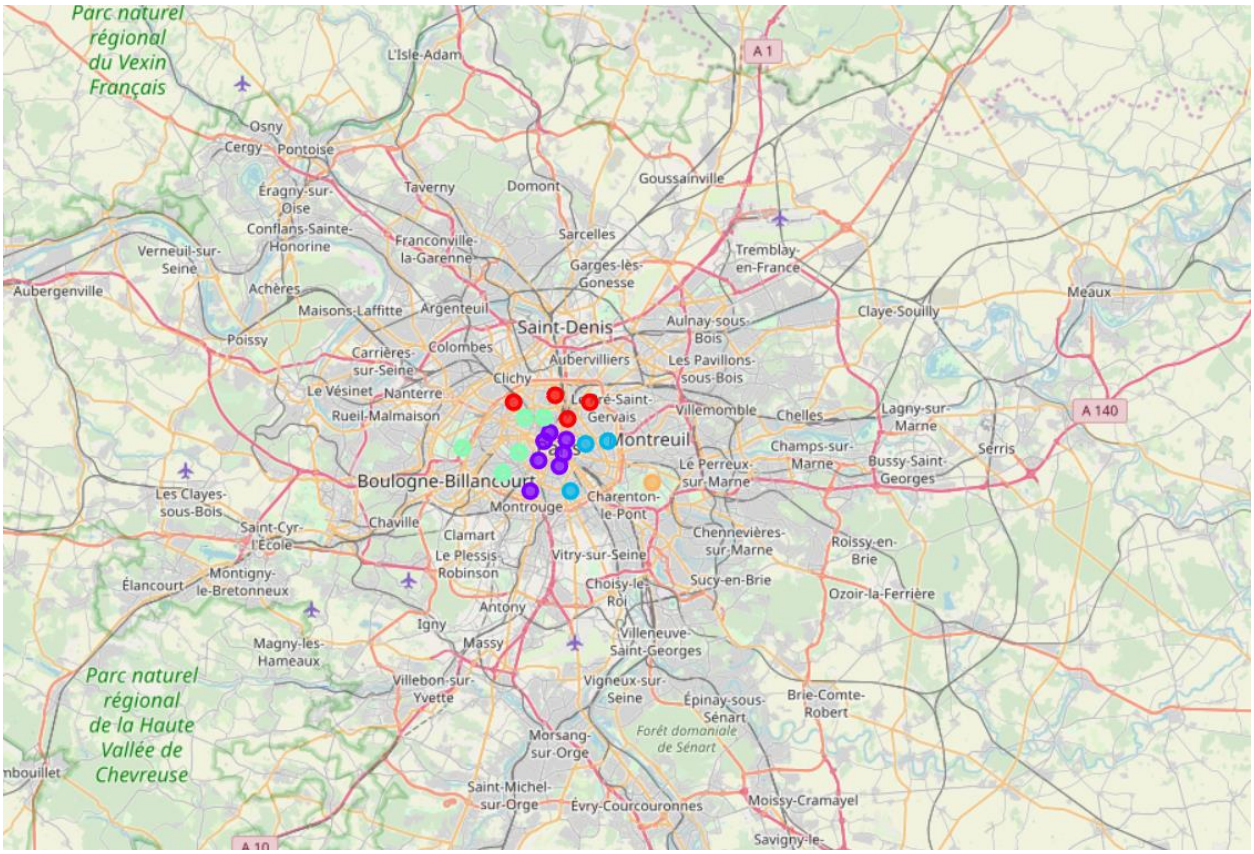
Cluster districts in Paris based on the similarity of art venues using k-means



This is the final project in IBM Data Science course. In this project, I get the Paris district data on Wiki using BeautifulSoup and their respective locations based on Geocode package. After that, I got a list of relevant venues of each district based on the FourSquare API, analyzed the data and use k-means clustering algorithm to cluster all districts into five different groups based on their similarity of venue categories. Finally, I visualized my results on a map using Folium libraries.

I. Introduction:

Paris is the capital of France, including 20 urban districts. There are many works of art, famous museums, and leisure activities for tourists. So every year, many tourists come to Paris to experience the culture as well as fun activities and festivals there. However, because there are so many places, travelers can't get enough information about the location offer which types of art options. Therefore, this project was created to solve that problem.

II. Bussiness Problem:

Business owners, artists or art fans and especially the tourists can find the result of this project useful in their decision-making process to visit Paris.

III. Data:

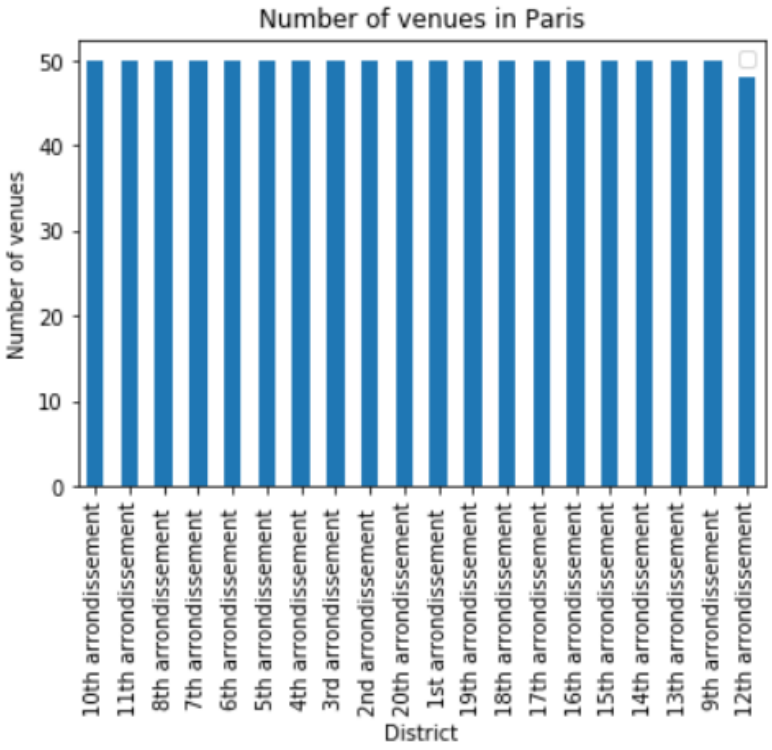
To solve this project's problem, I need to use the data of:

- The list of district names in Paris (scraped from Wikipedia using Beautiful Soup)
- The latitude and longitude of these neighborhoods (Geocoder package)
- Venues data related to the art section in each district to help find the location with more interest in art galleries (from Foursquare API)

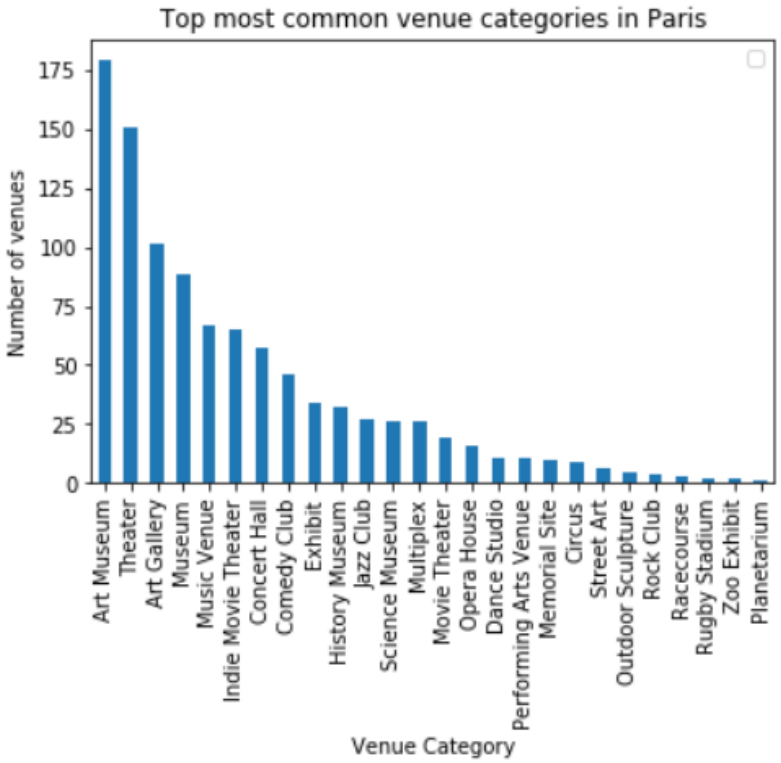
IV. Methodology:

Data from different sources is firstly integrated into a single data frame. Then I check the unique values of the venue categories and found that there are 26 venues categories, which are : Art Museum, Theater, Exhibit, Art Gallery, Museum, Concert Hall, Opera House, Indie Movie Theater, Jazz Club, Movie Theater, History Museum, Comedy Club, Music Venue, Memorial Site, Circus, Dance Studio, Science Museum, Multiplex, Performing Arts Venue, Outdoor Sculpture, Street Art, Rock Club, Zoo Exhibit, Racecourse, Planetarium, Rugby Stadium

Based on the following table, we will see that the number of venues in each district is equal (50), except that district 12 has a slightly smaller number of venues (48).



By observing the table of venue categories, we find that the 3-venue category at most are Art Museum (179), Theater (151), Art Gallery (101).



All categories were one-hot encoded to another data frame for further analysis and modeling. This prevented the count of each category bias towards others since it might be calculated as weights in machine learning models.

	District	Art Gallery	Art Museum	Circus	Comedy Club	Concert Hall	Dance Studio	Exhibit	History Museum	Indie Movie Theater	Jazz Club	Memorial Site	Movie Theater	Multiplex	Museum
0	1st arrondissement	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1st arrondissement	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1st arrondissement	0	0	0	0	0	0	1	0	0	0	0	0	0	0
3	1st arrondissement	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	1st arrondissement	0	1	0	0	0	0	0	0	0	0	0	0	0	0

After that, data were grouped by District and the mean of all counts for each category in a district was calculated.

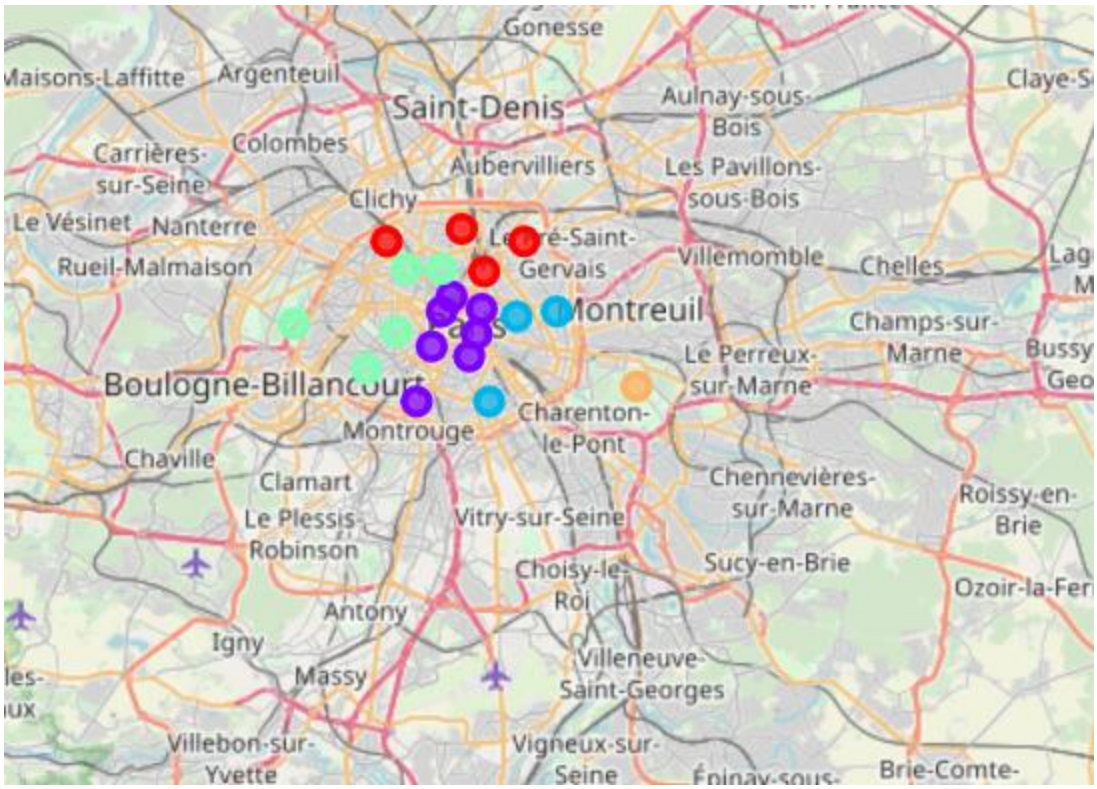
	District	Art Gallery	Art Museum	Circus	Comedy Club	Concert Hall	Dance Studio	Exhibit	History Museum	Indie Movie Theater	Jazz Club	Memorial Site	Movie Theater	Multiplex
0	10th arrondissement	0.1200	0.10	0.02	0.040000	0.080000	0.000000	0.02	0.020000	0.1000	0.04	0.000000	0.000000	0.040000
1	11th arrondissement	0.1200	0.10	0.02	0.060000	0.060000	0.020000	0.02	0.020000	0.0800	0.04	0.020000	0.000000	0.020000
2	12th arrondissement	0.0625	0.00	0.00	0.020833	0.083333	0.083333	0.00	0.020833	0.0625	0.00	0.041667	0.104167	0.041667
3	13th arrondissement	0.2000	0.14	0.00	0.020000	0.000000	0.020000	0.02	0.020000	0.0600	0.04	0.020000	0.000000	0.040000
4	14th arrondissement	0.0800	0.28	0.00	0.040000	0.000000	0.000000	0.06	0.040000	0.1000	0.06	0.000000	0.000000	0.040000
5	15th arrondissement	0.0800	0.32	0.00	0.060000	0.040000	0.000000	0.06	0.020000	0.0200	0.02	0.000000	0.020000	0.020000

Then I show the top 6 most common venues of each district. It represents the selected frequency of venues in each district, for example

	District	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
0	10th arrondissement	Theater	Art Gallery	Art Museum	Indie Movie Theater	Concert Hall	Music Venue	Museum	Comedy Club	Jazz Club	Multiplex
1	11th arrondissement	Art Gallery	Museum	Theater	Art Museum	Music Venue	Indie Movie Theater	Comedy Club	Concert Hall	Science Museum	Jazz Club
2	12th arrondissement	Theater	Music Venue	Movie Theater	Concert Hall	Dance Studio	Art Gallery	Indie Movie Theater	Museum	Memorial Site	Multiplex
3	13th arrondissement	Art Gallery	Museum	Art Museum	Theater	Music Venue	Science Museum	Indie Movie Theater	Jazz Club	Multiplex	Zoo Exhibit
4	14th arrondissement	Art Museum	Theater	Museum	Indie Movie Theater	Art Gallery	Science Museum	Exhibit	Jazz Club	Comedy Club	History Museum
5	15th arrondissement	Art Museum	Theater	Art Gallery	Museum	Comedy Club	Exhibit	Concert Hall	Music Venue	History Museum	Indie Movie Theater
6	16th arrondissement	Art Museum	Theater	Museum	Art Gallery	Music Venue	Movie Theater	Multiplex	Outdoor Sculpture	Racecourse	History Museum
7	17th arrondissement	Theater	Art Museum	Music Venue	Concert Hall	Art Gallery	Comedy Club	Museum	History Museum	Indie Movie Theater	Movie Theater
8	18th arrondissement	Theater	Music Venue	Concert Hall	Indie Movie Theater	Art Museum	Comedy Club	Multiplex	Jazz Club	History Museum	Opera House
9	19th arrondissement	Concert Hall	Theater	Music Venue	Multiplex	Art Museum	Art Gallery	Indie Movie Theater	Jazz Club	Circus	Museum

V. Result:

I set the number of K cluster equal 5 and the result is below:



VI. Conclusion:

This project only focuses on clustering the districts based on their art venues. However, in reality, many other factors can contribute to deciding which district is good to open an art gallery as well. In the scope and time frame of this project, I accept the result and hope to have further research to improve the solution.