to label new data
- Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined. Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

After the clustering of similar areas by venues category, we want to include demographic data such as the population density and mean age, movement of passengers in metro stations of each arrondissement and the share of hotel rooms per arrondissement. The data will be normalized by the maximum of each parameter.

Finally, we will run the K-means clustering, but this time the demographic data and touristic data will be used to cluster similar arrondissements. The arrondissements will be labelled by cluster number and we will be able to understand which features characterize each cluster and get insights about which arrondissements seem to require higher or lower density of scooters.

Results

From Foursquare API we retrived 11288 venues, distributed by 522 different categories. In the table below we represent the top 10 type of veneus per arrondissement. The arrondissements were clustered into 3 types as the present in the tables below.

Cluster 0

Arrondissement	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	48.862563	2.336443	0	Office	Jewelry Store	French Restaurant	Hotel	Art Gallery	Cosmetics Shop	Women's Store	Miscellaneous Shop	Café	Coffee Shop
3	48.862872	2.360001	0	Art Gallery	Office	Residential Building (Apartment / Condo)	Bar	Boutique	Coworking Space	Bakery	Salon / Barbershop	Café	Men's Store
4	48.854341	2.357630	0	French Restaurant	Office	Hotel	Art Gallery	Bar	Café	Arts & Crafts Store	Bakery	Clothing Store	Italian Restaurant
6	48.849130	2.332898	0	Art Gallery	Office	French Restaurant	Bookstore	Hotel	Bar	Furniture / Home Store	Restaurant	Café	Building
9	48.877164	2.337458	0	Hotel	Office	French Restaurant	Tech Startup	Bar	Italian Restaurant	Women's Store	Bakery	Asian Restaurant	Accessories Store

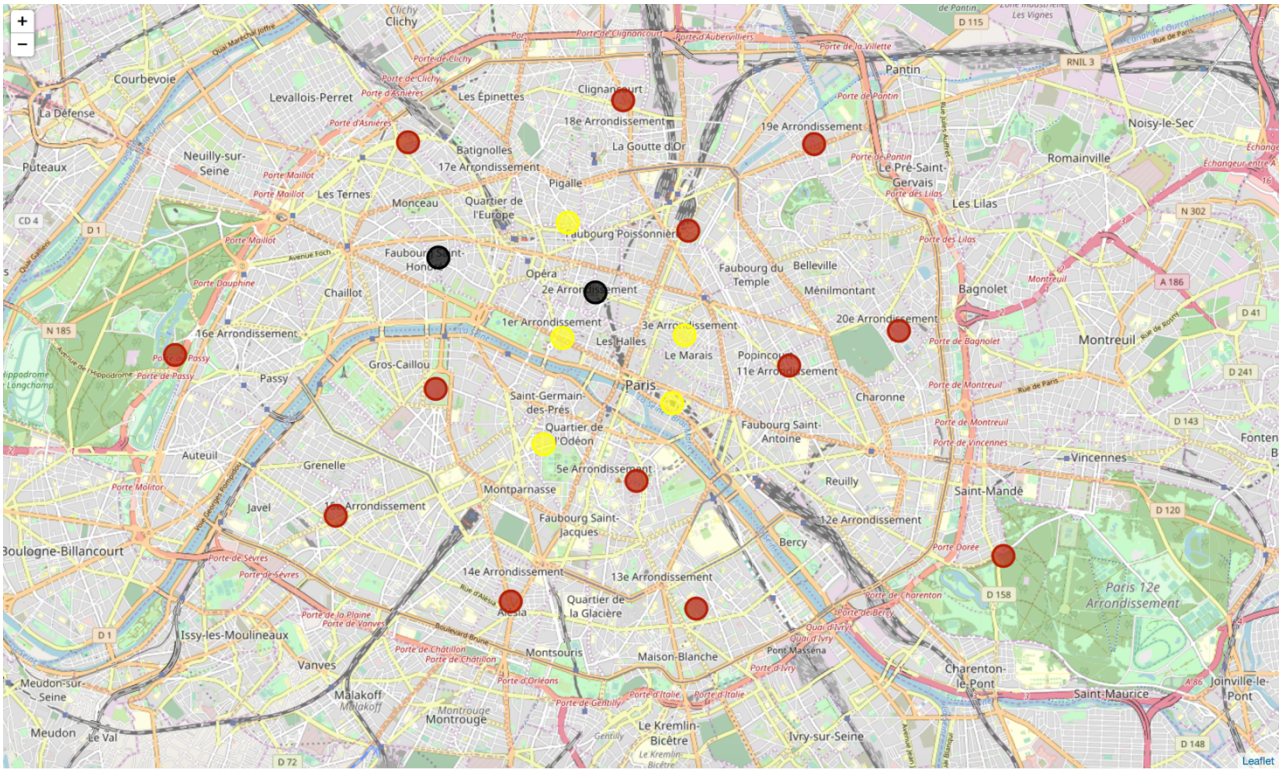
Cluster 1

Arrondissement	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	48.868279	2.342803	1	Office	Tech Startup	Building	French Restaurant	Coworking Space	Hotel	Japanese Restaurant	Salad Place	Salon / Barbershop	Sandwich Place
8	48.872721	2.312554	1	Office	French Restaurant	Building	Coworking Space	Hotel	Salad Place	Tech Startup	Residential Building (Apartment / Condo)	Boutique	Clothing Store

Cluster 2

Arrondissement	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
18	48.892569	2.348161	2	Tech Startup	Office	French Restaurant	Bar	Residential Building (Apartment / Condo)	Salon / Barbershop	Bakery	Pizza Place	Wine Shop	Café
17	48.887327	2.306777	2	Office	Residential Building (Apartment / Condo)	Bakery	French Restaurant	Coworking Space	Building	Supermarket	Tech Startup	Conference Room	Italian Restaurant
16	48.860392	2.261971	2	Office	French Restaurant	Tennis Court	Embassy / Consulate	Bank	Residential Building (Apartment / Condo)	Plaza	Government Building	Event Space	Bus Stop
15	48.840085	2.292826	2	Office	French Restaurant	Hotel	Residential Building (Apartment / Condo)	Bakery	College Classroom	Building	Salon / Barbershop	Doctor's Office	Bistro
14	48.829245	2.326542	2	Office	Salon / Barbershop	Doctor's Office	Residential Building (Apartment / Condo)	Bank	Hotel	Restaurant	Art Gallery	French Restaurant	Bar
10	48.876130	2.360728	2	Office	Tech Startup	Hotel	Bakery	French Restaurant	Building	Art Gallery	Indian Restaurant	Coworking Space	Bar
12	48.834974	2.421325	2	Office	French Restaurant	Bank	Salon / Barbershop	Bakery	Italian Restaurant	Restaurant	Hotel	Bike Rental / Bike Share	Café
11	48.859059	2.380058	2	Office	French Restaurant	Doctor's Office	Bar	Bank	Tech Startup	Hotel	Bakery	Japanese Restaurant	Chinese Restaurant
19	48.887076	2.384821	2	Office	College Classroom	Residential Building (Apartment / Condo)	French Restaurant	Bike Rental / Bike Share	Doctor's Office	Building	General Entertainment	Bakery	Non-Profit
7	48.856174	2.312188	2	Office	Government Building	French Restaurant	Café	Bank	Art Gallery	Bakery	Hotel	Residential Building (Apartment / Condo)	Salon / Barbershop
5	48.844443	2.350715	2	French Restaurant	Hotel	Office	Bookstore	College Auditorium	Building	Residential Building (Apartment / Condo)	Bar	Bike Rental / Bike Share	College Classroom
13	48.828388	2.362272	2	Office	Bakery	Hospital	Residential Building (Apartment / Condo)	French Restaurant	Italian Restaurant	Building	Dentist's Office	Grocery Store	Bar
20	48.863461	2.401188	2	Bakery	Bar	Cemetery	Office	Pizza Place	Bank	French Restaurant	Supermarket	Hotel	Bike Rental / Bike Share

The map below shows the clustered arrondissement. Yellow markers represent cluster 0, black cluster 1 and red cluster 2.

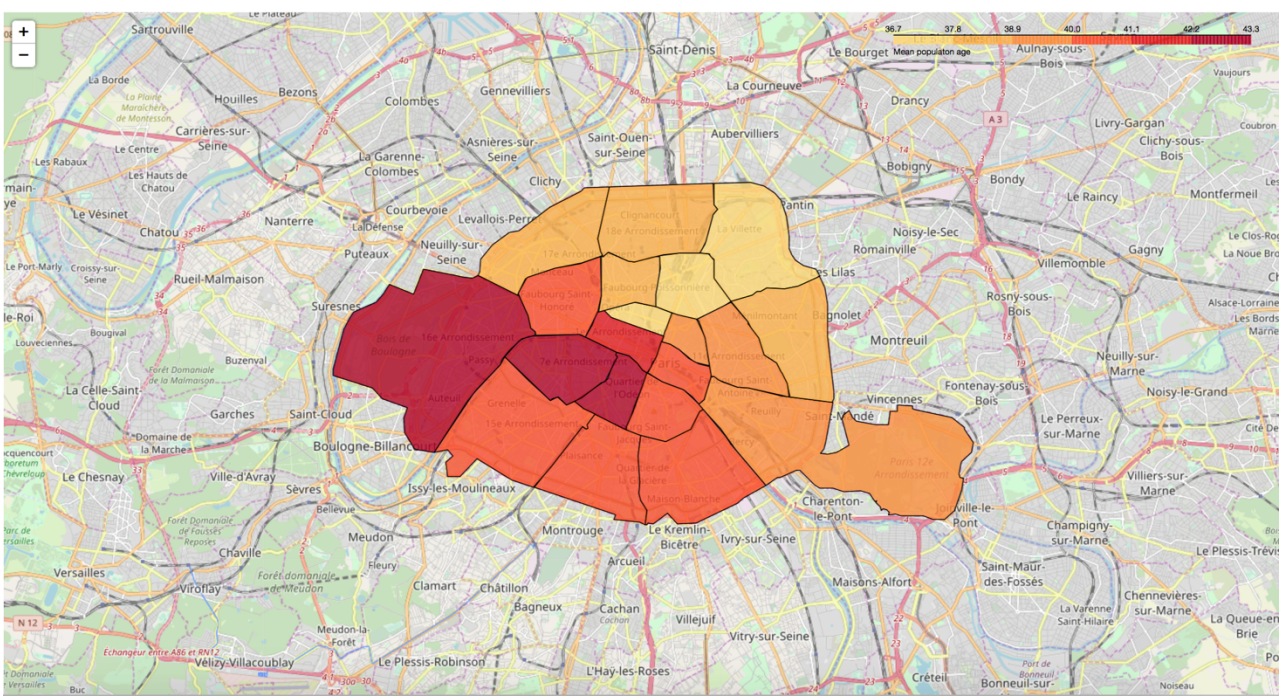


In the following table we present the dataframe will demographic data, touristic data, metro traffic and the labels from the venues clustering. After normalization this dataset was the input for the final clustering of the Arrondissements.

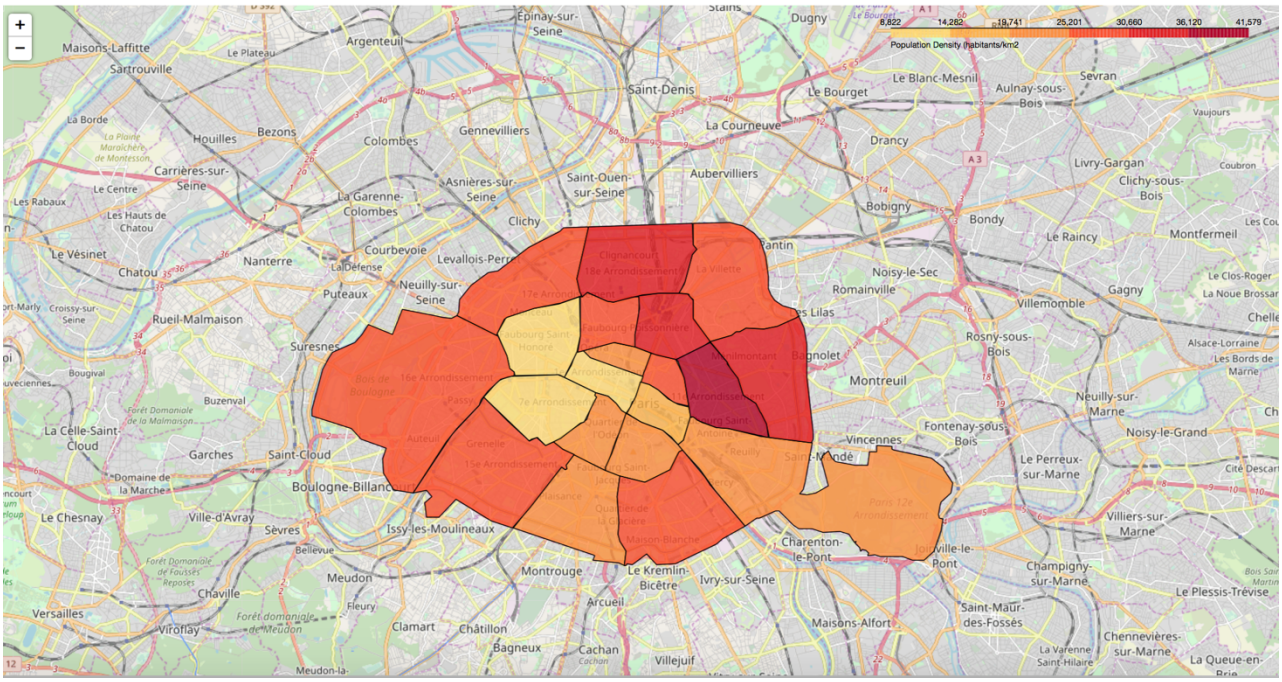
Arrondissement	mean age	density	hotel_rooms_share	metro_trafic	klabels_veneus
1	41.74	9143.48	5.1	46892170	1
2	36.72	21137.10	2.2	16335221	0
3	38.92	30529.46	1.2	10957857	2
4	40.91	17177.39	1.2	26589854	0
5	40.31	23692.25	3.7	19546551	0
6	42.47	20133.70	4.4	25105444	1
7	42.30	13488.26	2.6	23548365	2
8	40.10	9617.37	10.5	101780385	2
9	37.78	27583.75	10.3	66644419	2
10	36.90	32010.03	7.0	106929130	0
11	38.87	41258.32	4.5	58308937	0
12	39.76	22775.98	5.5	101752985	2
13	40.21	25867.76	2.7	74865649	2
14	40.10	25115.64	8.1	66280421	2
15	40.04	27886.14	8.0	53946732	2
16	43.22	26655.60	3.8	60649477	2
17	38.84	30022.93	9.3	59101557	2
18	38.00	33109.08	5.4	54150575	2
19	37.68	27568.67	2.8	48484851	2
20	38.86	32914.27	2.0	39759454	2

To better visualize this data we can create choropleth maps for each variable.

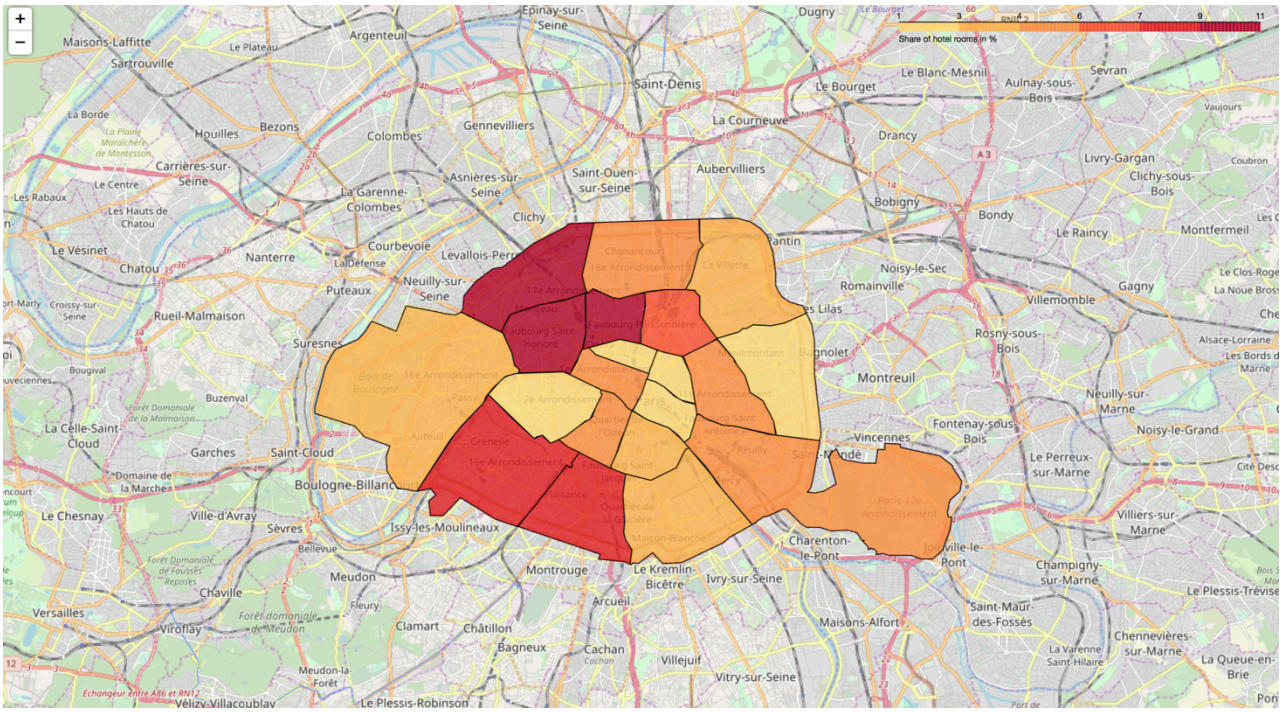
We start by looking at mean age of the population. Yellow represents younger population, red is the oldest. The arrondissements in the inner core and in left bank of the city have older population.



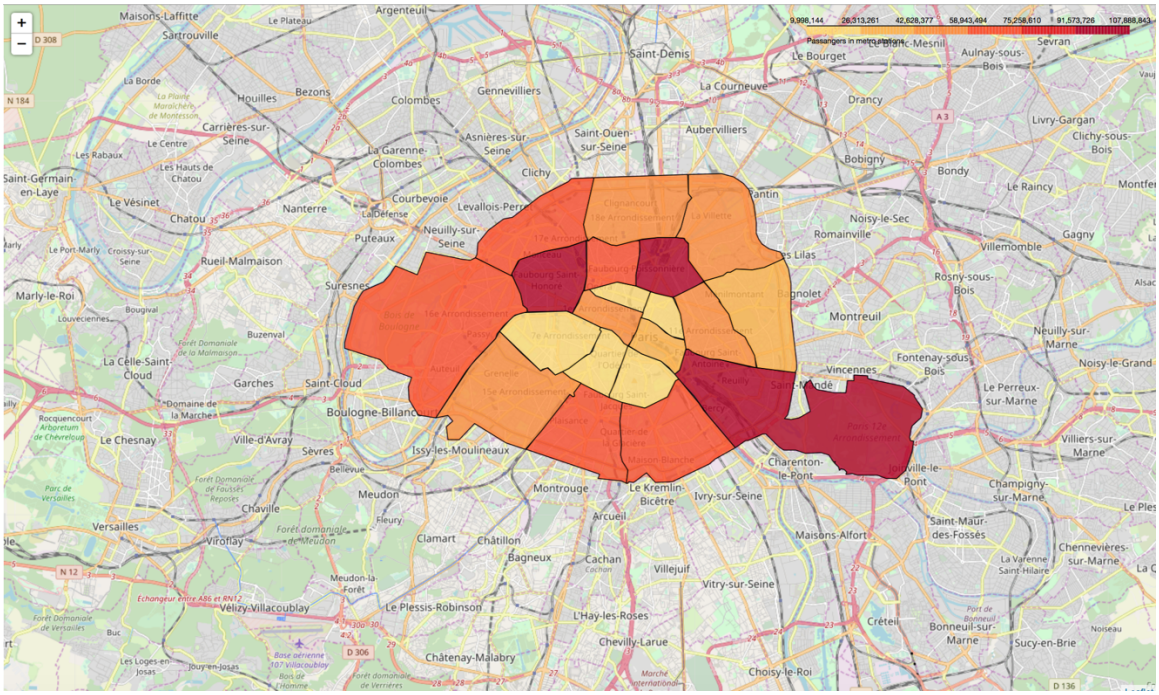
Regarding population density, the arrondissements outside the inner core of the city have the highest population density.



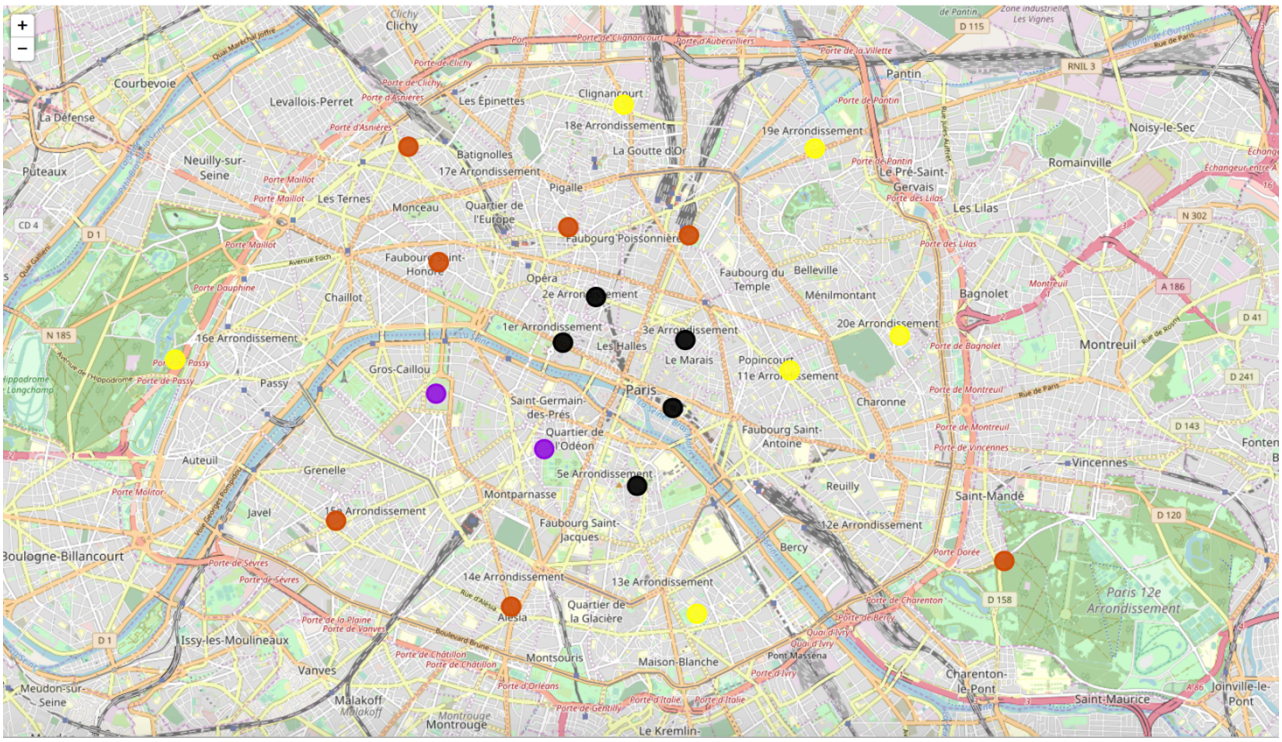
In what regards to the hotel rooms share, we present the results in the map below. The Arrondissements 8 and 9 have the highest number of hotel rooms in the city and the inner core Arrondissements have in general low shares.



Finally, let us look at the traffic in the metro station. The inner core Arrondissements have lower traffic. Arrondissement 8, 10 and 12 have the most movement metro stations.



The k-means procedure was again applied resulting into 4 different clusters. They are represented in the next map. Cluster 0 is represented with yellow, cluster 1 is black, cluster 2 is purple and cluster 3 is red.



Cluster 0 (Yellow)

Arrondissement	mean age	density	hotel_rooms_share	metro_trafic	veneus_klabels_veneus	Final Cluster Labels
11	38.87	41258.32	4.5	58308937	0	0
13	40.21	25867.76	2.7	74865649	2	0
16	43.22	26655.60	3.8	60649477	2	0
18	38.00	33109.08	5.4	54150575	2	0
19	37.68	27568.67	2.8	48484851	2	0
20	38.86	32914.27	2.0	39759454	2	0

Cluster 1 (Black)

Arrondissement	mean age	density	hotel_rooms_share	metro_trafic	veneus_klabels_veneus	Final Cluster Labels
1	41.74	9143.48	5.1	46892170	1	1
2	36.72	21137.10	2.2	16335221	0	1
3	38.92	30529.46	1.2	10957857	2	1
4	40.91	17177.39	1.2	26589854	0	1
5	40.31	23692.25	3.7	19546551	0	1

Cluster 2 (Purple)

Arrondissement	mean age	density	hotel_rooms_share	metro_trafic	veneus_klabels_veneus	Final Cluster Labels
6	42.47	20133.70	4.4	25105444	1	2
7	42.30	13488.26	2.6	23548365	2	2

Cluster 3 (red)

Arrondissement	mean age	density	hotel_rooms_share	metro_trafic	veneus_klabels_veneus	Final Cluster Labels
8	40.10	9617.37	10.5	101780385	2	3
9	37.78	27583.75	10.3	66644419	2	3
10	36.90	32010.03	7.0	106929130	0	3
12	39.76	22775.98	5.5	101752985	2	3
14	40.10	25115.64	8.1	66280421	2	3
15	40.04	27886.14	8.0	53946732	2	3
17	38.84	30022.93	9.3	59101557	2	3

Discussion

The results obtained in the final clustering revealed that the Arrondissements in the cluster 3 have a lower average age, high population density, high hotel room percentage and the highest traffic in the metro stations. In terms of venues this cluster is characterized by offices, hotels and residential buildings.

Cluster 2 is composed by higher average age people, lower population density, low hotel room share and low metro station traffic. The most common types of veneus are art galleries, offices, government buildings and French restaurants.

In cluster 1 we arrondissements with some of the lowest population densities (except Arrondissement 3). Hotel room share is also low metro station traffic. In what regards of venues, they are characterized by offices, art galleries and French restaurants.

In cluster 0 we find in general the youngest arrondissements (except Arrondissement 13, which actually has the highest average age). The arrondissements in this cluster have the highest population density, average hotel room shares and average traffic in metro stations. The veneus are mostly offices, French restaurant and residential buildings.

After this analysis of the characteristics of each cluster, we can observe that arrondissements in cluster 3 will potentially demand higher number of scooter vehicles. They have young people, the highest traffic in metro stations, high hotel room share, which means that a great percentage of tourists will start or end their days in one of these arrondissements. In contrast, Arrondissements 6 and 7 (cluster 2) are likely to demand the lowest number of vehicles. Arrondissements in the cluster 0, are also likely to require a high density of vehicles, but not so high as those in cluster 3. The arrondissements in the cluster 1, which represent the inner core of the city, are likely to require vehicles density similar to cluster 0, however the type of venues between this clusters are significantly different.

Conclusion

In this work we studied the Arrondissements of Paris with the goal of clustering the districts that potentially demand higher number of vehicles by e-scooter sharing companies. We successfully clustered the city in 4 different types using machine learning method k-means. The Arrondissements in cluster 3 (8,9,10,12,14, 15 and 17) are likely to have higher demands of scooters.