2. Data preparation

2.1. Approach

This project aims to solve the above problem with the following steps:

1. Collect a list of tourist attractions within a radius of 10km from the city center of Paris using Foursquare venue recommendation API [4]
2. Get number of likes and user rating for each venue in the list by Foursquare venue details API [5]
3. Filter the list to keep only venues with high reviews based on a threshold of likes and ratings.
4. Run a clustering algorithm on the remaining venues. Centroids of these clusters serve as the best locations to set up E-Scooter stations.

2.2. Data collection

2.2.1. Get the list of venues

Following the above approach, 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 the list of venues

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 [6]. 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:

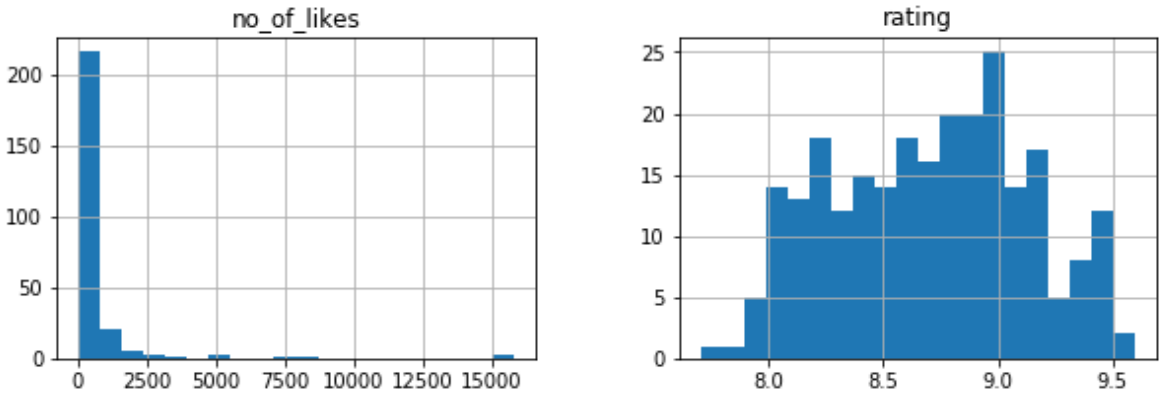
	id	name	categories	address	city	country	distance	no_of_likes	rating
0	4bf41231e5eba59334341f90	Place de l'Hôtel de Ville – Esplanade de la Li...	Plaza	Place de l'Hôtel de Ville	Paris	France	60	578.0	9.1
1	4adcda09f964a520e83321e3	Cathédrale Notre-Dame de Paris	Church	6 place du parvis Notre-Dame	Paris	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. Or in other words, all 250 venues have user likes and rating.

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

Out[40]: False

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

Out[41]: False
```

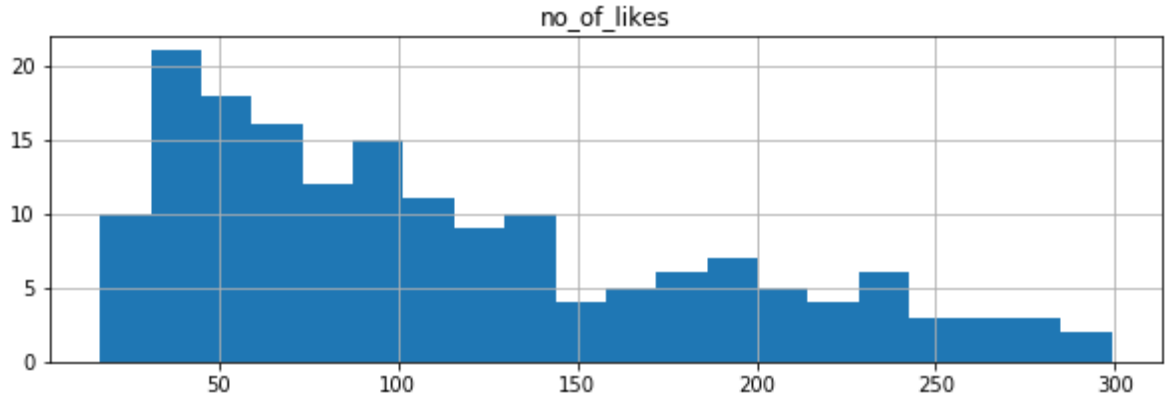
2.2.3. Define thresholds to filter the list of venues

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.

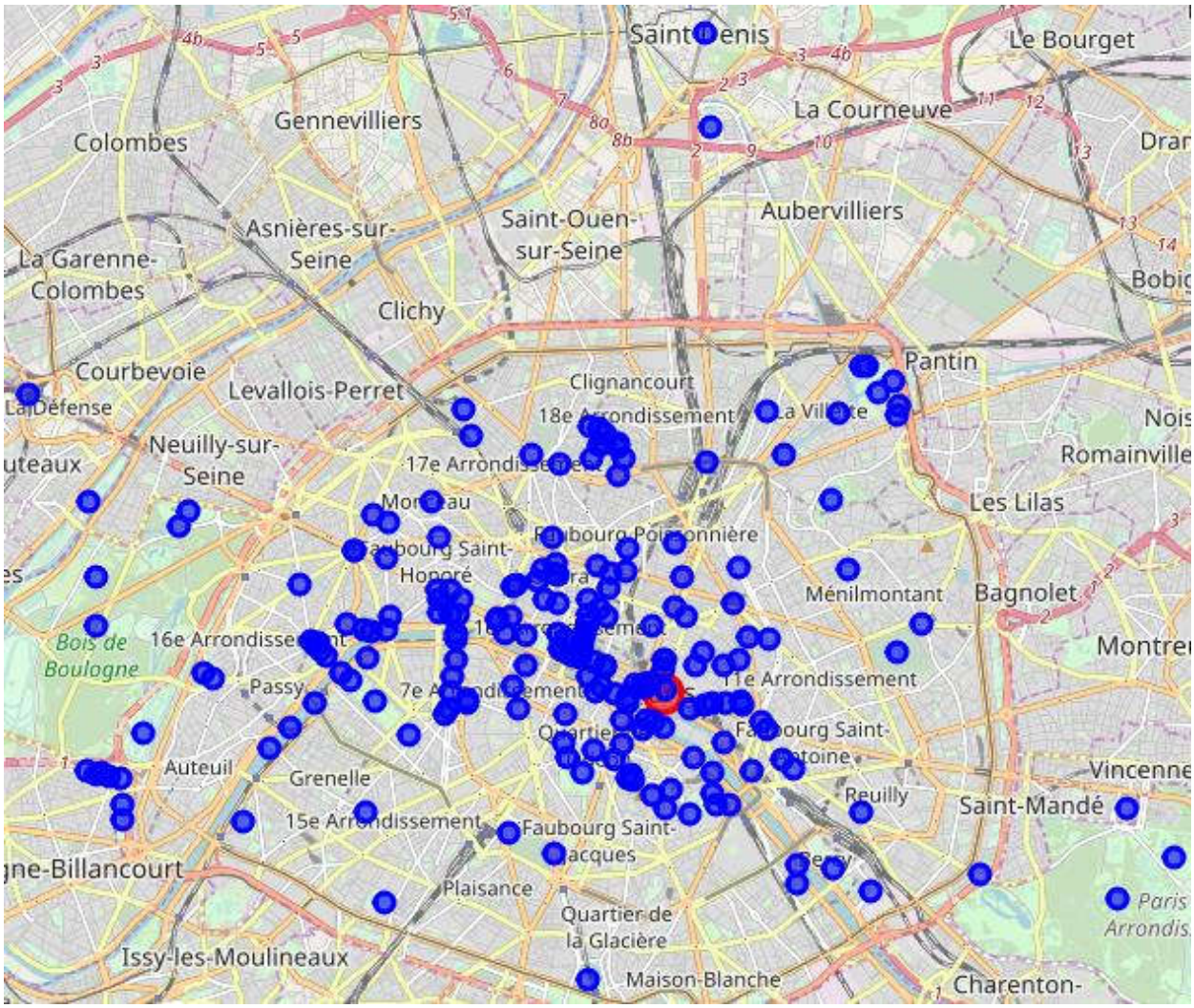
For *number of likes*, the plot is not that clear. It gives a rough idea that some particular venues (most likely famous ones like *Notre-Dame*) receive extraordinary high number of likes. Majority of other less well-known venues receive fewer than approximately 300 likes. Zooming into this range, we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

2.2.4. Plot filtered venues on a map

The dataset is ready for further centroid analysis. Before this, 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***



3. References

[1] https://www.parisdigest.com/paris/paris-facts.htm

[2] https://www.atlasobscura.com/places/the-stravinsky-fountain-paris-france

[3] https://foursquare.com/

[4] https://developer.foursquare.com/docs/api/venues/explore

[5] https://developer.foursquare.com/docs/api/venues/details

[6] https://developer.foursquare.com/docs/api/endpoints