The Battle of Neighborhood: Open a Cinema in Surabaya

Ishardina Cholifatul Hidayati

December 22, 2020

1. Introduction

Surabaya is the capital of the Indonesian province of East Java. Located on the northeastern border of Java island, on the Madura Strait, it is one of the earliest port cities in Southeast Asia. According to the National Development Planning Agency, Surabaya is one of the four main central cities of Indonesia, alongside Jakarta, Medan, and Makassar. The city has a population of 2.89 million within its city limits in 2019 and 9.5 million in the extended Surabaya metropolitan area, making it the second-largest metropolitan area in Indonesia.

There are many movie theaters in Surabaya, I will conclude where are the existing movie theaters. Then, I will use the clustering model to find similar areas in the city considering demographic data of each borough. The preferred area shall be minimum from existing movie theaters.

I will use data science tools to fetch the raw data, visualize it then generate a few most promising areas based on the above criteria. In the meanwhile, I will also explain the advantage and traits for the candidates, so that stakeholders can make the final decision base on the analysis.

2. Data

Based on the definition of our problem, factors that may impact our decision are:

- Demographic information, e.g. population, density, education, health, area.
- Number of existing shopping malls in the neighborhood and nearby
- Number of existing movie theaters in the neighborhood and nearby

In this project, we will fetch or extract data from the following data sources:

- Surabaya census information
- Shopping malls and movie theaters data in every neighborhood will be obtained using Foursquare API
- The coordinate center of Surabaya will be obtained using Positionstack Geocoding API of well known Surabaya location.
- Surabaya borough shapefile is obtained from Open Street Map.

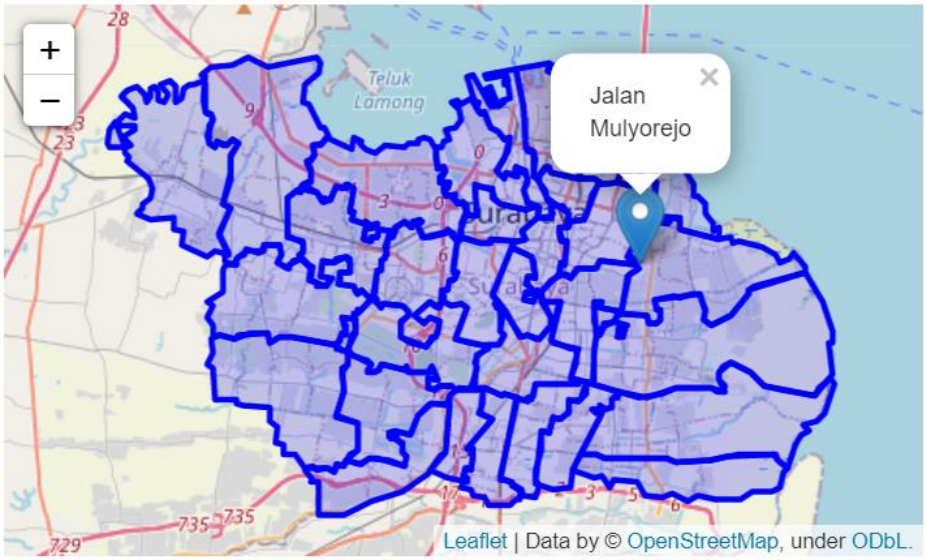
2.1 Census Data of Surabaya

To find areas region of Surabaya, I am looking on the internet and find the resources from BPS (Badan Pusat Statistik) or Statistical Center Data of Surabaya. I collect the data then make it to 1 file. I only focus on several basic census information: Population, Density, Area, Education, and Health. The area unit in the data frame is square meters. Thus, the population density unit becomes people / square meter.

	Borough	Area	Population	Density	Health	Education
0	Tegalsari	4300	85606	19.908372	78.5	94.3
1	Genteng	4100	46548	11.353171	78.1	94.4
2	Bubutan	3900	84465	21.657692	75.8	84.4
3	Simokerto	2600	79319	30.507308	74.7	83.2
4	Pabeancantikan	6800	69423	10.209265	73.0	84.6

2.2 Coordinate of Surabaya Neighborhood

Coordinate of Surabaya neighborhood will be obtained from the geojson file obtained from Open Street Map, while the center of Surabaya obtained using Positionstack Geocoding API of well known Surabaya location.




```
31 center coordinates are generated.
```

	Borough	Latitude	Longitude
0	Genteng	-7.258943	112.744854
1	Simokerto	-7.239466	112.753292
2	Semampir	-7.213868	112.748414
3	Kenjeran	-7.217600	112.768674
4	Bulak	-7.236022	112.788849
5	Krembangan	-7.228817	112.721646

2.3 Foursquare API

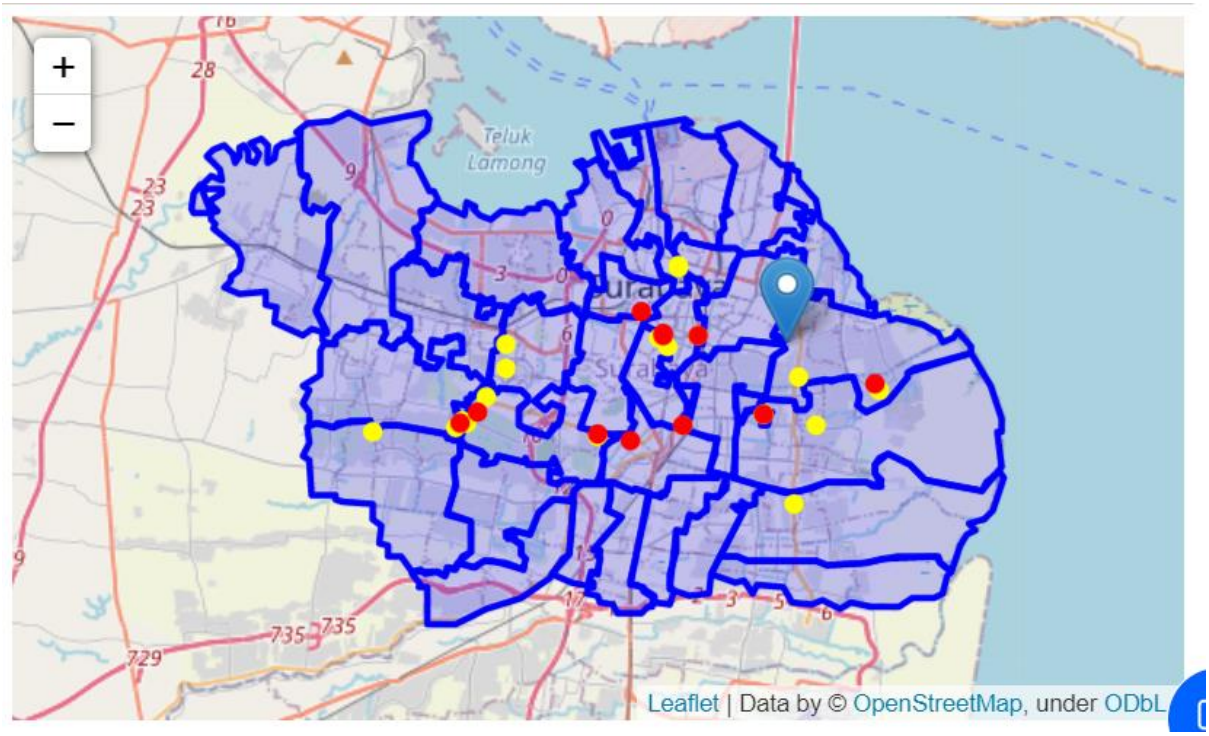
Now I generated all the candidate neighborhoods on Surabaya, I will get all movie theaters and mall information using Foursquare API. Let's fetch all the venue on Surabaya first. To do so, we will fetch movie theaters and shopping mall data in each borough.

```
There are 11 movie theater in Surabaya
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Distance Venue to Neighborhood
100	Genteng	-7.258943	112.744854	The Premiere	-7.261219	112.739538	Multiplex	644.278009
101	Genteng	-7.258943	112.744854	Tunjungan 5 XXI	-7.261345	112.739524	Movie Theater	651.339111
109	Genteng	-7.258943	112.744854	Grand City XXI	-7.261631	112.750340	Multiplex	680.773524
206	Bubutan	-7.248752	112.727749	CGV Cinemas	-7.254282	112.732692	Multiplex	826.926333
1719	Wonokromo	-7.297717	112.737909	Sutos XXI	-7.294064	112.729590	Multiplex	1012.292539

```
There are 23 mall in Surabaya
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Distance Venue to Neighborhood
106	Genteng	-7.258943	112.744854	Tunjungan Plaza 5	-7.261434	112.739563	Shopping Mall	651.531129
156	Genteng	-7.258943	112.744854	Grand City	-7.261586	112.750759	Shopping Mall	720.904447
142	Genteng	-7.258943	112.744854	Tunjungan Plaza	-7.263186	112.739825	Shopping Mall	733.345380
117	Genteng	-7.258943	112.744854	99 Ranch Market	-7.264796	112.741016	Supermarket	780.422172
105	Genteng	-7.258943	112.744854	Tunjungan Plaza 6	-7.262290	112.738271	Shopping Mall	822.879825



Unlike coffee shops, restaurants everywhere, there aren't lots of movie theaters in the region, it also makes sense since we don't expect movie theater in every block.

3. Methodology

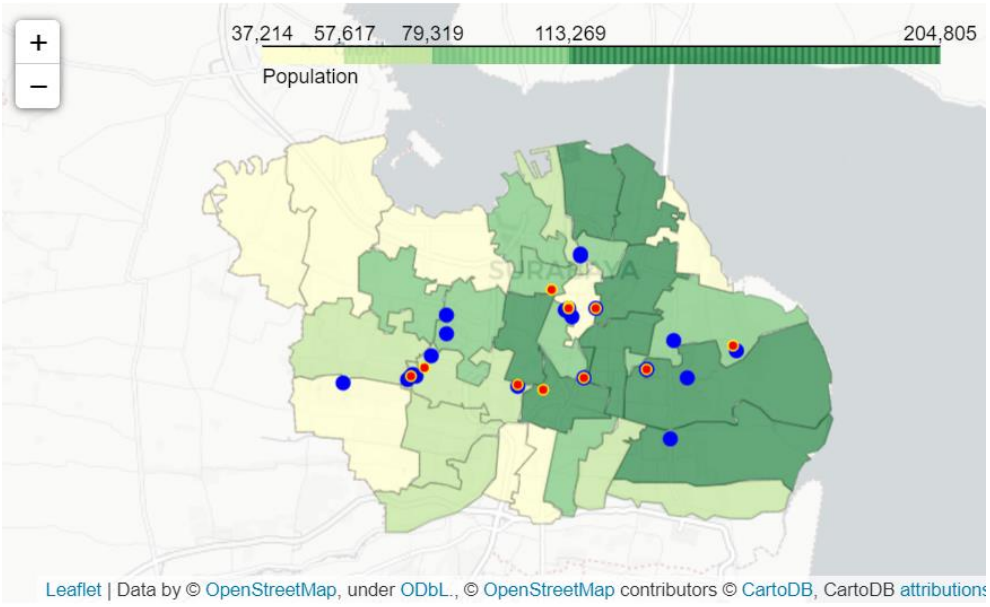
The business purpose of this project is to find a suitable place in Surabaya to open a movie theater. Now we retrieved the following data:

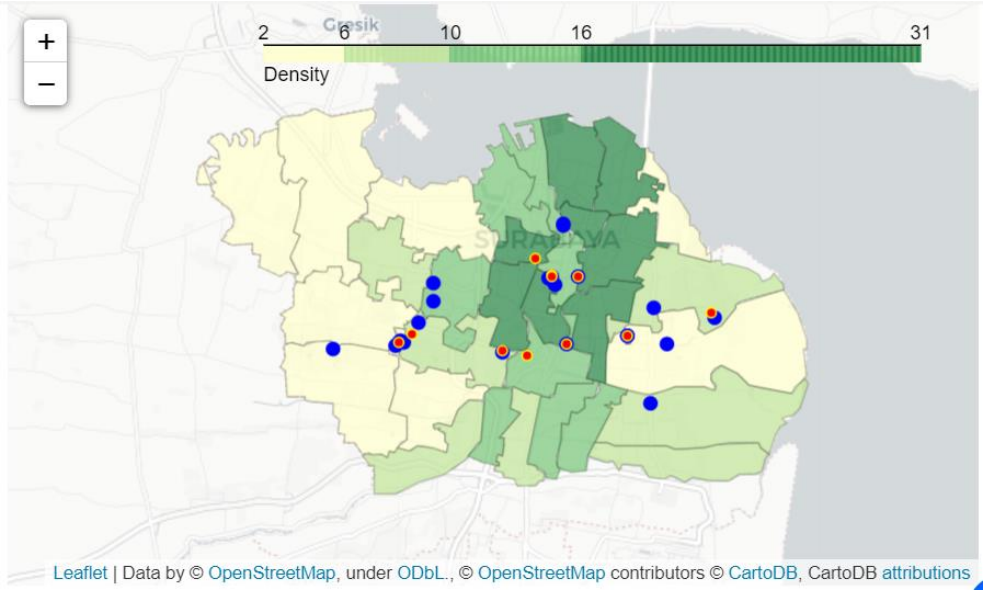
a. All movie theaters data in Surabaya
b. All shopping centers data in Surabaya
c. Surabaya census data for each borough, concretely, Population, Density, Area, Education, and Health data for each borough.
d. Boundary data of each borough in Surabaya

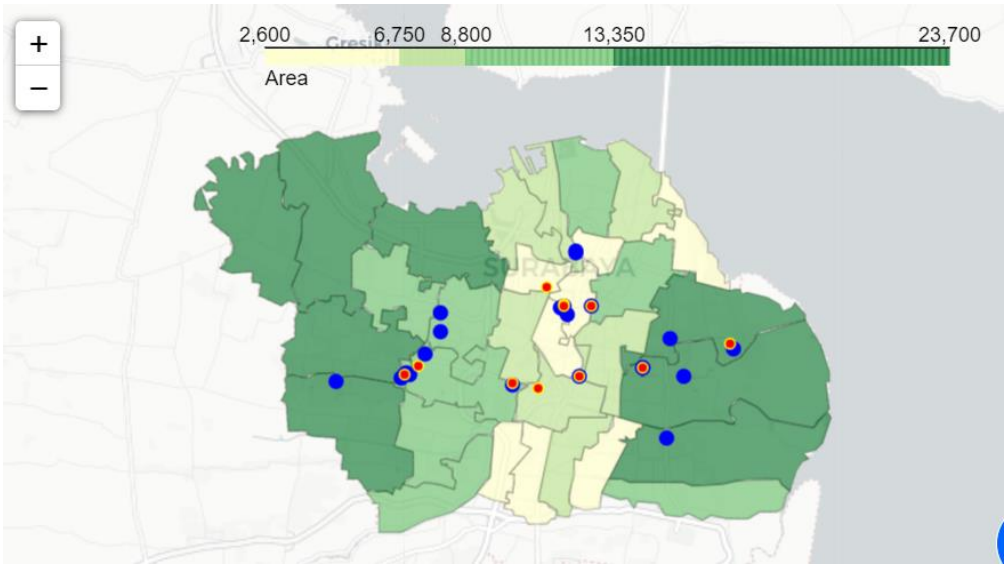
Based on the above raw data, we will try to explore that data more deeply then generate features accordingly. Then, using K-Means to helping us divide into the cluster so we can understand the most promising areas with more shopping malls and fewer movie theaters combine with census data.

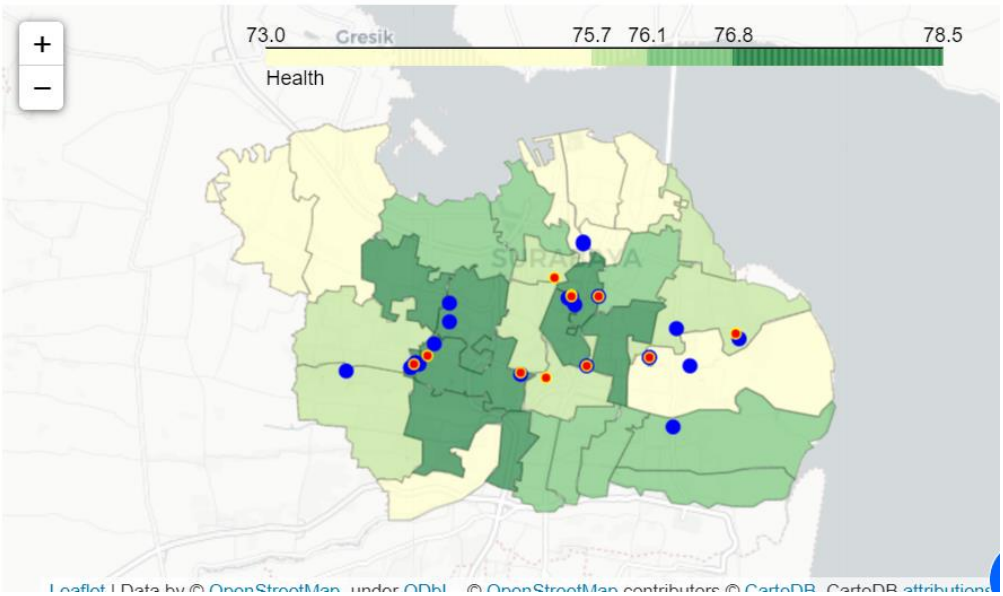
4. Analysis

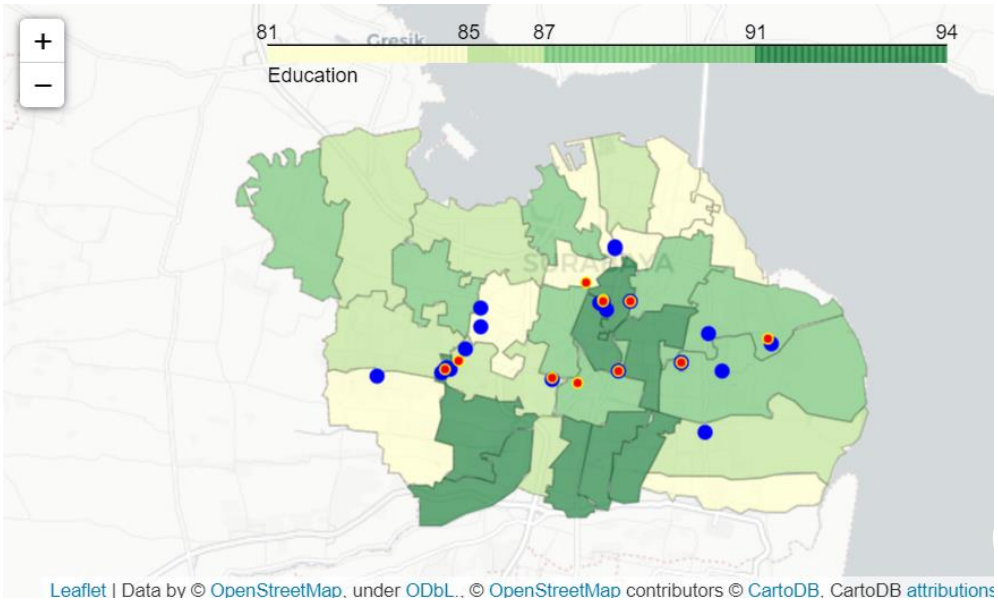
First of all, I will show you the census data distribution on a choropleth map and include the locations of theaters and malls. The first picture is the distribution of population, followed by density, area, health point, and the last education point.






2,600
6,750
8,800
13,350
23,700
Area


73.0
Gresik
75.7
76.1
76.8
78.5
Health

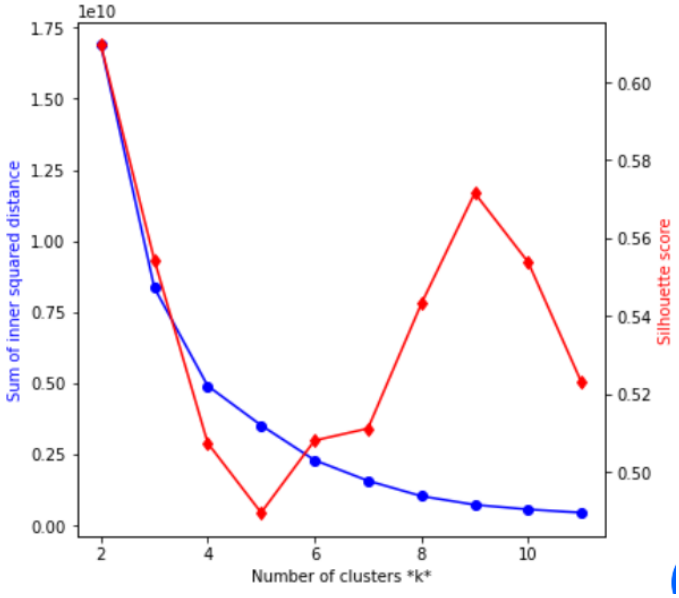

81
85
87
91
94
Education
Leaflet | Data by © OpenStreetMap, under ODbL., © OpenStreetMap contributors © CartoDB, CartoDB attributions

After that, I counted the number of cinemas and malls in every neighborhood.

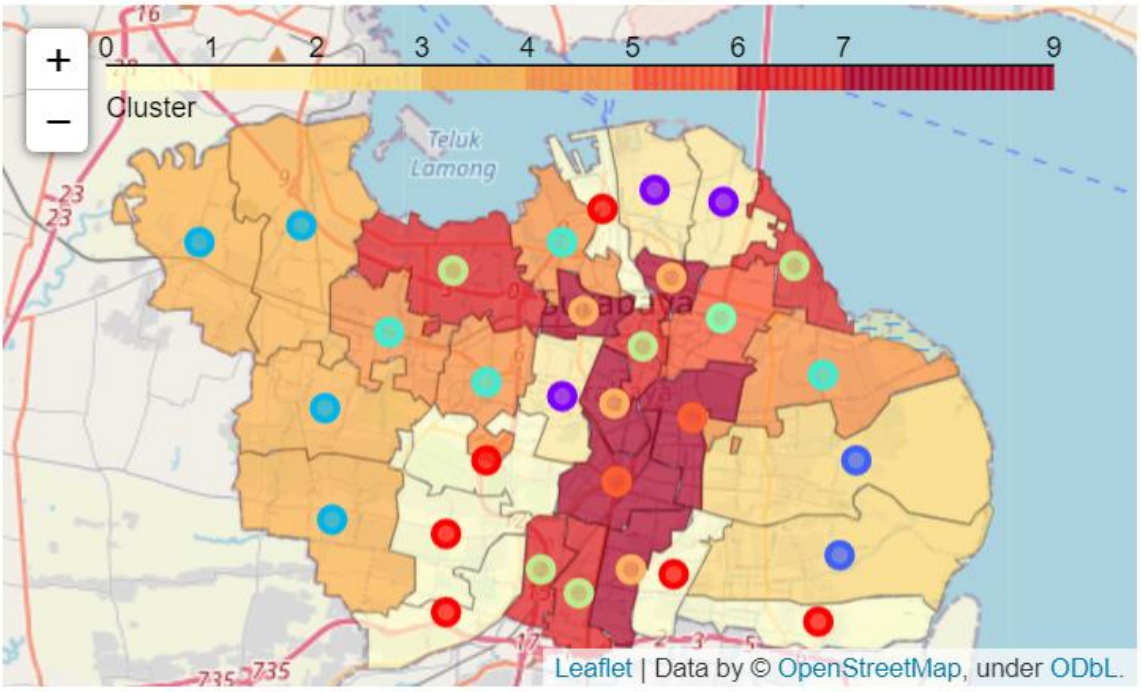
	Neighborhood	Cinemas in Borough
0	Bubutan	1
1	Dukuh Pakis	2
2	Genteng	3
3	Gubeng	1
4	Mulyorejo	1
5	Wonokromo	3

	Neighborhood	Malls in Borough
0	Dukuh Pakis	3
1	Genteng	5
2	Gubeng	2
3	Mulyorejo	1
4	Rungkut	1
5	Sambikerep	3
6	Simokerto	2
7	Sukolilo	2
8	Sukomanunggal	1
9	Wiyung	1
10	Wonokromo	2

Then, we using K-Means to helping us divide into the cluster so we can understand the most promising areas with more shopping malls and fewer movie theaters combine with census data. Also applying Elbow method and Silhouette score to obtain the optimal number of clusters.



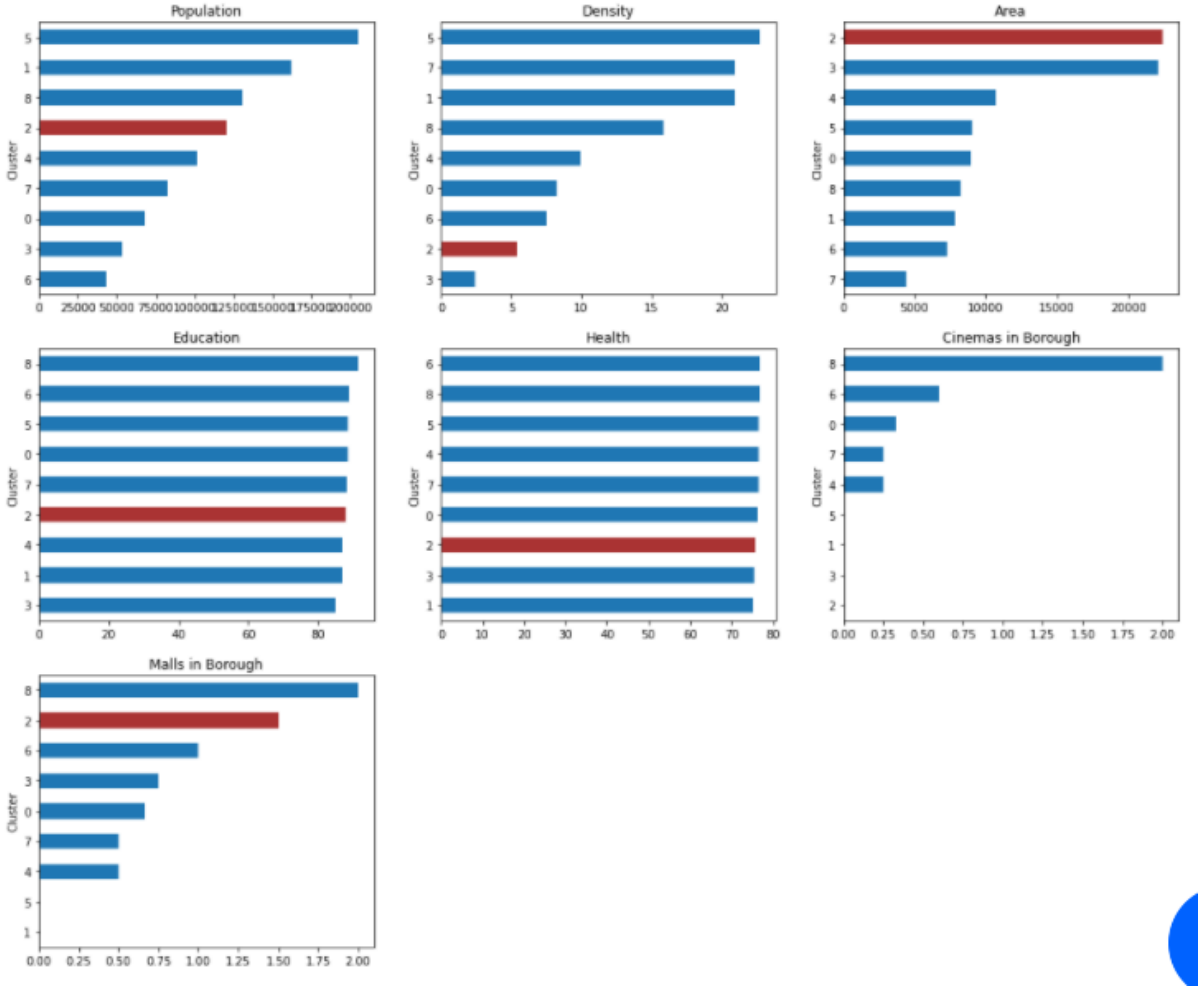
From the figure, we can see Sum of Squared Distance going down when K becomes bigger. Silhouette Score also going down until K=5 but going up start K=6, then going down again start k=10. we choose K=9 for this project, it's a balanced number for both Sum of Squared Distance and Silhouette Score. Let's run the K-Means algorithm again with K=9.



After we classify each region into 9 clusters. Now, we will calculate the score for each cluster to get an insight into which clusters have the potential to open theaters. Because our criterion is an area with a large number of malls but few cinemas, then we use the formula for the average number of malls in the cluster minus the average number of cinemas in the cluster. The following is the calculation result after sorted.

Cluster	Population	Density	Area	Education	Health	Malls in Borough	Cinemas in Borough	Mall Score	Cinema Score	Score
2	120478.500000	5.398255	22400.000000	87.900000	75.800000	1.500000	0.000000	1.500000	0.000000	1.500000
3	53458.250000	2.425411	22125.000000	85.125000	75.325000	0.750000	0.000000	0.750000	0.000000	0.750000
6	43122.600000	7.547607	7300.000000	88.820000	76.880000	1.000000	0.600000	1.000000	0.600000	0.400000
0	68119.166667	8.265838	8933.333333	88.450000	76.150000	0.666667	0.333333	0.666667	0.333333	0.333333
4	101272.000000	9.936251	10700.000000	86.975000	76.550000	0.500000	0.250000	0.500000	0.250000	0.250000
7	82416.500000	20.969667	4400.000000	88.375000	76.375000	0.500000	0.250000	0.500000	0.250000	0.250000
1	161824.000000	20.962264	7833.333333	86.833333	75.133333	0.000000	0.000000	0.000000	0.000000	0.000000
5	204805.000000	22.756111	9000.000000	88.600000	76.600000	0.000000	0.000000	0.000000	0.000000	0.000000
8	130669.000000	15.843879	8250.000000	91.450000	76.700000	2.000000	2.000000	2.000000	2.000000	0.000000

Cluster 2 has the highest score, it has more shopping malls and fewer movie theaters. Let's plot all clusters for comparison of each feature in bar charts using matplotlib.pyplot library. We highlight Cluster 2 which is our target cluster.



From the bar chart, we can see that Cluster 2 has the widest Area among all the clusters. Furthermore, it has fairly more shopping centers area and doesn't have any movie theaters.

Next, we sort all boroughs in Cluster 2 by Score in descending. They will be our first choice position to open a cinema.

	Borough	Population	Density	Area	Education	Health	Cinemas in Borough	Malls in Borough	Mall Score	Cinema Score	Score
14	Sukolilo	119873	5.057932	23700	89.6	75.5	0.0	2.0	2.0	0.0	2.0
11	Rungkut	121084	5.738578	21100	86.2	76.1	0.0	1.0	1.0	0.0	1.0

As the above statistics information, there are 1~2 shopping malls, but without any movie theater. Looks quite good selections.

5. Result and Discussion

I grouped sub-districts in the city of Surabaya into 9 clusters according to census data including population, density, area, education and health. Existing shopping center information and cinema information are also considered when running the clustering algorithm.

From data analysis and visualization, we can see that the cinema is always near the usual shopping center, which inspires us to know the areas with more shopping centers and fewer cinemas.

After the K-Means Clustering machine learning algorithm, we got a cluster with most of the closest shopping malls and on average fewer movie theaters. We also find other characteristics of the clusters. This shows that the cluster has the largest area.

I sorted all areas of the cluster by shopping mall and cinema info in descending order targeting to cover more shopping malls and fewer cinemas in local or nearby cells.

I draw our conclusions with the 2 most promising sub-districts that meet all our conditions, is **Sukolilo** and **Rungkut**. This recommended zone will be a good starting point for further analysis. Other factors that can be considered, eg. real traffic data and revenue of each cinema, the average income of citizens, and more specific locations. They will be of great help in finding more accurate results.

6. Conclusion

This project aims is to find a sub-district in the city of Surabaya to open a cinema. After taking data from multiple data sources and processing it into clean data frames, applying the K-Means clustering algorithm, we chose a cluster with more shopping malls and on average fewer cinemas. From these clusters, it was found that the two most promising sub-districts were used as starting points for final exploration by stakeholders.

The final decision on the optimal cinema location will be made by stakeholders based on the specific characteristics of the environment and location in each recommended zone, taking into account additional factors such as parking space at each location, existing cinema traffic in the cluster, and their revenue streams, etc.

7. Limitation

1. Cinema and mall data is only limited to data in the Foursquare API.
2. The census data presented in 2010.

8. Reference

1. https://surabayakota.bps.go.id/
2. https://openstreetmap.id/en/data-surabaya/
3. https://developer.foursquare.com/docs/places-api/
4. https://positionstack.com/documentation
5. https://python-visualization.github.io/folium/
6. https://plotly.com/python/choropleth-maps/
7. https://towardsdatascience.com/the-battle-of-the-neighborhoods-open-a-movie-theater-in-montreal-355cf5c679b8