Best locations to set up tourist E-Scooter stations in Paris

1. Introduction

1.1. Background

Tourism is a major source of income in Paris. In 2017, 33.8 million tourists visited the city and her region. Like any other tourism cities, Paris provides her visitors with a wide variety of choices for transportation and sight-seeing, including public transport (metro, tram, bus), canal boat trips and hop-on-hop-off buses. These traditional transportation, however, implies significant disadvantages. For example, tourists need to plan the trip very carefully considering transport timetables. Also, the transportation network is designed to cover main public attractions only.

Modern tourism services aim to create a more personalized experience for visitors, allowing them to discover the city in the own plan and pace. Car rental may not be the first choice for most people due to parking and driver license requirement. Bicycle rental is a more flexible option, but not everyone can manage cycling up and down the many hills of Paris.

Electronic scooter (E-Scooter) comes as an excellent choice for personal sight-seeing. This means of transport allows riders to freely roam the city without having to worry about time, physical endurance and parking place. In this project, we plan to set up a series of rental stations for E-Scooter in the central area of Paris.

1.2. Business problem

On a natural approach, stations for E-Scooter rental should be places from which nearby points of interest are reachable by e-scooters within a reasonable radius. For a complete tourism experience for our customers, this project aims to *identify the best locations for E-Scooter stations to cover as many tourist attractions as possible*.

For this purpose, it is vital to have a comprehensive list of tourist attractions in Paris central. However, most tourist guides of Paris, such as this one www.planetware.com/tourist-attractions-/paris-f-p-paris.htm, tend to recommend only famous ones. There are many "hidden gems", for example *the Stravinsky Fountain* (https://www.atlasobscura.com/places/the-stravinsky-fountain-paris-france), that makes up the charm of Paris but are never captured in guide books. Luckily, such venues can be discovered based on users' reviews and ratings in public search-and-discovery location services like FourSquare.

1.3. Assumptions

The project is conducted under the following assumptions:

1. The company wants to identify 5 best locations to set up a chain of E-Scooter stations within Paris metropolitan area (i.e 10km from the city center)
2. The series of 5 locations is considered "best" to set up E-Scooter stations when:
    - Each location covers as much as possible the surrounding venues of interest within e-scooter distance
    - No venues of interest is too far from the nearest E-Scooter station.
3. The company has sufficient financial capabilities to build stations in every recommended location. In other words, factors like space or office rental cost of each areas will not be considered.

2. Data preparation

2.1. Data sources

To solve the business problem, this project collect data from the following sources:

1. A list of tourist attractions within a radius of 10km from the city center of Paris using FourSquare venue recommendation API
2. Number of likes and user rating for each venue in the list from FourSquare venue details API
3. After best locations have been identified, we will need to reverse geocoding coordinates of these locations to physical addresses.

2.2. Data collection

2.2.1. Get the list of venues

First step is to define the central point of Paris. This can be done by using FourSquare agent to translate the representative address *"Paris"* to longitude and latitude.

From this central point, we use the FourSquare venue recommendation API to explore venues within a radius of 10km. This API method returns all recommended visits around a given point, including different categories (food, drink, arts, outdoors, etc.). To make sure only tourist attractions are captured, parameter *query=tourist* is injected to the request URL. Result is a list of **250** points of interests:

	id	name	categories	address	city	country	distance
0	4bf41231e5eba59334341f90	Place de l'Hôtel de Ville – Esplanade de la Li...	Plaza	Place de l'Hôtel de Ville	Paris	France	60
1	4adcda09f964a520e83321e3	Cathédrale Notre-Dame de Paris	Church	6 place du parvis Notre-Dame	Paris	France	413
2	4b5c7d1ff964a5205f3229e3	Tour Saint-Jacques	Historic Site	88 rue de Rivoli	Paris	France	248
3	4adcda0af964a520623421e3	Centre Pompidou – Musée National d'Art Moderne	Art Museum	Place Georges Pompidou	Paris	France	458
4	4adcda0af964a520353421e3	Sainte-Chapelle	Church	8 boulevard du Palais	Paris	France	496
5	4bae535af964a520f5a23be3	Maison Européenne de la Photographie	Art Museum	5 rue de Fourcy	Paris	France	569
6	4cca7e73c4d06dcbb72d6303	Fontaine Stravinsky	Fountain	Place Stravinsky	Paris	France	330
7	4dbd336b6a23e294ba405cfa	Square de la Tour Saint-Jacques	Park	88 rue de Rivoli	Paris	France	245

These attractions belong to 48 distinct categories, including church, art museum, park, theater, etc.

```
In [33]: explore_df['categories'].unique()

Out[33]: array(['Plaza', 'Church', 'Historic Site', 'Art Museum', 'Fountain',
               'Park', 'Theater', 'Memorial Site', 'Garden', 'Museum',
               'Pedestrian Plaza', 'Bridge', 'Art Gallery', 'Concert Hall',
               'Monument / Landmark', 'History Museum', 'Opera House',
               'Botanical Garden', 'Sculpture Garden', 'Circus', 'Science Museum',
               'College Library', 'Canal', 'Trail', 'Event Space', 'Zoo',
               'General Entertainment', 'Arcade', 'Comedy Club', 'Library',
               'Cemetery', 'Street Art', 'Pool', 'Vineyard',
               'Performing Arts Venue', 'Theme Park Ride / Attraction', 'Island',
               'Outdoor Sculpture', 'Castle', 'Forest', 'Radio Station',
               'Rugby Stadium', 'TV Station', 'Soccer Stadium', 'Tennis Court',
               'Shopping Plaza', 'Racecourse', 'Stadium'], dtype=object)
```

2.2.2. Get number of likes and average user ratings for each venue

Next step is to get number of likes and user rating for each of the 250 venues in the list. Likes and rating are elements of Venue Details, which consumes a premium call for each venue. A FourSquare personal account allows 500 premium calls per day, which is sufficient to cover these 250 venues in a single batch.

After being loaded from FourSquare API, *number of likes* and *user rating* are appended as two additional columns to the dataset:

	id	name	categories	address	city	country	distance	no_of_likes	rating
0	4bf41231e5eba59334341f90	Place de l'Hôtel de Ville – Esplanade de la Li...	Plaza	Place de l'Hôtel de Ville	Paris	France	60	578.0	9.1
1	4adcda09f964a520e83321e3	Cathédrale Notre-Dame de Paris	Church	6 place du parvis Notre-Dame	Paris	France	413	8513.0	9.4
2	4b5c7d1ff964a5205f3229e3	Tour Saint-Jacques	Historic Site	88 rue de Rivoli	Paris	France	248	272.0	8.6
3	4adcda0af964a520623421e3	Centre Pompidou – Musée National d'Art Moderne	Art Museum	Place Georges Pompidou	Paris	France	458	5322.0	9.1
4	4adcda0af964a520353421e3	Sainte-Chapelle	Church	8 boulevard du Palais	Paris	France	496	583.0	9.1
5	4bae535af964a520f5a23be3	Maison Européenne de la Photographie	Art Museum	5 rue de Fourcy	Paris	France	569	358.0	9.0
6	4cca7e73c4d06dcbb72d6303	Fontaine Stravinsky	Fountain	Place Stravinsky	Paris	France	330	103.0	8.5
7	4dbd336b6a23e294ba405cfa	Square de la Tour Saint-Jacques	Park	88 rue de Rivoli	Paris	France	245	41.0	8.4
8	4b8ebd9df964a5203c3433e3	Théâtre du Châtelet	Theater	1 place du Châtelet	Paris	France	404	228.0	8.6
9	4adcda15f964a520a13721e3	Théâtre de la Ville	Theater	2 place du Châtelet	Paris	France	271	76.0	8.4

On a closer look, *number of likes* and *user rating* does not have any Null values.

```
In [40]: explore_df['no_of_likes'].isnull().any()
Out[40]: False

In [41]: explore_df['rating'].isnull().any()
Out[41]: False
```

Our raw data, containing 250 venues, all having user likes and rating, is now ready for further analysis.

3. Methodology

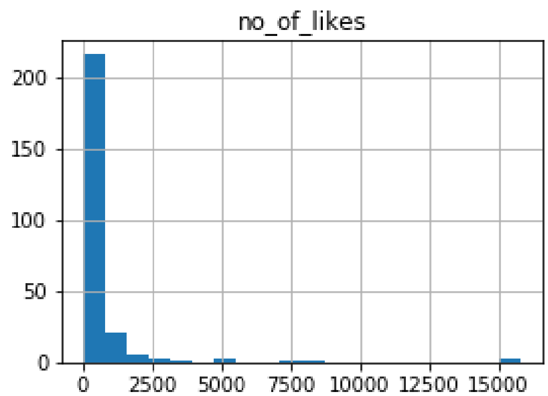
3.1. Data analysis approach

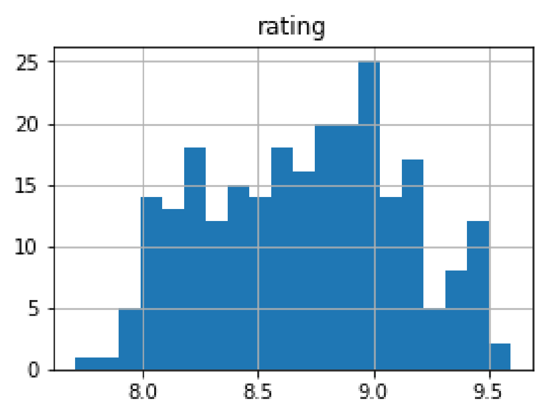
From the collected raw data, we will apply the following steps to identify 5 best locations for E-Scooter stations:

1. Explore columns *number of likes* and *rating* in the dataset. Define minimum threshold values of these fields for a venue to be considered. Filter the list to keep only venues with high reviews based on this threshold.
2. Run k-means clustering algorithm with k=5 on the remaining venues. Centroids of these clusters serve as the best locations to set up E-Scooter stations.
3. Reverse geocoding to translate coordinates of the 5 centroids into physical addresses.

3.2. Define thresholds to filter the list of venues

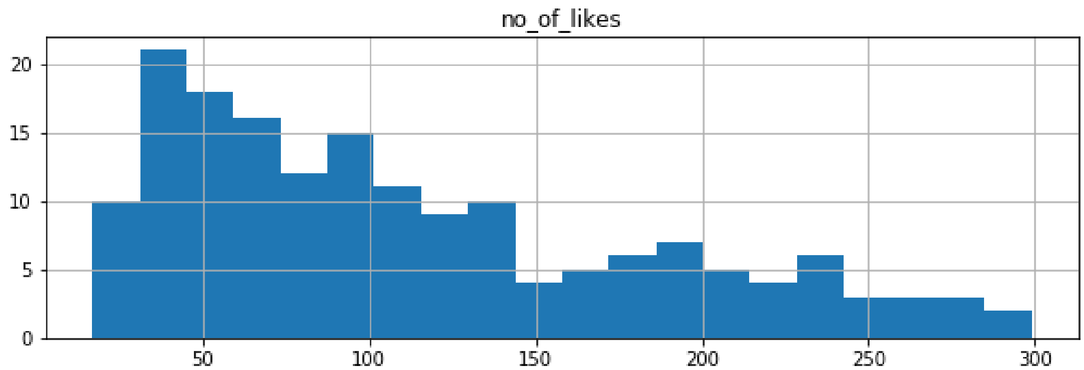
We want our customers to have the best tourist experience, so we are only interested in venues that are widely liked by FourSquare users. This means we will consider venues having relatively high *number of likes* and *user rating* in our dataset. Distribution of these fields are as follows:





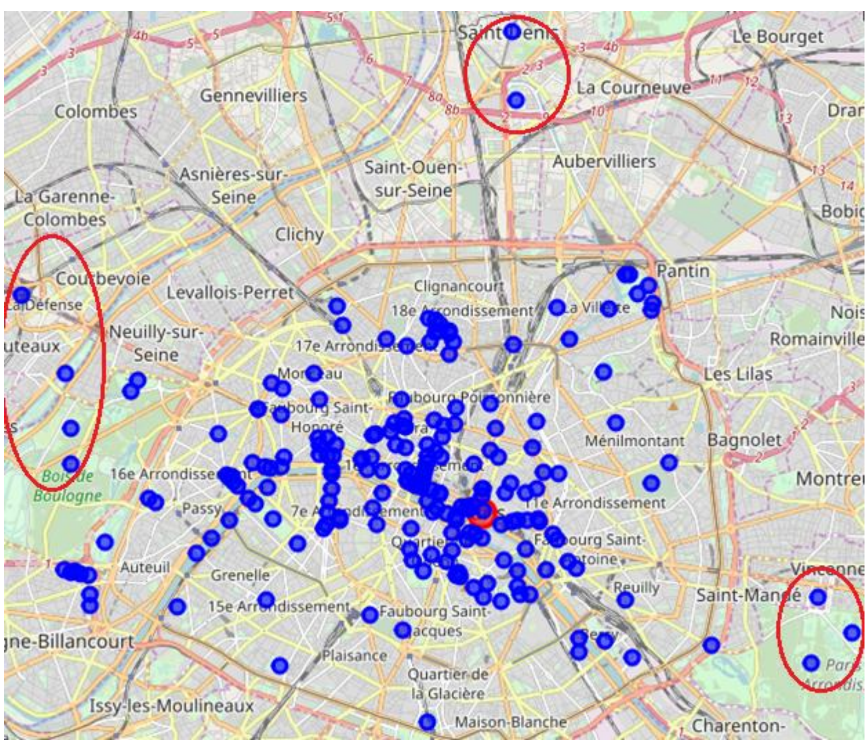
For *user rating*, the histogram is pretty close to a standard "bell shape" with majority of the values larger than 8.0. Therefore, minimum rating = 8.0 should be a reasonable threshold.

For *number of likes*, the plot is not that clear. It gives a rough idea that some particular venues (most likely famous ones like *Notre-Dame*) receive extraordinary high number of likes. Majority of other less well-known venues receive fewer than approximately 300 likes. Zooming into this range, we have the following histogram:



For this distribution, 50 should be an acceptable choice for the threshold of *number of likes*. Applying these two thresholds to the dataset (*number of likes >= 50* and *user rating >= 8.0*), we have a subset of **209** venues from the original dataset.

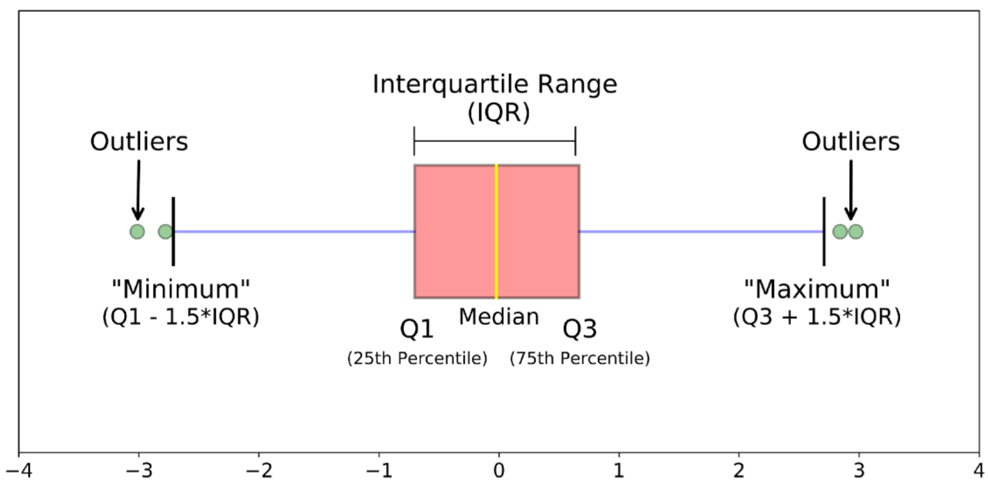
Let's plot all the 209 venues on a Paris map to have an initial idea of their distribution within the city. In the figure below, the red dot in the middle is the center of Paris identified in step ***2.2.1***



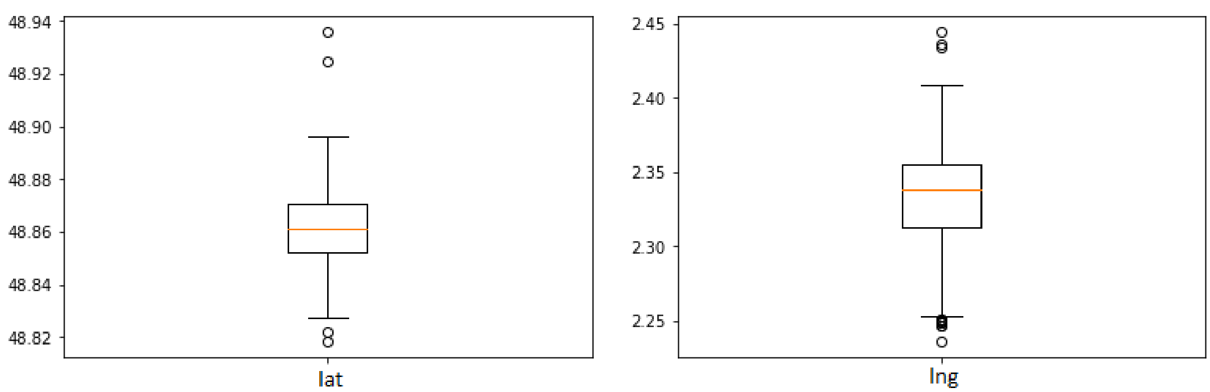
At a quick glance of the map, we can see that there are some venues on the north, west and south-east areas (marked with red circles) that are quite far away from the majority of other venues. Setting up E-Scooter stations to cover these venues seems not to be economically efficient, as they tend to “drag” the latest E-Scooter station towards them when running the clustering step. On an algorithm perspective, they can be considered as outliers that significantly affect centroids locations.

3.3. Outliers detection and removal

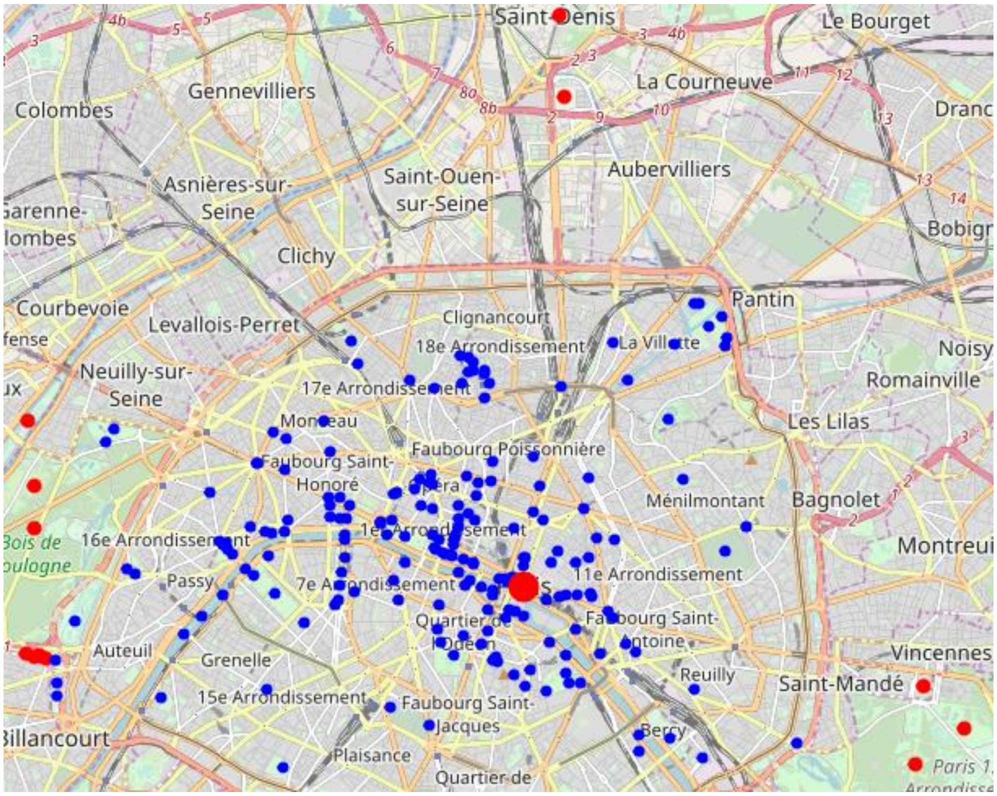
In our dataset, outliers are isolated venues that have exceptional coordinates compared to other venues. Based on this, we can draw box plots of venue *latitudes* and *longitudes*. Every point that fall outside standard whiskers (±*1.5 IQR*) will be considered as outliers.



https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51



After removing data having exceptional *latitude* and *longitude* values, **195** out of **209** venues remains in our dataset. As shown in the new map below, 14 venues considered as outliers (marked in light red smaller points) have been removed from the list of venues.



3.4. Machine learning application: K-Means clustering

Coming back to the ultimate purpose of the project: identify the most appropriate five locations to set up our E-Scooter stations. We want our customers to be able to quickly reach venues surrounding each E-Scooter station with minimal travelling time. We will therefore try to group venues based on their addresses, in a logic that nearby venues should belong to the same group. Once this is done, best locations for the stations should be "local centers" of 5 groups of venues.

Clustering is an appropriate choice of algorithm for this type of problem due to the following reasons:

- Nature of the problem is to group observations and find local central points of each group
- This is an unsupervised machine learning technique that works for datasets without pre-define labels such as our list of venues.

Clustering algorithm used in this case will be K-Means with K = 5. Input of the algorithm is the list of 195 venues with high likes/ rating scattered within 10km from Paris center that we have prepared above. Features being used in the algorithm are only coordinates (latitudes, longitudes) of venues. Output will be 5 clusters, centroids (central points) of those can serve as the best 5 locations for E-Scooter stations we are looking for.

After applying K-Means clustering, each venue is identified with a cluster label from 0-4. The resulting dataset is as follows:

```
In [61]: explore_df_no_outliers[['id','name','clusterLabel','categories','address','city','country','no_of_likes','rating']].head(10)
```

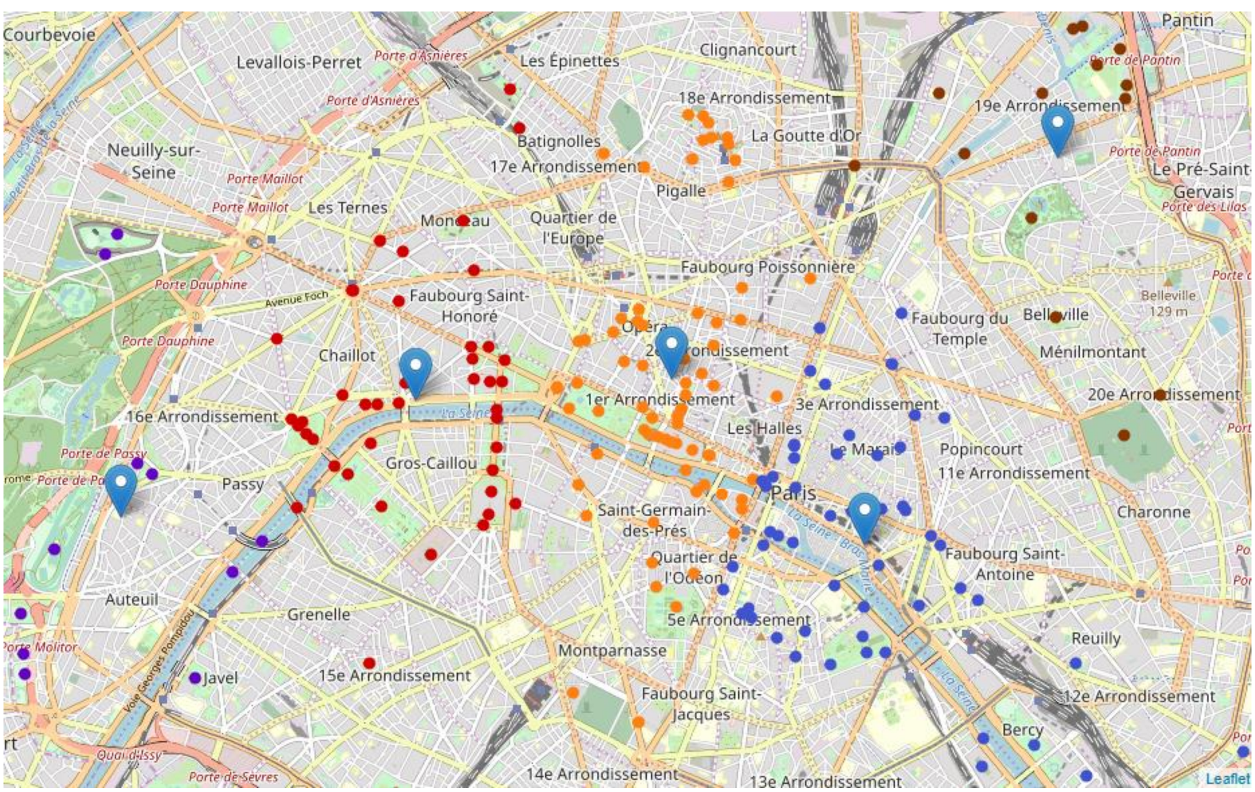
Out[61]:

	id	name	clusterLabel	categories	address	city	country	no_of_likes	rating
0	4bf41231e5eba59334341f90	Place de l'Hôtel de Ville – Esplanade de la Li...	1	Plaza	Place de l'Hôtel de Ville	Paris	France	579.0	9.1
1	4adcda09f964a520e83321e3	Cathédrale Notre-Dame de Paris	1	Church	6 place du parvis Notre-Dame	Paris	France	8518.0	9.4
2	4b5c7d1ff964a5205f3229e3	Tour Saint-Jacques	1	Historic Site	88 rue de Rivoli	Paris	France	272.0	8.7
3	4adcda0af964a520623421e3	Centre Pompidou – Musée National d'Art Moderne	1	Art Museum	Place Georges Pompidou	Paris	France	5325.0	9.2
4	4adcda0af964a520353421e3	Sainte-Chapelle	1	Church	8 boulevard du Palais	Paris	France	585.0	9.1
6	4cca7e73c4d06dcbb72d6303	Fontaine Stravinsky	1	Fountain	Place Stravinsky	Paris	France	103.0	8.5
7	4b8ebd9df964a5203c3433e3	Théâtre du Châtelet	1	Theater	1 place du Châtelet	Paris	France	228.0	8.6
8	4bae535af964a520f5a23be3	Maison Européenne de la Photographie	0	Art Museum	5 rue de Fourcy	Paris	France	358.0	9.0
9	4adcda15f964a520a13721e3	Théâtre de la Ville	1	Theater	2 place du Châtelet	Paris	France	76.0	8.4
10	4c00046cc30a2d7fbeda111d	Parvis Notre-Dame — Place Jean-Paul II	1	Plaza	Parvis Notre-Dame	Paris	France	223.0	8.5

4. Results

4.1. Visualize the five clusters

5 clusters of venues of interest resulting from the K-Means clustering algorithm are plotted to the map below. Each cluster is plotted with a dedicated color. Centroids the five clusters are also visualized as place markers on the map. These centroids are the best 5 best locations to set up e-Scooter stations we are looking for.



4.2. Reverse geocoding centroid locations

Coordinates of the 5 identified centroids are as follows:

```
centroids = kmeans.cluster_centers_
centroids
```

```
array([[48.85022711,  2.36611467],
       [48.86425525,  2.33842997],
       [48.85450928,  2.26530539],
       [48.88482816,  2.3852195 ],
       [48.86413735,  2.30325019]])
```

For a meaningful conclusion, we will apply reverse geocoding to translate these coordinates to physical address. Geocoding service being used is *geocode.farm* (https://geocode.farm/). The final result is the following 5 addresses:

```
print('Recommended locations to set up E-Scooter stations:')

i=0
for c in centroids:
    i+=1
    g = geocoder.geocodefarm(list(c), method='reverse')
    address = g.json['address']
    print('{}. {}'.format(i,address))
```

```
Recommended locations to set up E-Scooter stations:
1. 14 Rue Crillon, 75004 Paris, France
2. 13 Rue de Valois, 75001 Paris, France
3. 18 Rue du Docteur Blanche, 75016 Paris, France
4. 10 Rue André Danjon, 75019 Paris, France
5. 51 Cours Albert 1er, 75008 Paris, France
```

5. Discussion

5.1. Recommendations

As suggested by the outcome of the clustering algorithm, E-Scooter stations should be set up at the following 5 addresses:

```
1. 14 Rue Crillon, 75004 Paris, France
2. 13 Rue de Valois, 75001 Paris, France
3. 18 Rue du Docteur Blanche, 75016 Paris, France
4. 10 Rue André Danjon, 75019 Paris, France
5. 51 Cours Albert 1er, 75008 Paris, France
```

5.2. Observations

Another key observation on the outcome of the clustering algorithm can be obtained by a visual inspection of the cluster map: 3 clusters surrounding Paris center (red, orange and blue) have higher density of tourist attractions. The other 2 clusters (purple and brown) are located farther away from the city center and have sparer venues.

This can be taken into consideration to equip proper types of e-scooters for the stations. For the central 3 stations, convenient e-scooters are more appropriate. Speed is not a priority because venues are located close to each other. For the other 3 stations, faster e-scooters should be equipped as users need to travel a longer distance from one venue to another.

6. Conclusion

In this project, we identified the best locations to set up a series of 5 tourist E-Scooters stations in the central area of Paris. Best locations are based on the distribution of tourist attractions within a radius of 10km from Paris center. The list of venues is collected from FourSquare API and filtered for highly rated venues according to number of user likes and average rating. Data exploration with histograms and box plots was applied to define appropriate thresholds for filtering values. Finally, we apply an unsupervised learning (K-Means clustering) to group the venues into 5 clusters. Coordinates of the clusters' centroids are reversed geocoded to recommend 5 physical addresses in Paris where the E-Scooter stations should be set up.