Hanoi venues and real estate correlations analysis

Vi Thanh Dat

March 2019

1 Introduction

1.1 Description of the problem and background

Hanoi is the capital of Vietnam, the political, economic, cultural, commercial and tourist center of the country. Currently, It is one of the most fast-growing cities in the world with over 8 million in population and population density of 2,392 people per square kilometer. It is the 3rd most dynamic city in the world according to [1]. I decided to use Hanoi in my project since I myself am a resident of the city.

For such a dynamic city, sometimes it's really hard to choose the area where apartments price is low. And at the same time, geographically convenient. I think this is the problem that many parties interested in such as investors, city's residents and even the government.

My plan in this project is to create a correlation graph between apartments price and venues distribution in each district, and cluster districts then examine the correlation between each clustered and the price range.

1.2 Data preparation

Data used for the project:

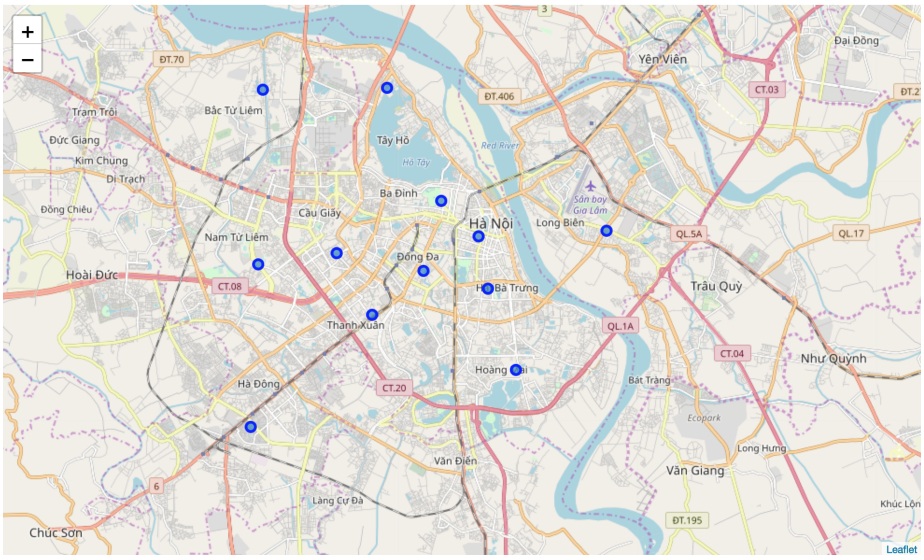
1. Hanoi's 12 urban districts and their coordinates: I scraped the data from [3] and it's coordinate section.
2. I used Foursquare API to get common venues in each districts.
3. I scraped the web [2] to get the price of about 10000 apartments. And then I take the average per square price for each district.

2 Methodology

A row of my original data has the components: district name, area, population, latitude, longitude, apartment rent price and number of apartments for rent of the district.

	District	Area	Population	Latitude	Longitude	Price	Count
0	Ba Đình	9.224	247,100	21.036667	105.836111	252132.671250	1624
1	Bắc Từ Liêm	43.350	333,300	21.074832	105.770597	91049.913941	812
2	Cầu Giấy	12.040	266,800	21.018907	105.797624	212135.645161	406
3	Đống Đa	9.960	420,900	21.012862	105.829642	176190.476190	1218
4	Hai Bà Trưng	10.090	318,000	21.006483	105.853338	151364.555256	812
5	Hà Đông	47.917	319,800	20.959251	105.765959	115615.296807	620
6	Hoàn Kiếm	5.290	160,600	21.024443	105.849847	147143.598834	72
7	Hoàng Mai	41.040	411,500	20.978733	105.863400	65789.473684	406
8	Long Biên	60.380	291,900	21.026478	105.896822	102968.115281	162
9	Nam Từ Liêm	32.270	236,700	21.014968	105.768715	98611.111111	812
10	Tây Hồ	24.000	168,300	21.075463	105.816021	181683.748262	255
11	Thanh Xuân	9.110	285,400	20.997596	105.810656	117857.142857	812

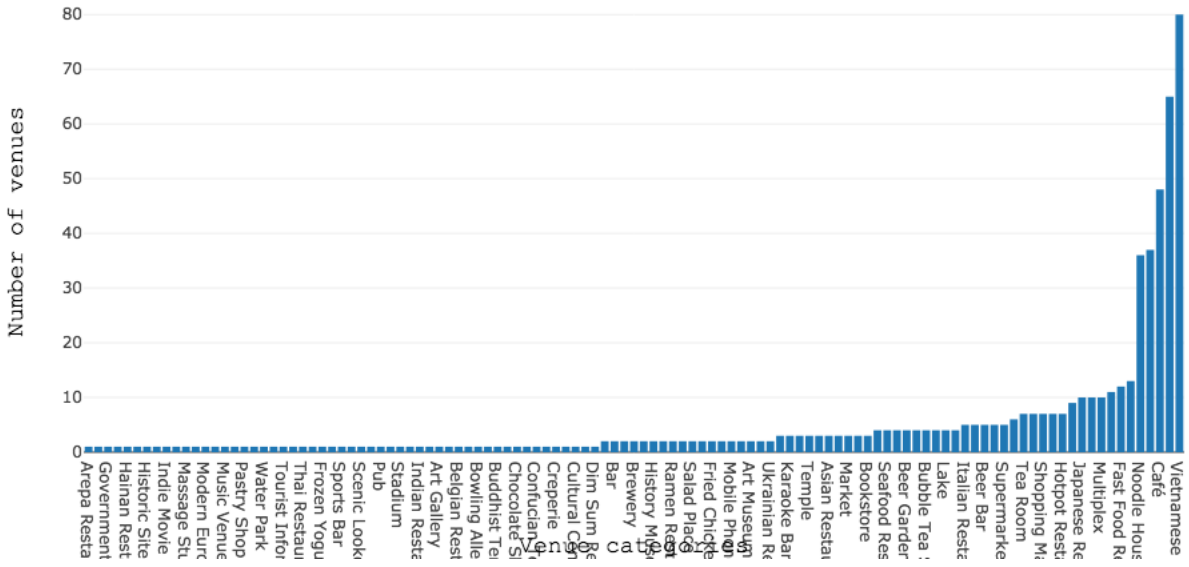
To get a clearer view of the geographical location of the districts of Hanoi, I used python Folium and Hanoi districts coordinates to create a map where each point represent one district of Hanoi.



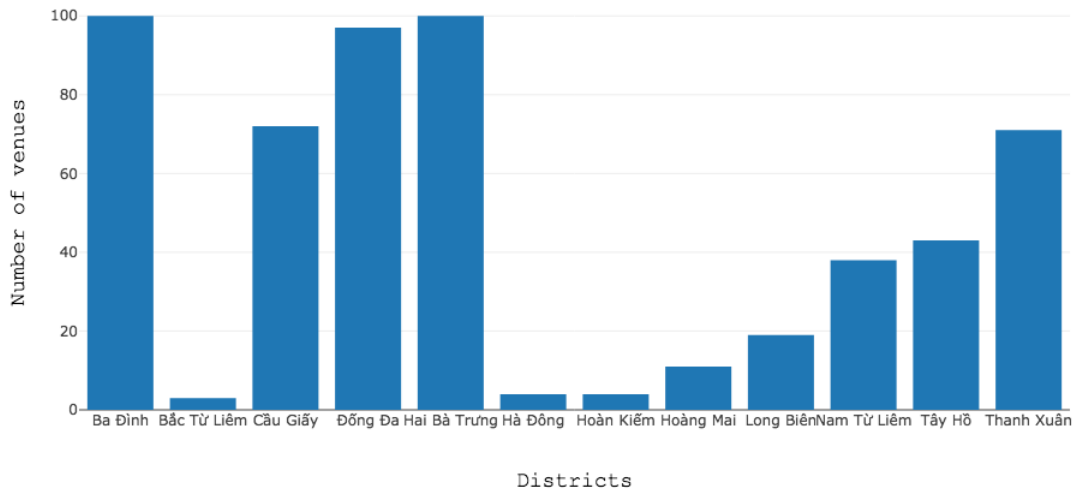
For each district, I used Foursquare API to fetch nearby venues by passing the district's coordinate and parameters like limit: 100 venues, radius: 1500m. Here's how the fetched data look like:

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ba Đình	21.036667	105.836111	Lăng Chủ Tịch Hồ Chí Minh (Ho Chi Minh Mausoleum)	21.035525	105.834720	Monument / Landmark
1	Ba Đình	21.036667	105.836111	Cộng Càphê	21.033504	105.838189	Coffee Shop
2	Ba Đình	21.036667	105.836111	Đền Quán Thánh	21.043024	105.836395	Temple
3	Ba Đình	21.036667	105.836111	Văn Miếu Quốc Tử Giám (Temple of Literature) (...	21.028707	105.836005	Confucian Temple
4	Ba Đình	21.036667	105.836111	Hotel La Siesta Trendy	21.032244	105.845727	Hotel

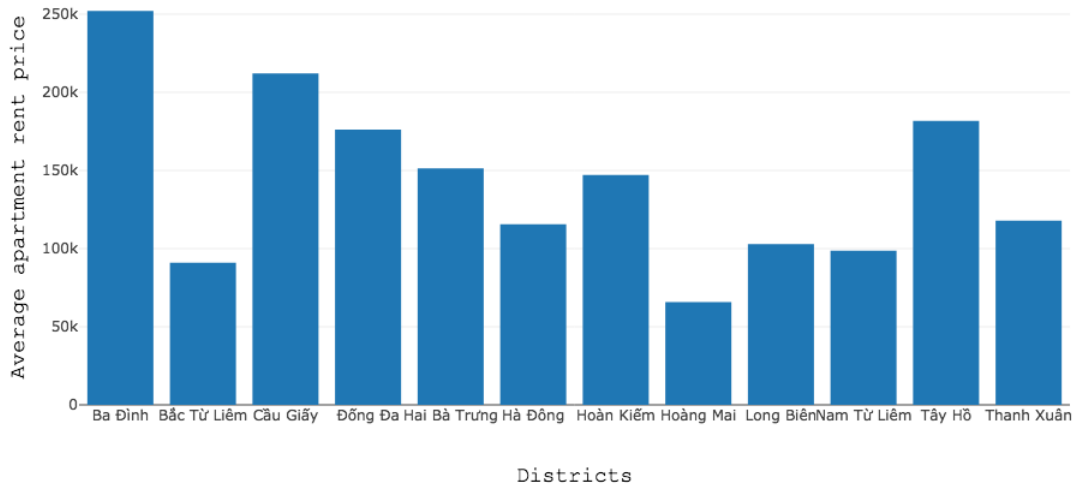
I made some plot to get a closer look at the data I prepared and the data We got from Foursquare API. First, the distribution of about 600 Venues over 113 categories.



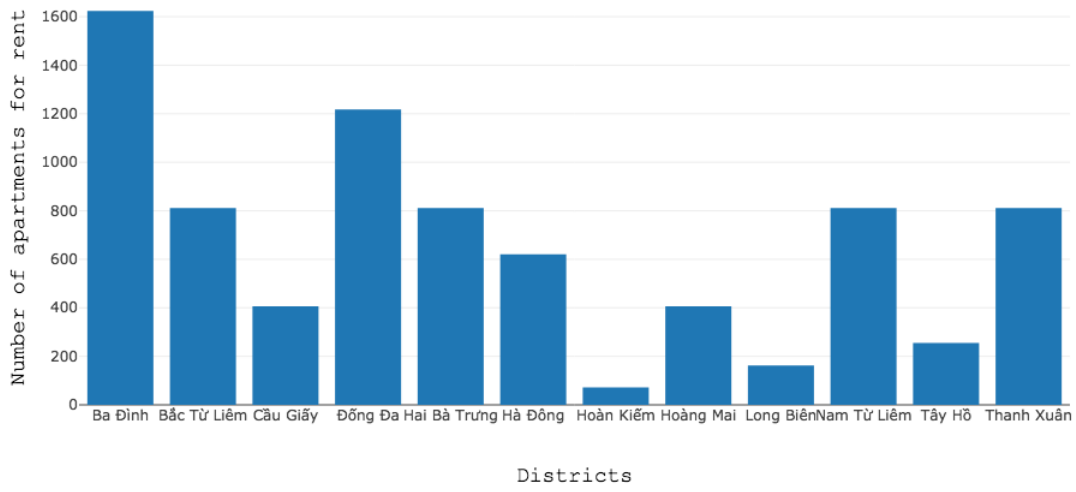
Bar chart shows number of venues of each district.



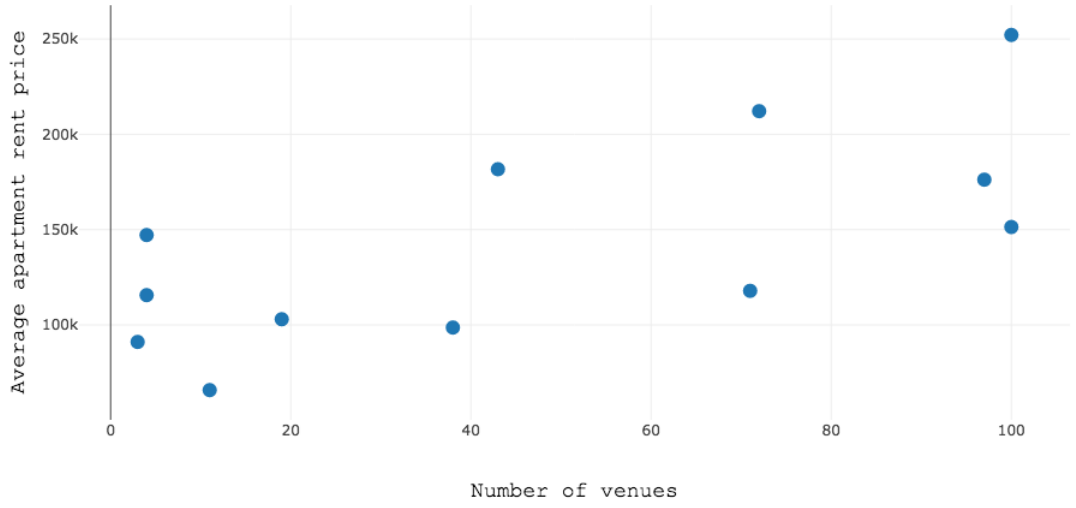
Average apartment rent price of each district.



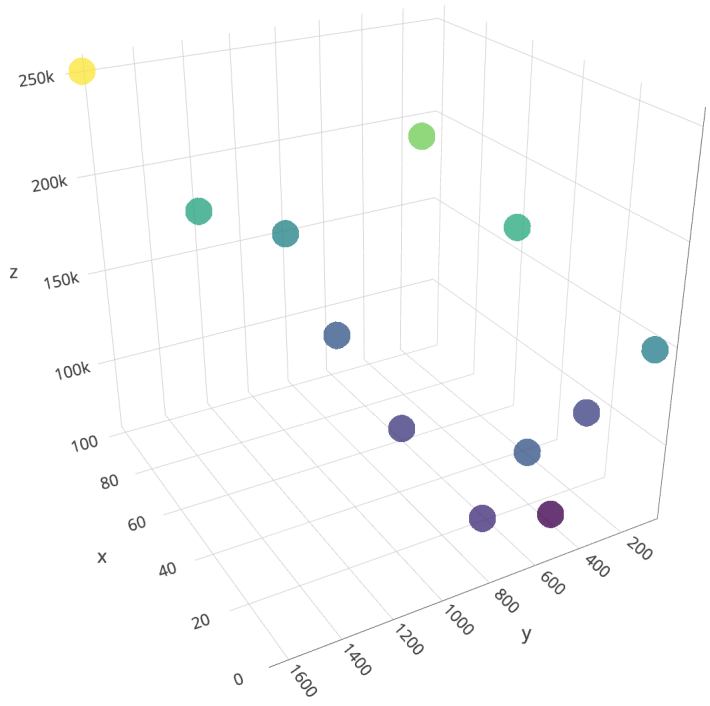
Number of apartment for rent of each district.



2d scatter chart of number of venues and rent. Notice that the relationship between number of venues and average rent is really linear-like.

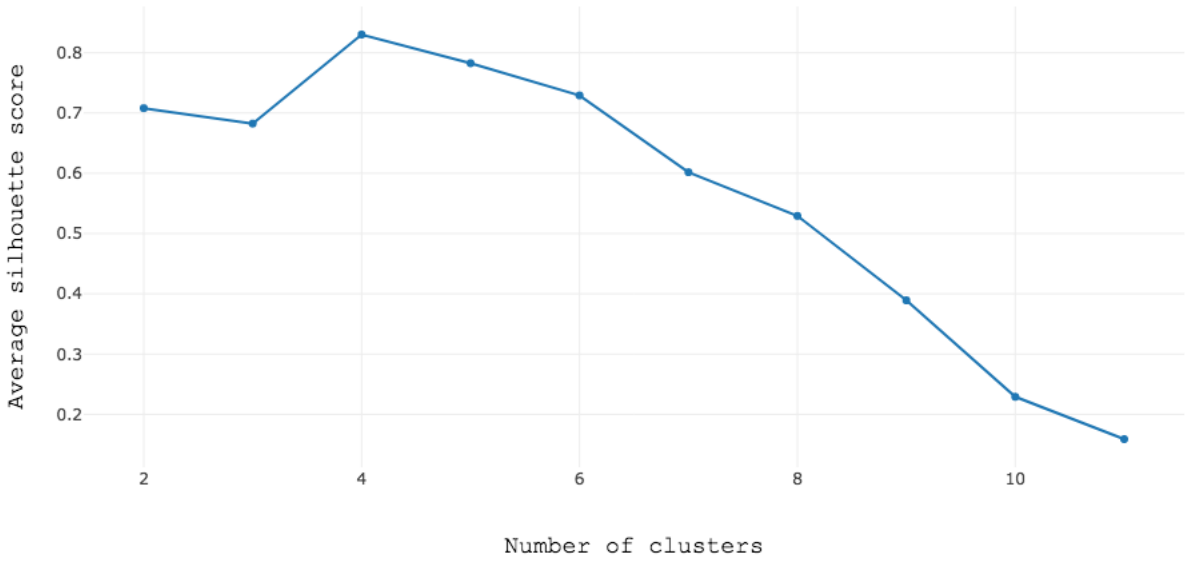


3D scatter chart of all three rent, number of venues, number of apartments for rent.



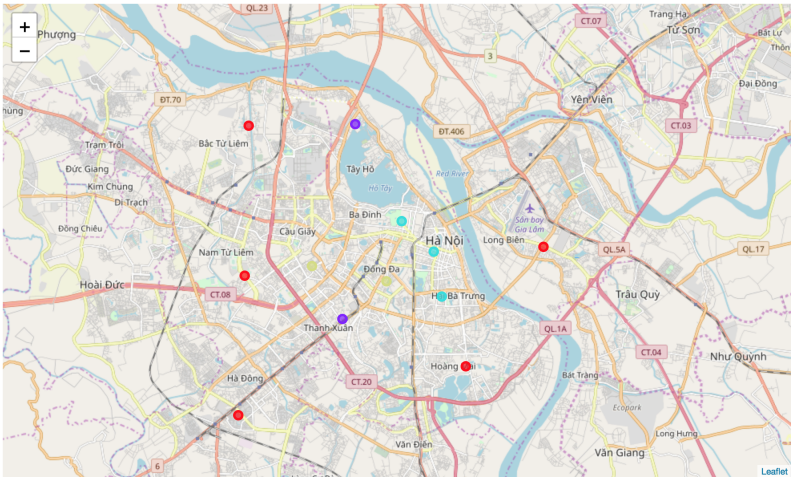
As you can see from the above charts, some districts are really similar in term of price range. I decided to use K-means algorithm to cluster the districts. To choose the number of K, average silhouette score is calculated for each value of K ranging from 2 to 11 since there's only 12 district.

The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. The average silhouette method computes the average silhouette of observations for different values of K. The optimal number of clusters K is the one that maximizes the average silhouette over a range of possible values for K. In our case, the optimal number of clusters is 4. Here's a chart to describe the average silhouette score for each district.

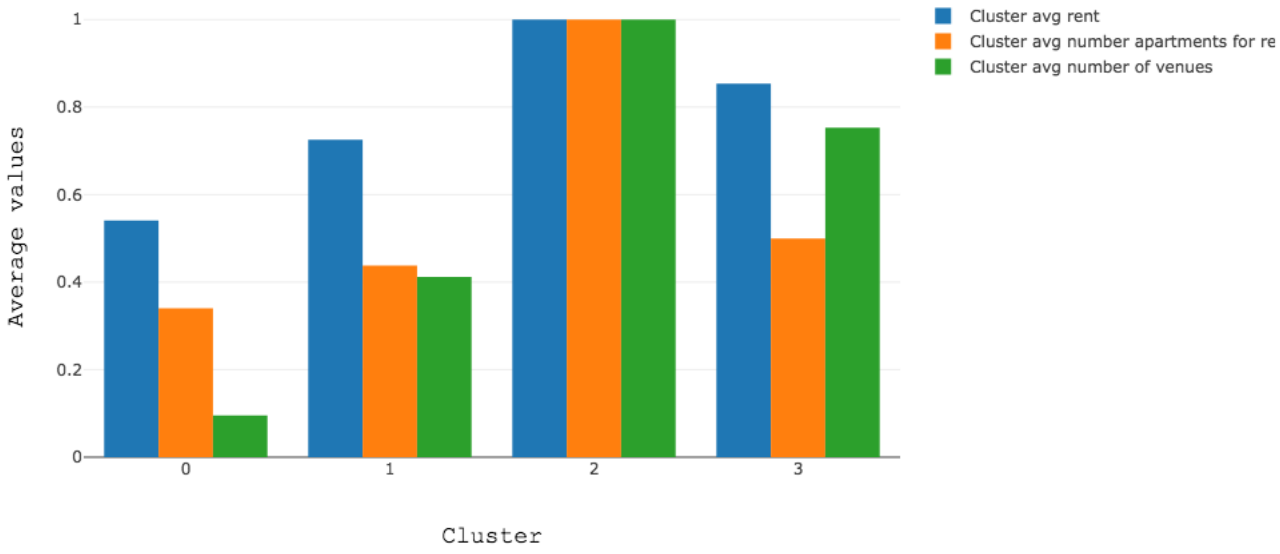


3 Results

Here's how the clustered districts look like on the map of Hanoi.



To summarize the project, I created a grouped bar chart for clusters. The first bar of each cluster represents the cluster average rent. The second one is average number of apartments for rent in cluster. The last one is average number of venues of cluster.



From the bar chart and Folium map, we can conclude that cluster 2 and 3, which locate at the center of the city has the most expensive rent price and also packed with venues. In the other hand, cluster 0 and 1 consist of district those are at the edge of the city where has very low population density and little common venues. Therefore the average rent price of cluster 0 and 1 is lower than cluster 2 and 3.

4 Discussion

Hanoi is a big city with 30 districts. But I only analyze 12 most centered district since the rest are rural districts with large area, low population density and most importantly lack of venue data.

In the project, I used K-means clustering because it is a well-known clustering algorithm and very easy to implement. Besides K-means, We can definitely try other clustering algorithms in the future. For choosing K, I used average silhouette method. I believe for 12 districts, 4 is sufficient for the number of clusters. We can also choose K by other methods like elbow method, gap statistic method, ...

About the dataset, I think my acquired data is pretty small, only 12 district, about 600 venues. So on the way to final results, it is really hard to work on those limited data. Con-

trarily, I have a very large dataset (10000 rows) of apartments which has information about apartment address, rent price, and apartment area. I would probably perform data analysis on that dataset in the near future.

5 Conclusion

With a small dataset and simple data exploration and data clustering, we can easily point out the correlation between venues density and other factors. A better outcome could be achieved if we can access to better and larger sources of data. And not only students, investors and city managers can benefit from similar kind of data analysis.

References

[1] Chinadaily. World's top 10 most dynamic cities, 2019.

[2] https://batdongsan.com.vn. Vietnamese real estate website.

[3] Wikipedia. Hanoi, 2019.