Location and Price Analysis for Housing in Indonesia

Jonathan Tanone

April 2021

1. Introduction

1.1. Background

Indonesia is one of the most populous countries in the world, having a 270 million population. Property demand in Indonesia has been increasing lately. Although the pandemic has halted the demand for a while, many property analysts are optimistic for its growth in 2021[1]. In this era, searching for new property is simple, mainly because of the internet convergence. There is a lot of information out there, but its sheer quantity may overwhelm many Indonesian. Therefore, I am interested in providing analysis to help readers in starting their property search.

1.2. Problem

There is a lot of information presented online, but people cannot transform it into meaningful insights. As a result, people usually start their property search randomly. This project aims to help readers in their search of property by providing a good search starting point.

1.3. Interest

People who want to search for a new property would be interested combine this knowledge with their own, especially if they have any preferences about the venue profile around the area.

2. Data acquisition and processing

2.1. Data Sources

I scraped the data for this analysis is procured through Lamudi[2]. There are several options of other property listings websites, but only Lamudi has the most location variety. To complement this data, I requested the location coordinates using Gecoders and the list of venues near the locations using the foursquare API[3]. I used Geojson data from this Github repository for the map [4].

2.2. Data processing

Through the scraping process from Lamudi, I acquired the location address and the price of the property. Below is the table preview.

	Location	Price	Area
0	JlCisarantenArcamanikCisarantenKulon,Bandung	Rp550.000.000	Bandung
1	TanahKusir,JakartaSelatan	Rp5.500.000.000	JakartaSelatan
2	Cilandak,JakartaSelatan	Rp3.900.000.000	JakartaSelatan
3	Indonesia,JakartaSelatan,JalanBangkaIIIBangka,...	Rp5.500.000.000	JakartaSelatan
4	Serpong,TangerangSelatan	Rp700.000.000	TangerangSelatan

Unfortunately, the location address format is very inconsistent. Some of the listings provide the detailed location, and some of them only display the city name. I decided to group the listings based on the city and average the listings price for that city.

After grouping them, I realized that there are cities that have only one or two listings. In my opinion, the number is not sufficient to create valuable information. I decided to drop the areas that have fewer than three listings. After appending and visualizing the data points using Folium, I noticed that a few of the coordinates are false. I manually corrected it. Below is the finished group table.

	Location	Listings	Latitude	Longitude	Mean_Price
0	Tangerang	402	-6.176031	106.638447	4.545191e+09
1	Jakarta Selatan	330	-6.283818	106.804863	9.785526e+09
2	Bekasi	259	-6.234986	106.994544	5.438207e+09
3	Bandung	240	-6.934469	107.604954	2.390222e+09
4	Bogor	229	-6.596299	106.797242	4.291715e+09
...	...	...	...	...	...
62	Ciamis	3	-7.326661	108.353095	3.006500e+10
63	Madiun	3	-7.611888	111.673193	6.186667e+08
64	Lamongan	3	-7.122912	112.328216	2.275000e+08
65	Cilegon	3	-6.017389	106.053769	3.258333e+08
66	Maros	3	-4.965502	119.692843	1.700000e+08

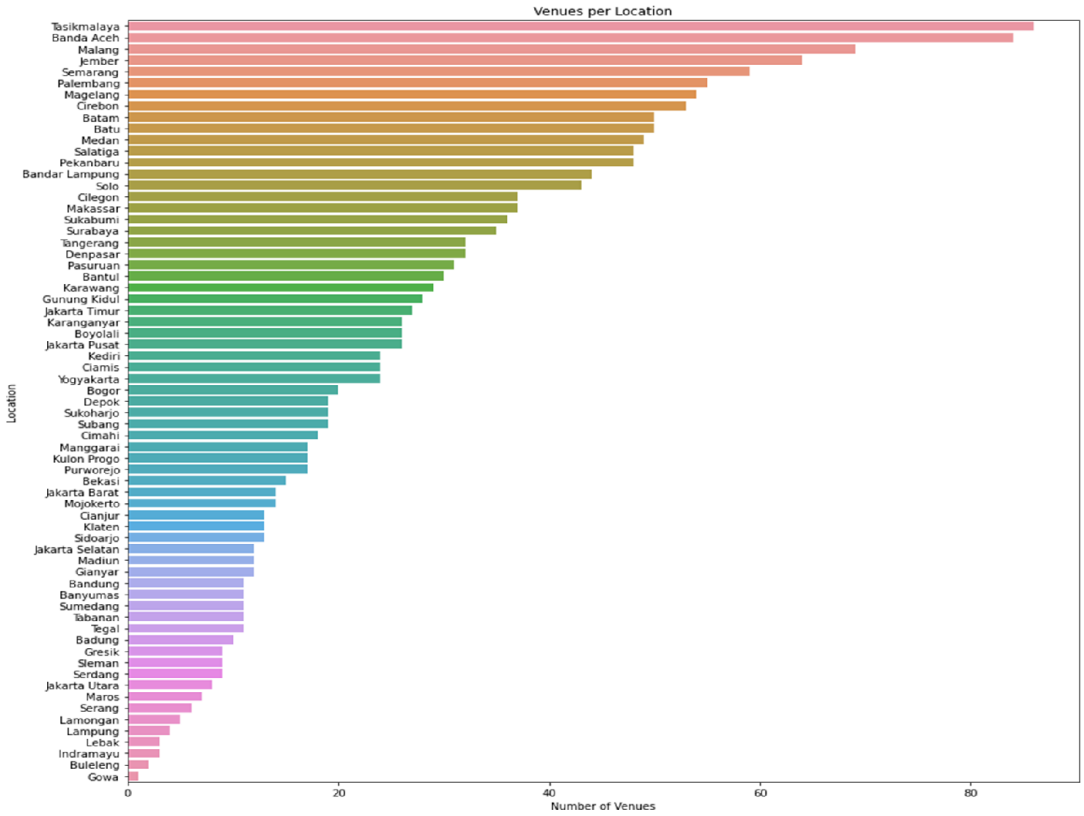
67 rows × 5 columns

The next step is requesting the coordinates using geocoders and the venues list using the foursquare API. For the venues, I plan to list 100 venues near the location coordinates. I collected only the category and the number of venues. The table above transformed into the table below.

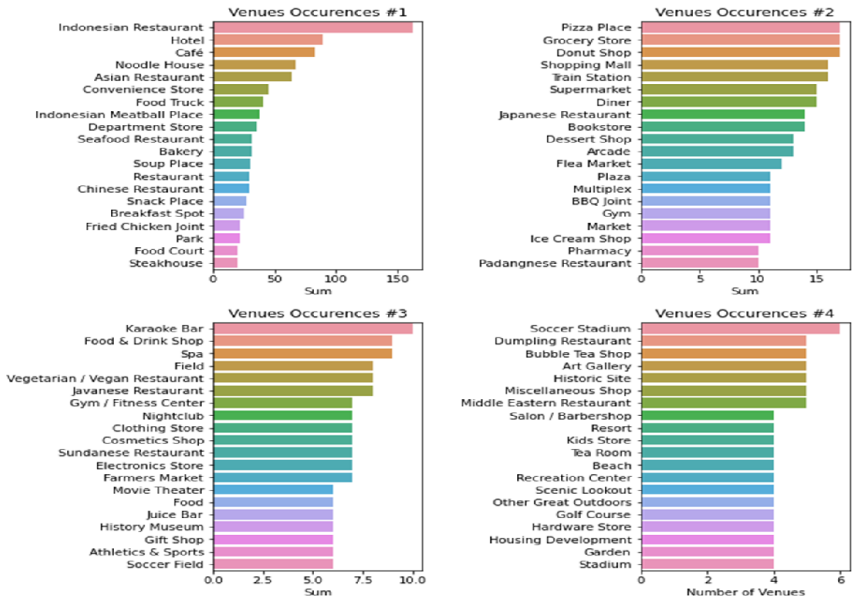
	Unnamed: 0	Unnamed: 0.1	key_0	Location	Listings	Latitude	Longitude	Mean_Price	Fast Food Restaurant	Coffee Shop	Café	Noodle House	Donut Shop	Indonesian Meatball Place	Bookstore	Soc Plac
0	0	0	Tangerang	Tangerang	402	-6.176031	106.638447	4.545191e+09	2.0	2.0	3.0	1.0	1.0	1.0	1.0	1
1	1	1	Jakarta Selatan	Jakarta Selatan	330	-6.283818	106.804863	9.785526e+09	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1
2	2	2	Bekasi	Bekasi	259	-6.234986	106.994544	5.438207e+09	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	3	3	Bandung	Bandung	240	-6.934469	107.604954	2.390222e+09	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0
4	4	4	Bogor	Bogor	229	-6.596299	106.797242	4.291715e+09	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0

5 rows × 218 columns

To make it easier to understand, I created a figure to visualize the table. In the figure below, the total number of venues per location the API requested is not uniform. The number ranges from two to ninety. I believe that clustering the data points first will yield the best information quality. Furthermore, there is a lot of venue categories collected in this data.



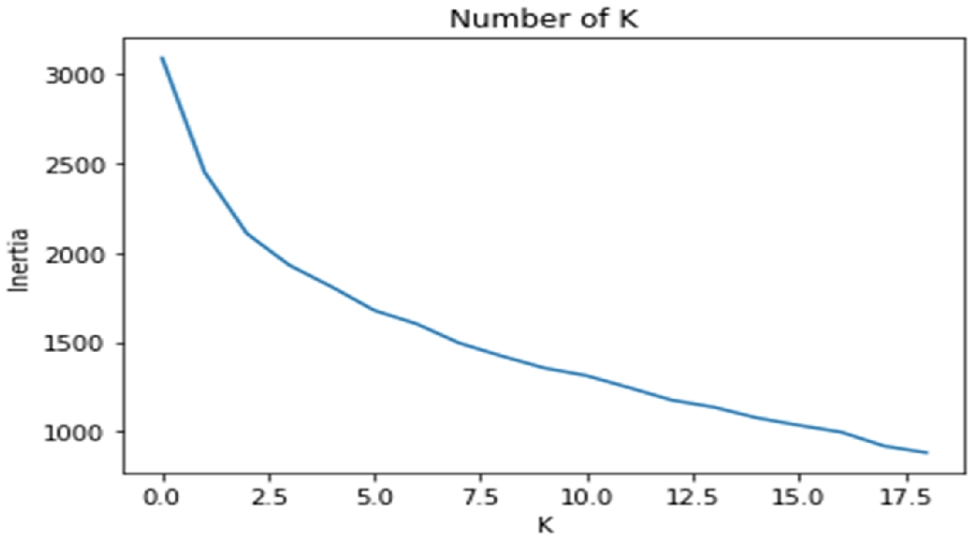
Ranking the venue occurrence will provide the most informative result. In the venue category occurrence graph below, I divided the data into bins with 20 members each. There are around 200 venue categories recorded, but venues in “Venue Occurrences #4" have small numbers. Thus, venues ranked lower than those in "Venue Occurrences #4" will not contribute much to the cluster profile. On the other hand, those in Figures #1 to #3 will contribute to the cluster profile significantly.



3. Methodology

3.1.Clustering

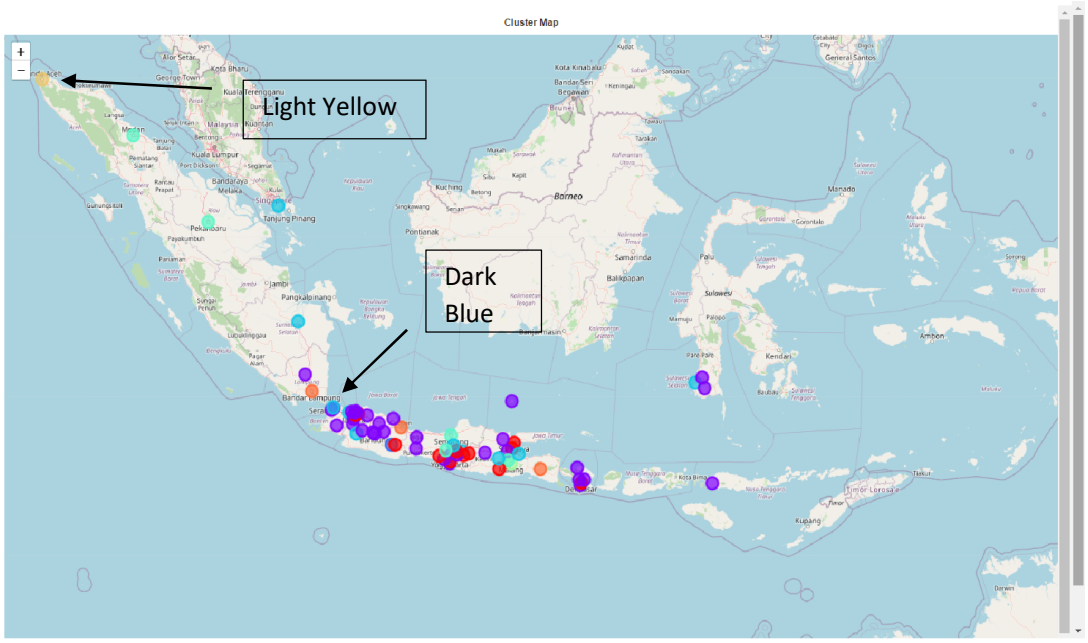
I cluster the data into several clusters using the K-Means clustering algorithm. Variables that I used for the clustering are the venue number per category. The goal is to provide the expected general venue profile t when buying properties in those areas. Using the elbow method, I visualize the inertia and choose the best number of K. Below is the visualization.



The inertia has a steep curve until 7.5. Thus, I decided to divide the data into 8 clusters. The resulting clusters have a weird distribution.

	Cluster	Number of Members	Color
1	0	34	Purple
6	1	1	Dark Blue
0	2	10	Light Blue
5	3	5	Cyan
3	4	1	Light Green
7	5	1	Dark Yellow
4	6	3	Orange
2	7	12	Red

Interestingly, only three of the clusters have ten or more members. Plotting the data to the map will help in determining the quality of the information provided. Below is the visualization of the cluster in the map.



As you can see, there is a big part of the map that does not have data points. The data source does not contain any information about those areas. The reason is that the rate of technological penetration and the education quality is not the same for each location. Java is the most developed island in Indonesia[5]. Compared to people from other areas, people in Java are more comfortable using the Internet in their daily life. As a result, the likelihood of online property listings located in Java is much larger than in the rest.

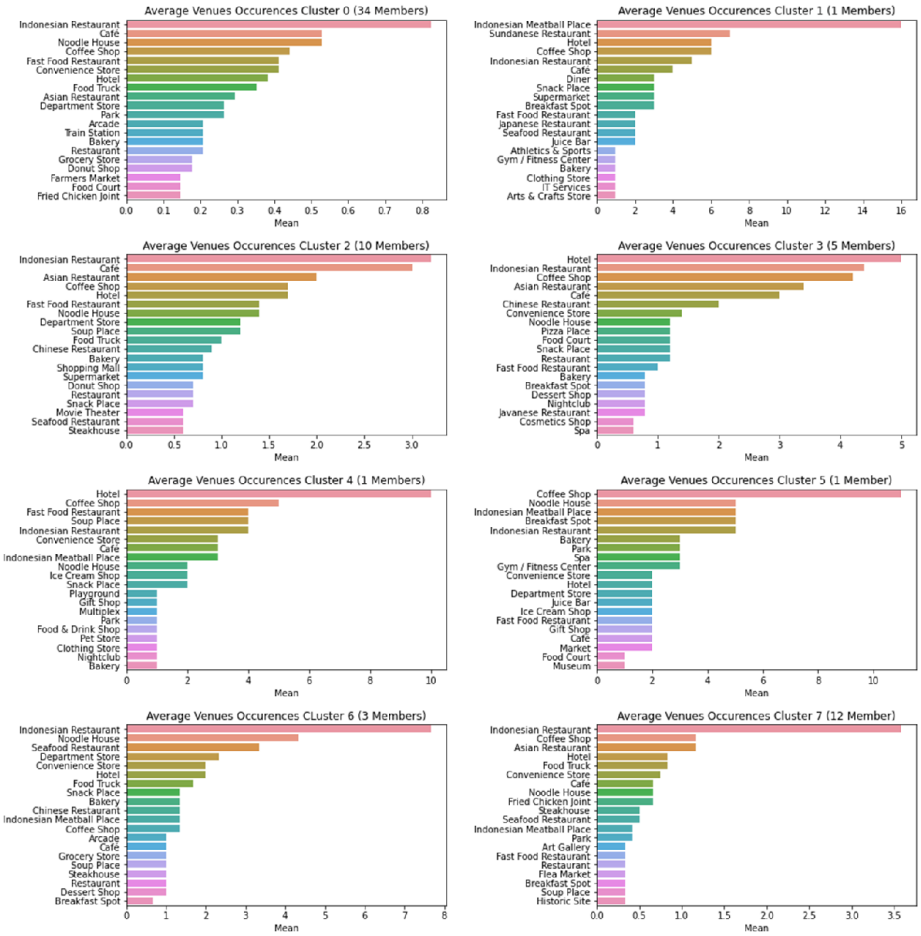
In the Map, Purple cluster members are more dispersed than the other. Also, the Purple cluster has the highest member count. It signifies distribution that surpasses geographical boundaries. On the other hand, the Red cluster aggregated at Java. Other clusters, such as Cyan and Light Blue, are distributed throughout Java and Sumatra. And the rest of the cluster only consists of one or two cities.

It is important to note that clusters that have only a few members are not in itself unique. Several factors need to be assessed before making the classification. For example, Cluster Light Yellow is a cluster with one member because there is a lack of data points on its surrounding. On the other hand, Cluster Dark Blue is surrounded by Purple Cluster, making it a unique cluster.

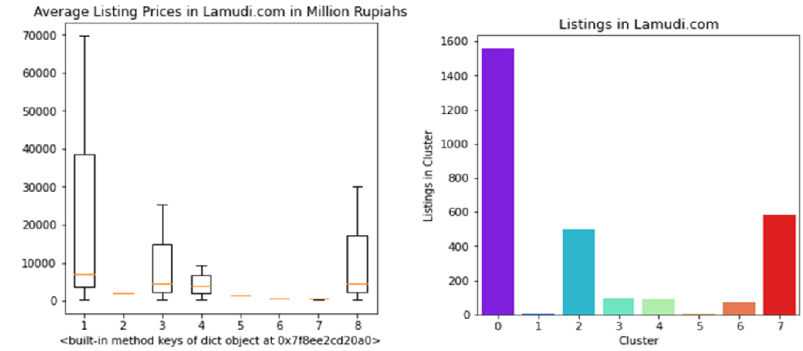
4. Discussion

4.1. Cluster Data Analysis

After the clustering process, we can start visualizing the venue profile of each cluster. These venue profiles will act as a guide for people when purchasing a property. Below is the graph of the venue ranking.



The graph represents the average venue count for each cluster in descending order. I displayed the top twenty venues within the cluster. While it is true that the top venues for each cluster are similar, the rest of the venues are different. These differences will be useful when people want to search for a place based on their venue preference.

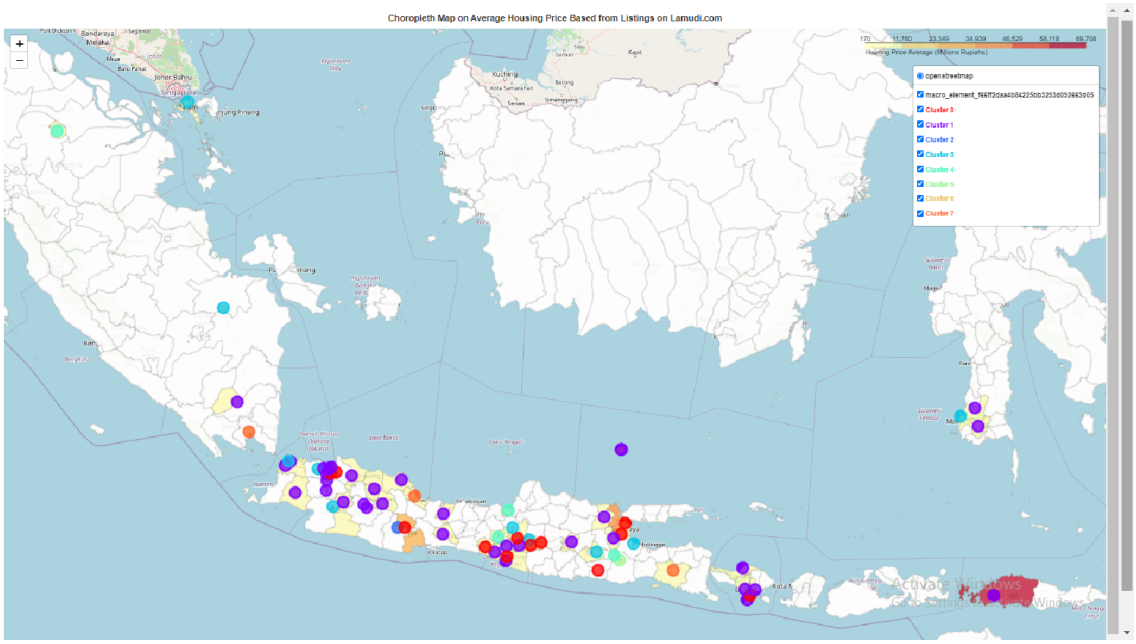


To complement the graph above, I also created the listings quantity and the price distribution for each cluster. In the figure above, the price range for some clusters is significant considering the average listing prices. It is normal because the cluster is based on venue similarity and not based on the listing price.

To make the information easier to understand, I created a choropleth map that displays the average listing price of the location with the cluster points. Below is the choropleth map.

The choropleth map contains the information of:

1. Average property listing price
2. Location cluster
3. Cluster Name



Using the choropleth map, readers can easily retrieve the cluster and the price of each location. By referring to the figure in the discussion above, readers can get informative data of the venues. It will create a good starting point for readers to start their search for property in Indonesia.

5. Conclusion

This study has succeeded in producing the venue profile for each location cluster in Indonesia. However, there are still some issues that need to be tackled in future studies. For one, increasing the number of data points in locations outside Java will benefit the analysis significantly. It will enable the researcher to explain a unique cluster and spot a pattern in the venue profiles based on the location. Currently, it is hard to procure data outside Java because the lack of internet proficiency of people outside Java makes those people avoid online transactions.

6. References

[1]https://tekno.sindonews.com/read/260144/207/prediksi-kebutuhan-properti-2021-versi-rumahcom-1607357560/15

[2]https://www.lamudi.com

[3]https://developers.foursquare.com

[4]https://raw.githubusercontent.com/rifani/geojson-political-indonesia/master/IDN_adm_2_kabkota.json

[5]https://www.thejakartapost.com/news/2020/11/11/indonesian-internet-users-hit-196-million-still-concentrated-in-java-apjii-survey.html