Battle of Neighborhoods

The Philippines

BY
Mark Angelo Ruz

Problem Statement

- To recommend the best neighborhood to live, to buy a house, to rent an apartment or build a restaurant etc in the Philippines .
- To understand the similarities and differences between the neighborhoods using Unsupervised K-Mean Clustering Algorithm.

Objective

- Collecting the top trending venues in the using Foursquare API(Beautiful Soup, http request)
- Forming neighborhood clusters based on venue categories using unsupervised k-mean clustering algorithm(sklearn)
- Identifying and understanding the similarities and differences between two chosen neighborhoods to retrieve more insights and to conclude which neighborhood wins over other.

Python packages and Dependencies:

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library

Work flow

- Web Scraping and Data Wrangling
- Top Trending Places Extraction and Clustering
- Decision Making based on the clustered neighborhoods, Population Distribution, School Ratings, Median House Price Analysis

Web Scraping and Data Wrangling

Beautiful Soup

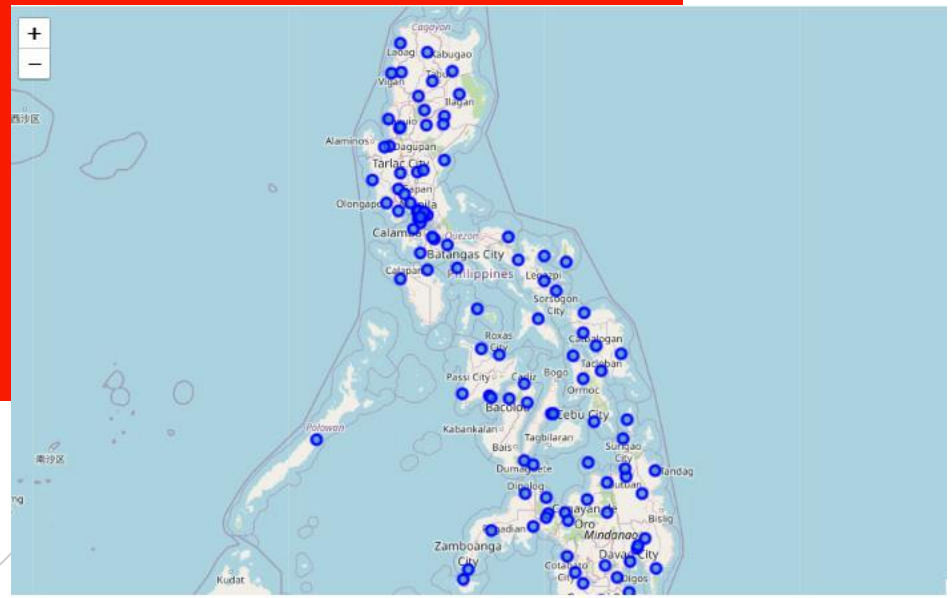
Collecting Neighborhood/Postal code

Google Maps API

To Collect Geographical Data

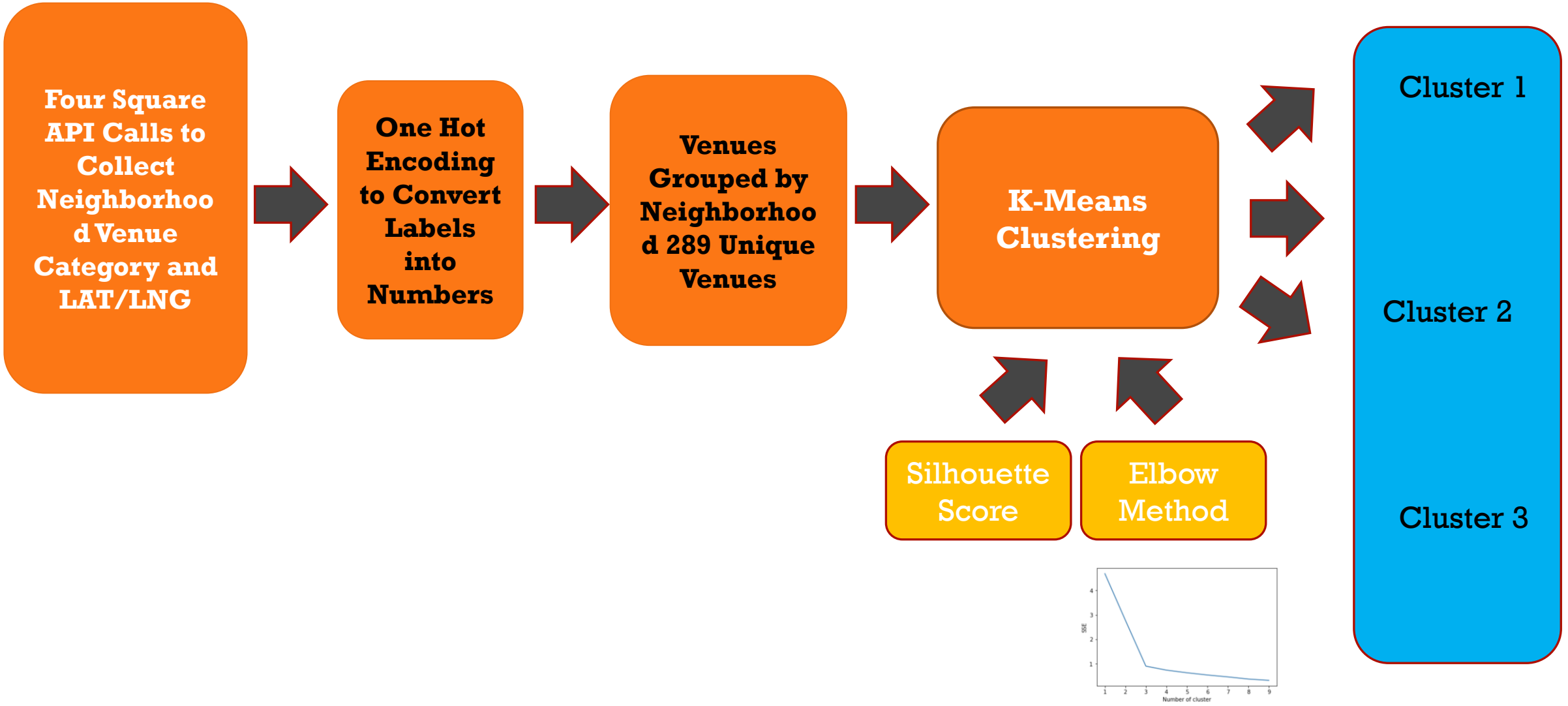
	city	Neighborhood	lat	lng	country	iso2	iso3	admin_name	capital	population	id
0	Manila	Manila	14.6042	120.9822	Philippines	PH	PHL	Manila	primary	11100000.0	1608618140
1	Quezon City	Quezon City	14.6504	121.0300	Philippines	PH	PHL	Quezon	admin	2761720.0	1608974097
2	Davao	Davao	7.1100	125.6300	Philippines	PH	PHL	Davao	admin	1402000.0	1608906877
3	Cagayan de Oro	Cagayan de Oro	8.4508	124.6853	Philippines	PH	PHL	Cagayan de Oro	admin	1121561.0	1608831546
4	General Santos	General Santos	6.1108	125.1747	Philippines	PH	PHL	General Santos	admin	950530.0	1608171585

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	Ayala Triangle Gardens	[{'id': '4bf58dd8d48988d163941735', 'name': 'P...	14.556471	121.023204
1	The Peninsula Manila	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	14.555066	121.025466
2	Escolta	[{'id': '4eb1bd1c3b7b55596b4a748f', 'name': 'F...	14.555485	121.025509
3	Banapple Pies & Cheesecakes	[{'id': '4bf58dd8d48988d1c4941735', 'name': 'R...	14.556634	121.023619
4	Top of the Citi by Chef Jessie	[{'id': '4bf58dd8d48988d1c4941735', 'name': 'R...	14.558932	121.025147



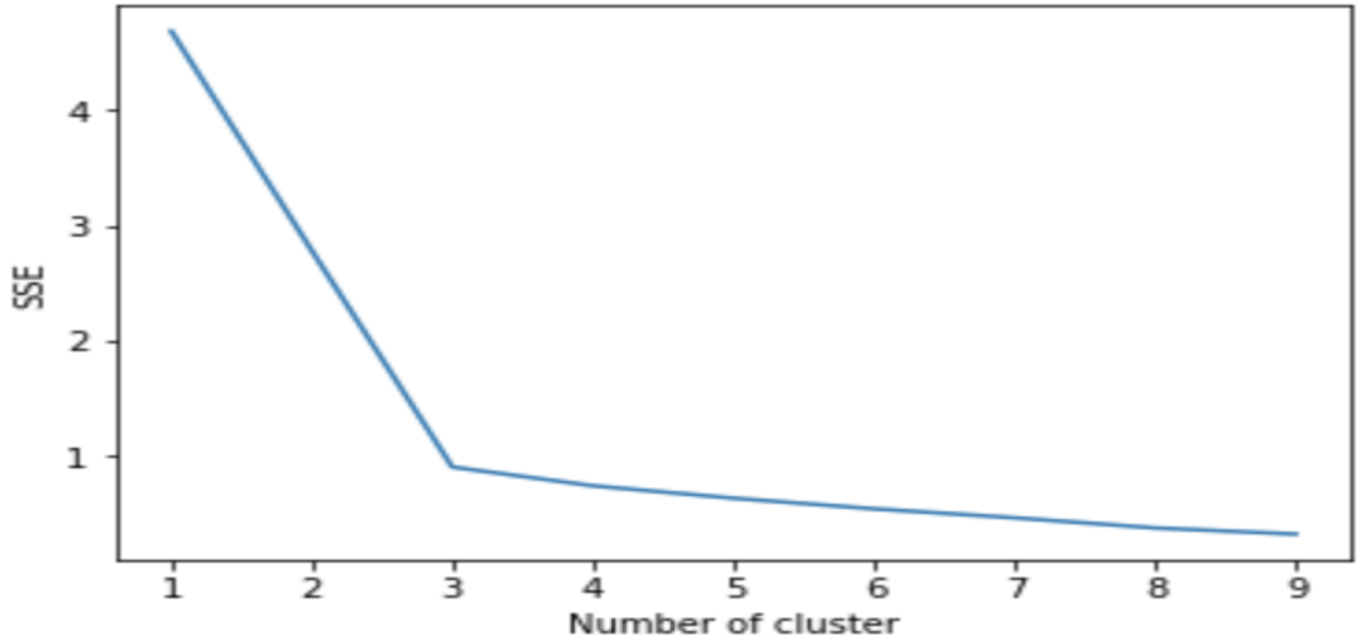
Folium Visualization for the Philippines Neighborhood

Venues Extraction using Four Square API and Clustering

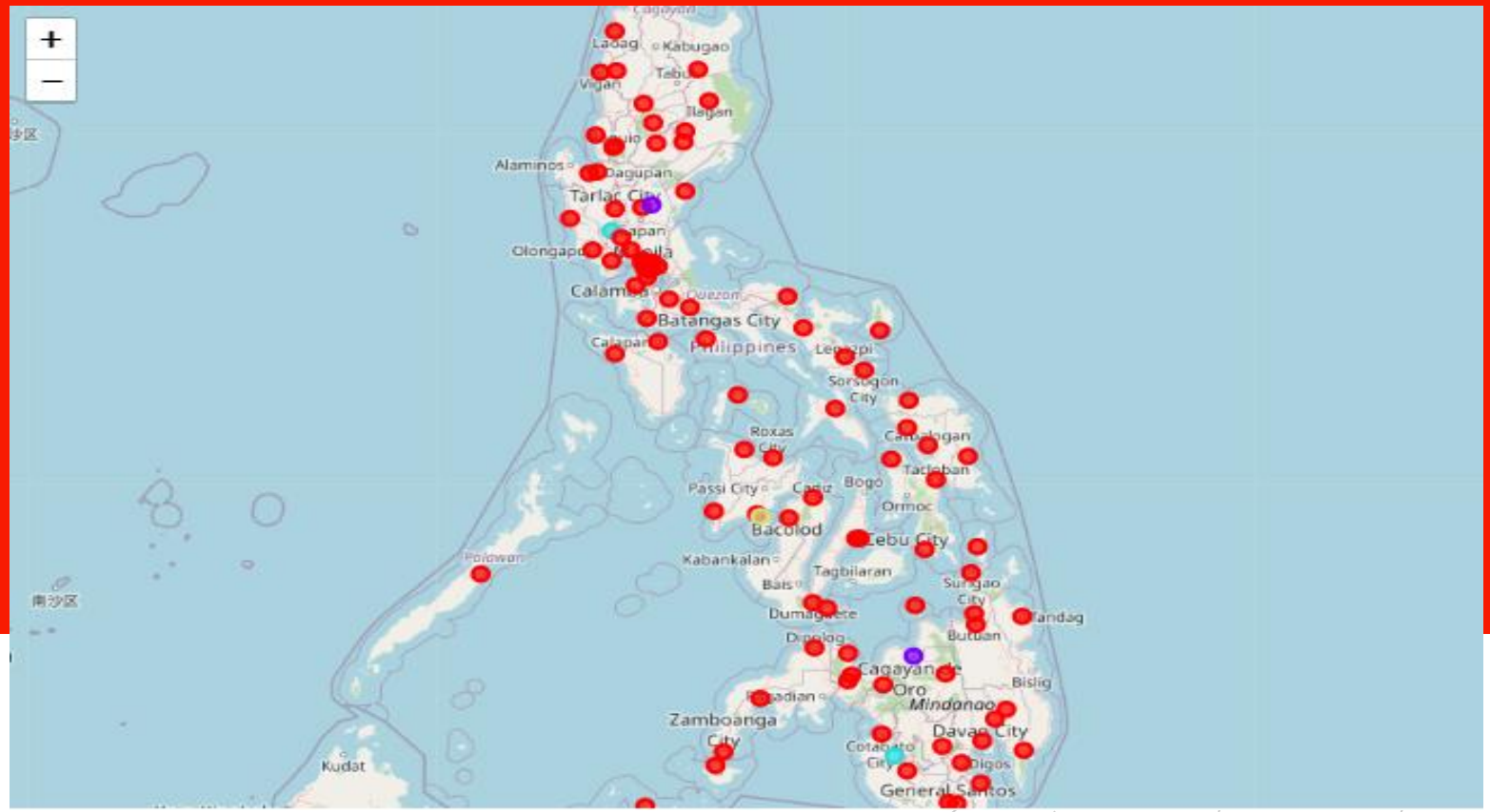


Elbow Criterion Method

Elbow method is to run k-means clustering on a given dataset for a range of values of k and for each value of k and calculate sum of squared errors (SSE).



Cluster Neighborhood



sklearn.metrics.silhouette_score

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample.

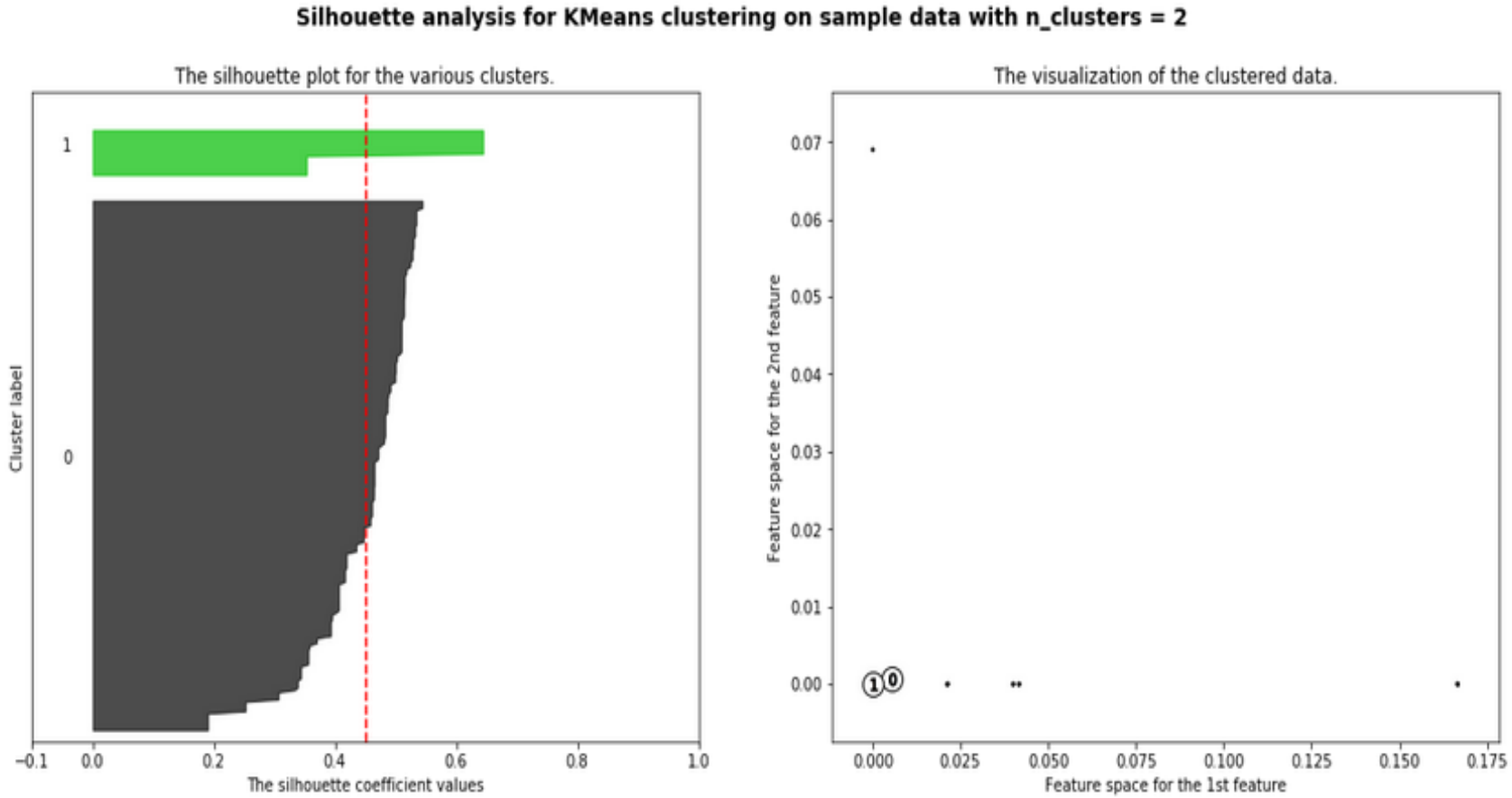
The formula for the Silhouette Coefficient of a sample is

(b - a) / max(a, b).

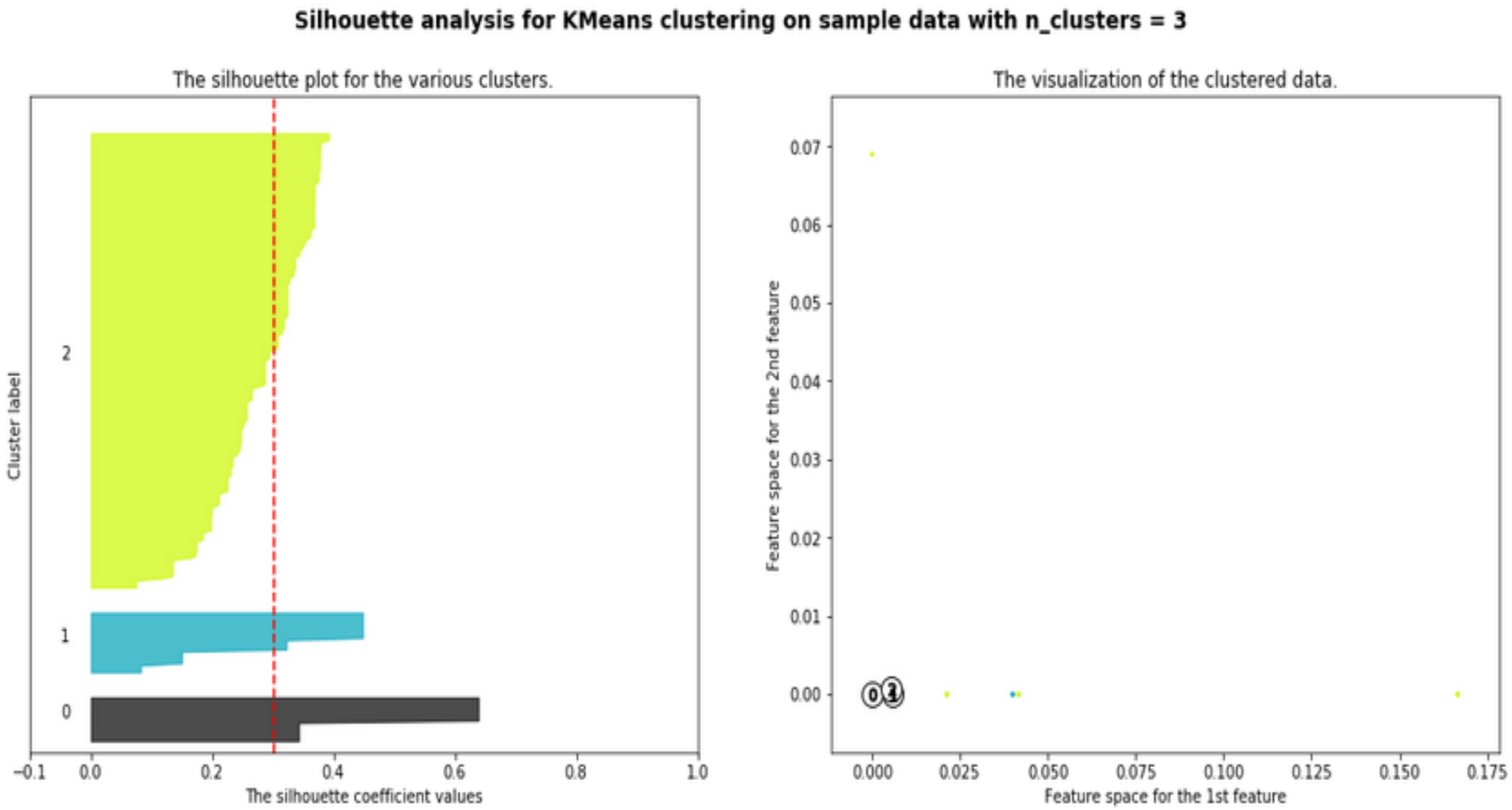
The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster.

Siluhouette Score and Cluster Visualizations

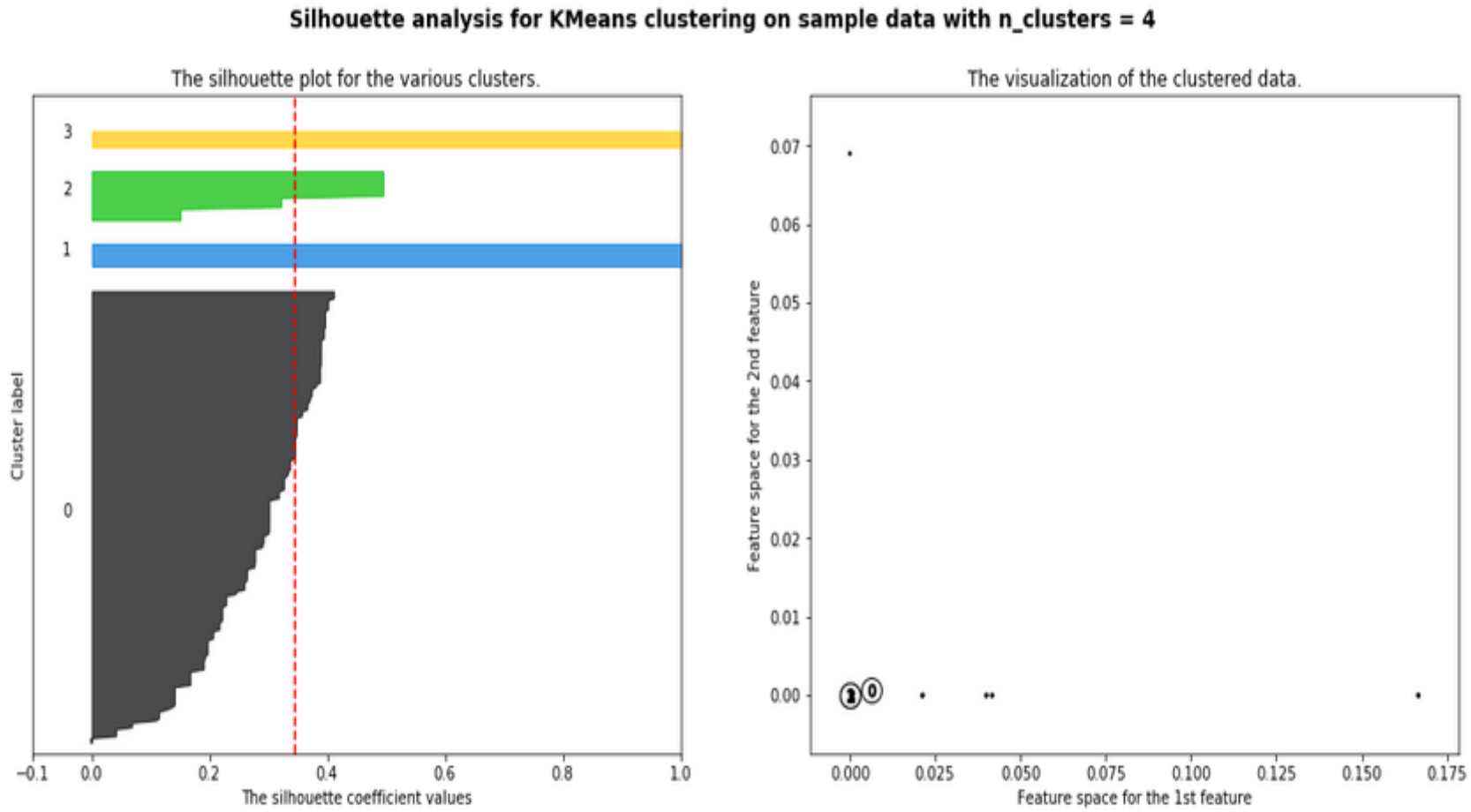
```
For 2  Clusters  the average silhouette_score is : 0.45129474688303467
For 3  Clusters  the average silhouette_score is : 0.30224267891889856
For 4  Clusters  the average silhouette_score is : 0.3449303029949109
For 5  Clusters  the average silhouette_score is : 0.3858319904775224
```



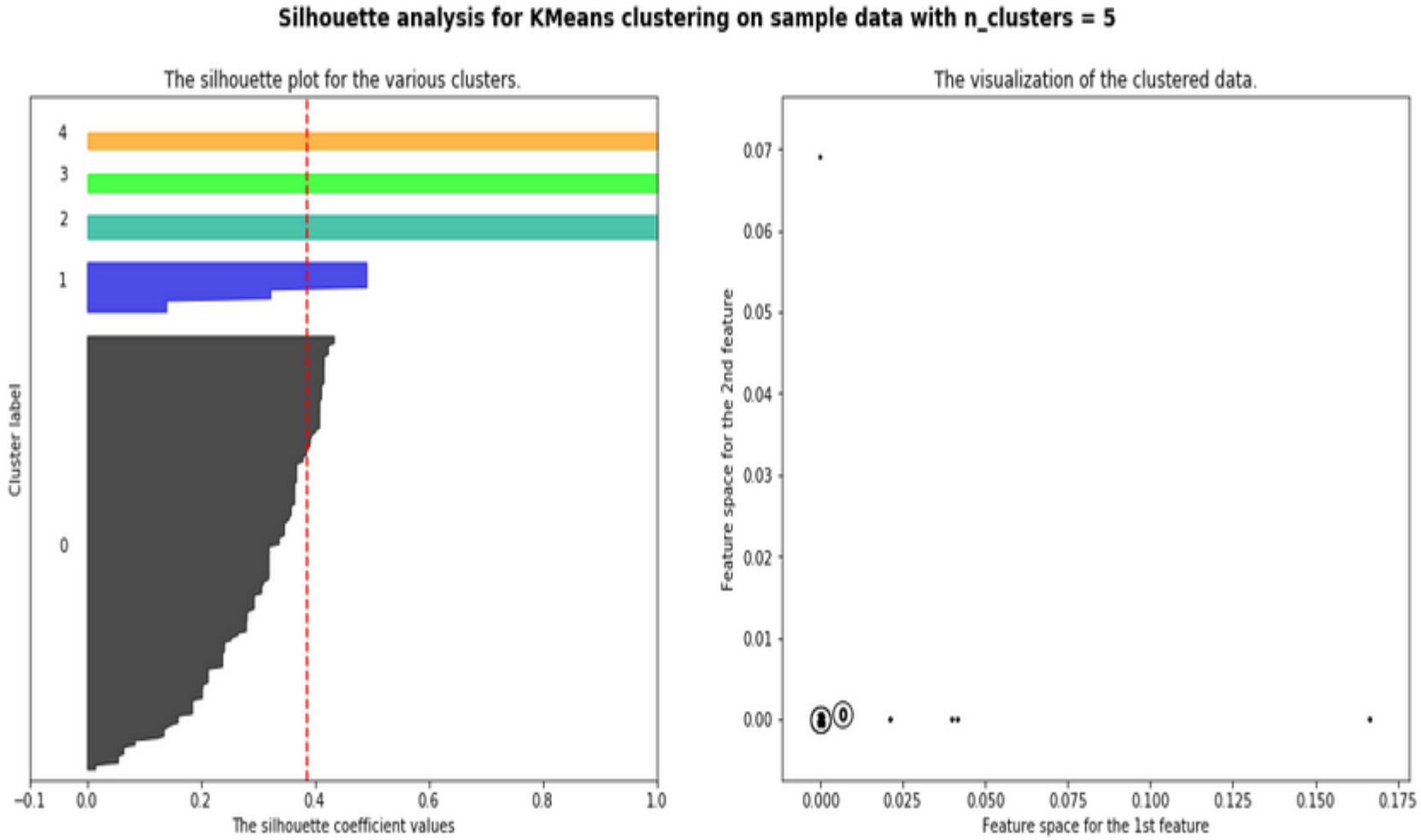
Siluhouette Score and Cluster Visualizations



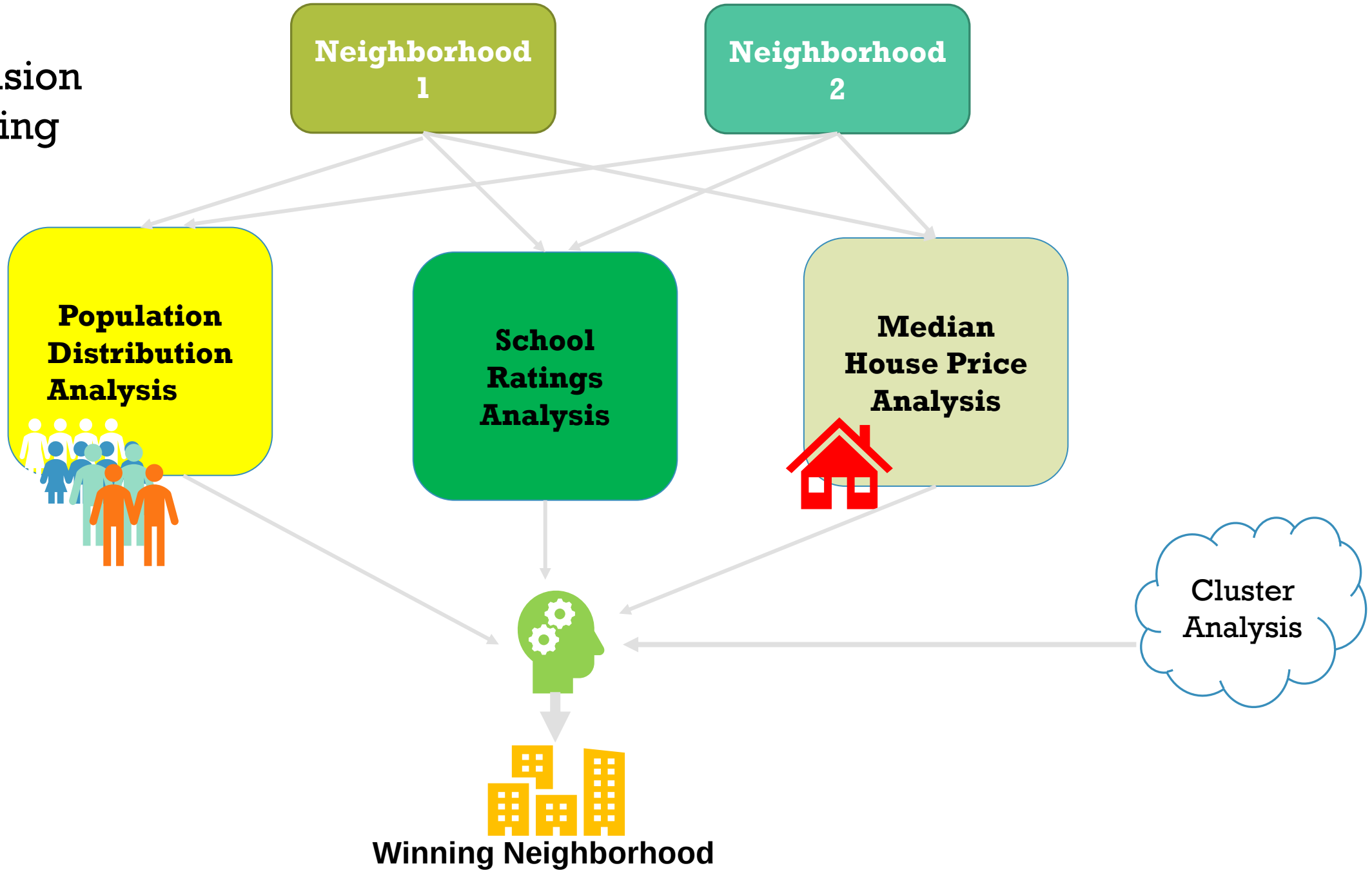
Siluhouette Score and Cluster Visualizations



Siluhouette Score and Cluster Visualizations



Decision Making



Comparison between Cities - Philippines

There are 3 classification for cities in the Philippines based on the following characteristics:

- Cluster 1: Mostly urbanized cities which are densely populated. Restaurants, shopping malls, and accommodation are the mostly recognized establishments.
- Cluster 2: Tourism based cities where establishments such as accommodation, transport hub, and recreational centers are mostly common.
- Cluster 3: Cities with large land area and less populous. Agriculture is the main source of living.

Conclusion

- The project will recommend the resulting classification for further research in cities of the Philippines.
- Homebuyers who are expecting to move in the cities will have an idea on the common establishment found for each classification.

Thank You