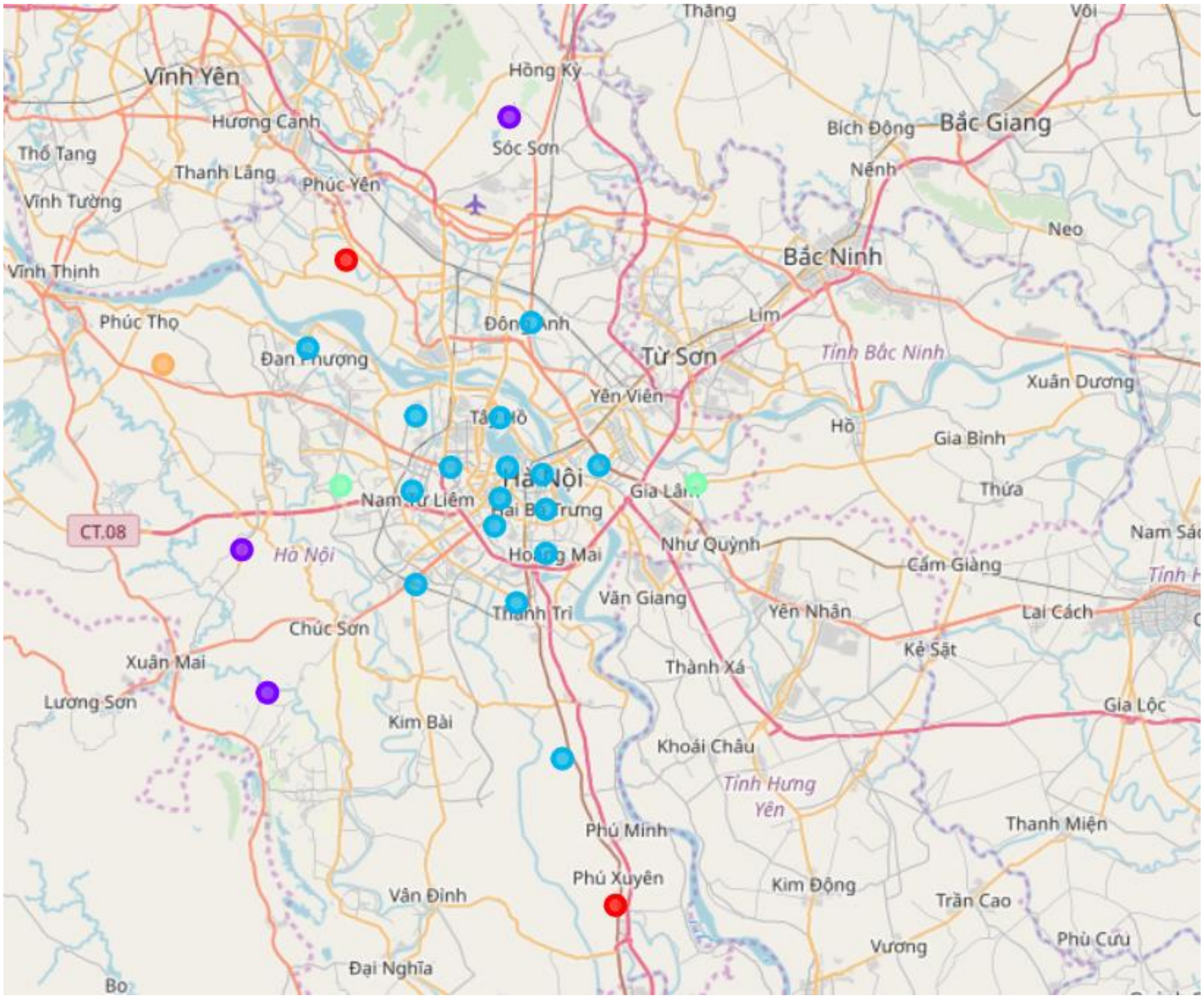
Cluster districts in Hanoi based on the similarity of art venues using k-means



This is the final project in IBM Data Science course. In this project, I get the Hanoi 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:

1. Introduction:

Hanoi is the capital of Vietnam, including 30 urban districts. There are many works of art, famous museums, and leisure activities for tourists. So every year, many tourists come to Hanoi 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.

2. Business problem:

Business owners, artists or art fans can find the result of this project useful in their decision-making process. Business owners can see areas that have many art venues and come up with further decisions. Tourists with great interest in Vietnamese artworks can look at the project’s result to choose their destination for their next trip to Hanoi.

II. Data:

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

- ➢ The list of district name in Hanoi (scraped from Wikipedia using BeautifulSoup)
- ➢ 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 (FourSquare API)

III. 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 19 of them, which are Art Gallery, Museum, Movie Theater, Multiplex, Jazz Club, Music Venue, Opera House, Art Museum, History Museum, Rock Club, Theater, Concert Hall, Memorial Site, Performing Arts Venue, Public Art, Dance Studio, Zoo Exhibit, Amphitheater and Piano Bar.

Based on Table 1 below we can see that the urban districts like Ba Dinh, Hoan Kiem, Tay Ho ... have the highest number of venues (50). Among the districts, Gia Lam, Me Linh, Phu Xuyen, and Phuc Tho have the least art venues. There is only one venue found for each district within 10 kilometers from the district center.

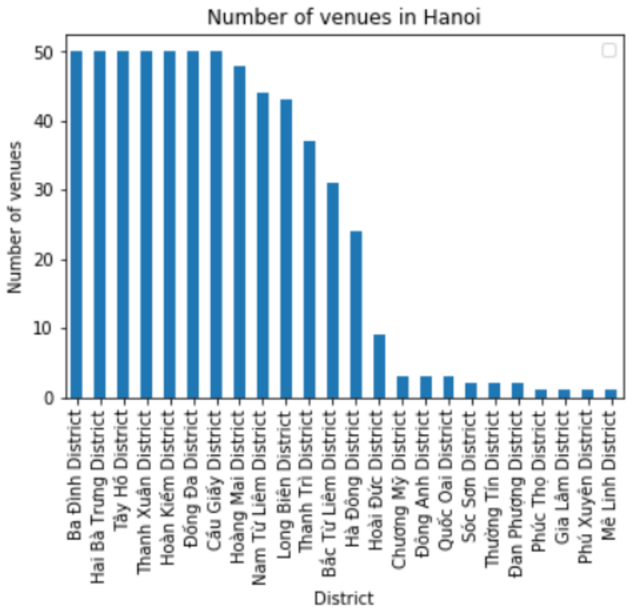


Fig 1: Number of venues in Hanoi

Among the Venue categories, it can be seen that 'Multiplex' has the highest number (155), followed by History Museum (92), Movie Theater (61). Amphitheater is the least venue category with only one (Fig 2):

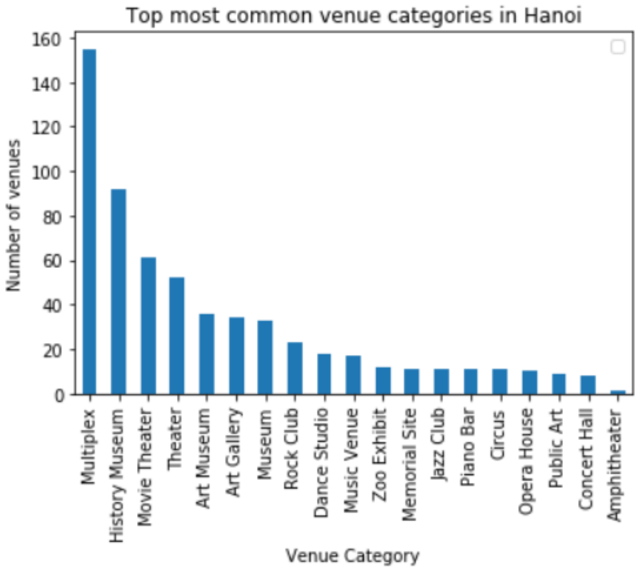
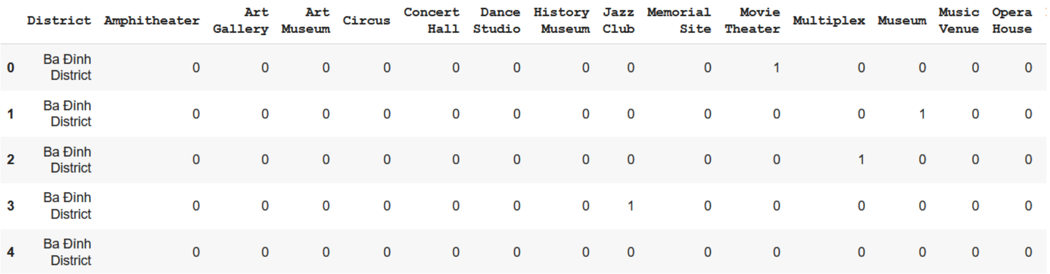


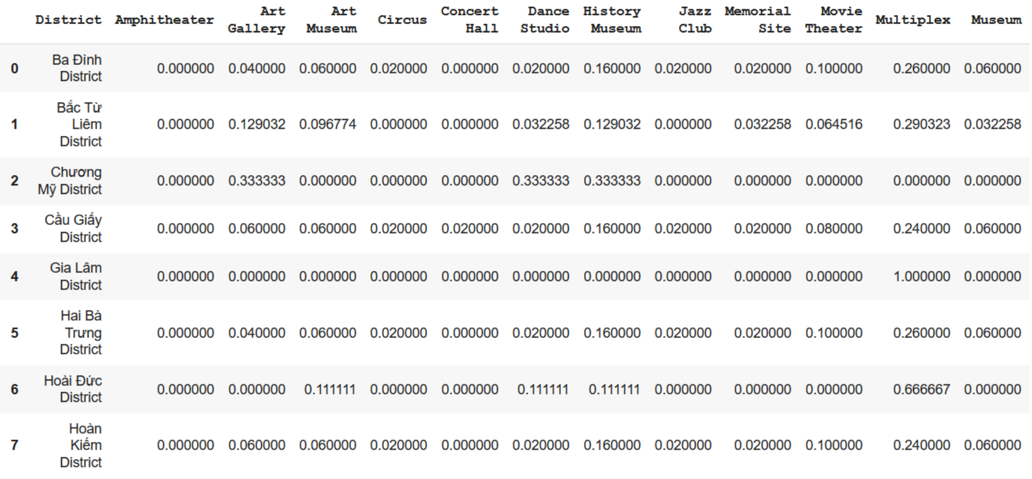
Fig 2: Top most venue category in Hanoi

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.



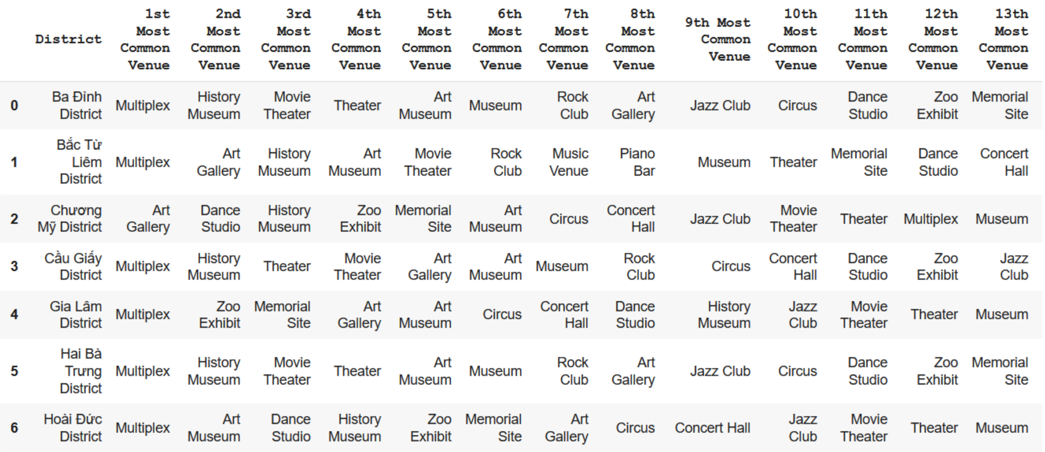
	District	Amphitheater	Art Gallery	Art Museum	Circus	Concert Hall	Dance Studio	History Museum	Jazz Club	Memorial Site	Movie Theater	Multiplex	Museum	Music Venue	Opera House
0	Ba Đình District	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	Ba Đình District	0	0	0	0	0	0	0	0	0	0	0	1	0	0
2	Ba Đình District	0	0	0	0	0	0	0	0	0	0	1	0	0	0
3	Ba Đình District	0	0	0	0	0	0	0	1	0	0	0	0	0	0
4	Ba Đình District	0	0	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	Amphitheater	Art Gallery	Art Museum	Circus	Concert Hall	Dance Studio	History Museum	Jazz Club	Memorial Site	Movie Theater	Multiplex	Museum
0	Ba Đình District	0.000000	0.040000	0.060000	0.020000	0.000000	0.020000	0.160000	0.020000	0.020000	0.100000	0.260000	0.060000
1	Bắc Từ Liêm District	0.000000	0.129032	0.096774	0.000000	0.000000	0.032258	0.129032	0.000000	0.032258	0.064516	0.290323	0.032258
2	Chương Mỹ District	0.000000	0.333333	0.000000	0.000000	0.000000	0.333333	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000
3	Cầu Giấy District	0.000000	0.060000	0.060000	0.020000	0.020000	0.020000	0.160000	0.020000	0.020000	0.080000	0.240000	0.060000
4	Gia Lâm District	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
5	Hai Bà Trưng District	0.000000	0.040000	0.060000	0.020000	0.000000	0.020000	0.160000	0.020000	0.020000	0.100000	0.260000	0.060000
6	Hoài Đức District	0.000000	0.000000	0.111111	0.000000	0.000000	0.111111	0.111111	0.000000	0.000000	0.000000	0.666667	0.000000
7	Hoàn Kiếm District	0.000000	0.060000	0.060000	0.020000	0.000000	0.020000	0.160000	0.020000	0.020000	0.100000	0.240000	0.060000

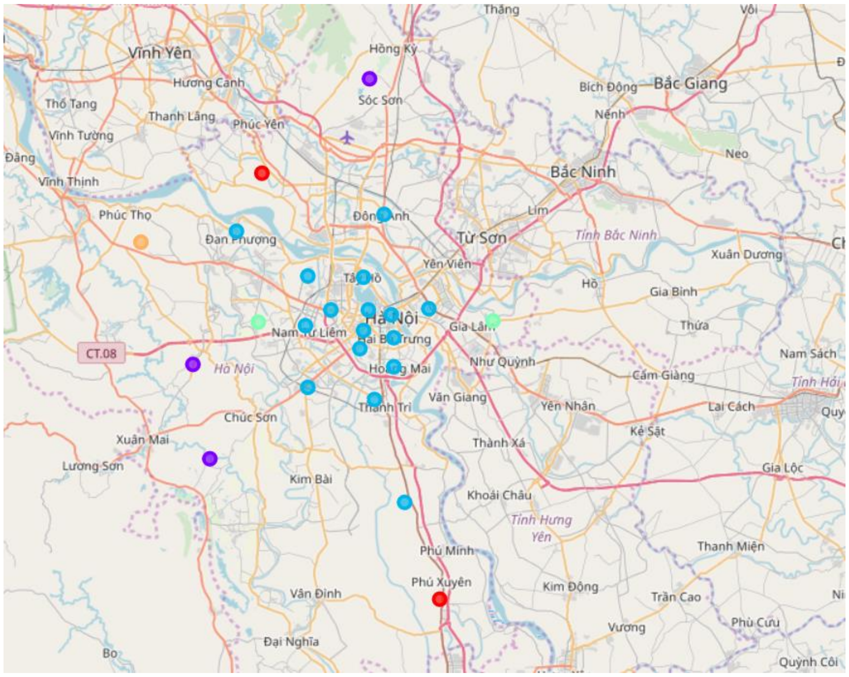
Then I show the top 6 most common venues of each district. It represents the selected frequency of venues in each district, for example:



	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue
0	Ba Đình District	Multiplex	History Museum	Movie Theater	Theater	Art Museum	Museum	Rock Club	Art Gallery	Jazz Club	Circus	Dance Studio	Zoo Exhibit	Memorial Site
1	Bắc Từ Liêm District	Multiplex	Art Gallery	History Museum	Art Museum	Movie Theater	Rock Club	Music Venue	Piano Bar	Museum	Theater	Memorial Site	Dance Studio	Concert Hall
2	Chương Mỹ District	Art Gallery	Dance Studio	History Museum	Zoo Exhibit	Memorial Site	Art Museum	Circus	Concert Hall	Jazz Club	Movie Theater	Theater	Multiplex	Museum
3	Cầu Giấy District	Multiplex	History Museum	Theater	Movie Theater	Art Gallery	Art Museum	Museum	Rock Club	Circus	Concert Hall	Dance Studio	Zoo Exhibit	Jazz Club
4	Gia Lâm District	Multiplex	Zoo Exhibit	Memorial Site	Art Gallery	Art Museum	Circus	Concert Hall	Dance Studio	History Museum	Jazz Club	Movie Theater	Theater	Museum
5	Hai Bà Trưng District	Multiplex	History Museum	Movie Theater	Theater	Art Museum	Museum	Rock Club	Art Gallery	Jazz Club	Circus	Dance Studio	Zoo Exhibit	Memorial Site
6	Hoài Đức District	Multiplex	Art Museum	Dance Studio	History Museum	Zoo Exhibit	Memorial Site	Art Gallery	Circus	Concert Hall	Jazz Club	Movie Theater	Theater	Museum

IV. Result:

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



V. Conclusion:

This project aims to cluster 30 districts of Hanoi into 5 areas based on the districts common in art venues. By scraping district info and coordinates from the internet and using data analysis and machine learning models, the project finds that districts in the city center are more art-oriented with more art galleries, whereas the districts that are near rural areas have fewer options.