Clustering Cities in the Philippines

Applied Data Science Capstone – Coursera/IBM
by D. Fenix

Introduction

The year 2020 will be remembered as the year when the COVID-19 pandemic changed the world. Its impact covers all levels and aspects of society. As of December 14, 2020, the World Health Organization reported that over 71 million people have contracted the virus and over 1.6 million people have died. In my home country, the Philippines, a total of 449,400 cases have been reported and 8,733 deaths.

While governments continue to impose restrictions on human activity in order to limit the spread of the virus, they are also looking ahead for ways to address its economic and societal impact. On May 6, 2020, Philippine President Rodrigo Duterte issued Executive Order No. 114, instituting the '*Balik Probinsya'* program. This goal of program is to "develop the quality of life in the rural areas, in effect decongesting the

densely populated areas of the country such as Metro Manila by encouraging people to move to the countryside once COVID-19-related quarantine measures are lifted."[1]

While most migrants to the National Capital Region (NCR) in the Philippines might consider going back to their home cities or provinces as a first choice, a presentation of other options could be invaluable. This is where a clustering exercise of Philippine cities can help.

The audience that can benefit from this project are:

- Metro Manila residents who want to identify rural areas/provinces that are potentially good options for relocation.
- Provincial governments that can use the output as a reference for what establishments they can promote and incentivize to make their cities more attractive for relocation.

Data

The following data will be used in clustering Philippine cities.

- **List of Philippine cities** from https://en.wikipedia.org/wiki/List_of_cities_in_the_Philippines. This data will be read into Python using pandas. This is the list that will be clustered based on common venues. Coordinates for the different cities are also available from the backend table in Wikipedia.
- **Forsquare API** will be used to get the most common venues of each city. From doing data exploration of the Foursquare data, it might be necessary to group the venue types into larger buckets. For example, venue types like American Restaurant and Japanese Restaurant can be grouped into Restaurant.

Methodology

Data Acquisition and Cleaning

The list of Philippine cities is sourced from Wikipedia. This is read into a pandas dataframe.

[1] https://en.wikipedia.org/wiki/Balik_Probinsya

A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V · W · X · Y · Z

City	Population (2015) [4]	Area [5][i]	Density (2015)	Province [5][ii]	Region	Legal class [6]	Charter [iii]	Date of	
								Approval [iv]	Ratification [v]
Alaminos	89,708	164.26 km² (63.42 sq mi)	550/km² (1,400/sq mi)	Pangasinan	I	CC	RA 09025 [7]	March 5, 2001 [7]	March 28, 2001 [8]
Angeles	411,634 [vi]	60.27 km² (23.27 sq mi)	6,200/km² (16,000/sq mi)	*Pampanga*	III	HUC	RA 03700 [9]	June 22, 1963 [9]	January 1, 1964
Antipolo	776,386	306.10 km² (118.19 sq mi)	2,500/km² (6,500/sq mi)	Rizal	IV-A	CC	RA 08508 [10]	February 13, 1998 [10]	April 4, 1998
Bacolod	561,875	162.67 km² (62.81 sq mi)	3,500/km² (9,100/sq mi)	*Negros Occidental*	VI	HUC	CA 326 [11]	June 18, 1938 [12]	October 19, 1938 [13]
Bacoor	600,609	46.17 km² (17.83 sq mi)	13,000/km² (34,000/sq mi)	Cavite	IV-A	CC	RA 10160 [14]	April 10, 2012 [14]	June 23, 2012 [15]
Bago	170,981	401.20 km² (154.90 sq mi)	430/km² (1,100/sq mi)	Negros Occidental	VI	CC	RA 04382 [16]	June 19, 1965 [16]	February 19, 1966 [17]
Baguio	345,366	57.51 km² (22.20 sq mi)	6,000/km² (16,000/sq mi)	*Benguet*	CAR	HUC	Act 1963	September 1, 1909	*September 1, 1909*
Bais	76,291	319.64 km² (123.41 sq mi)	240/km² (620/sq mi)	Negros Oriental	VII	CC	RA 05444 [18]	September 9, 1968 [18]	September 9, 1968
Balanga	96,061	111.63 km² (43.10 sq mi)	860/km² (2,200/sq mi)	Bataan	III	CC	RA 08984 [19]	December 5, 2000 [19]	December 30, 2000
Batac	55,201	161.06 km² (62.19 sq mi)	340/km² (880/sq mi)	Ilocos Norte	I	CC	RA 09407 [20]	March 24, 2007 [20]	June 23, 2007 [21]

Below is the pandas dataframe. While the geographic coordinates are not displayed in the website, it is available in the backend table.

	Coordinates	City	Population	Area	Density	Province	Region
0	16°09′23″N 119°58′49″E / 16.1565°N 119.9804°E	Alaminos	89708	164.26 km2(63.42 sq mi)	550/km2(1,400/sq mi)	Pangasinan	I
1	15°08′24″N 120°35′16″E / 15.1399°N 120.5879°E	Angeles	411,634 [vi]	60.27 km2(23.27 sq mi)	6,200/km2(16,000/sq mi)	Pampanga	III
2	14°35′13″N 121°10′33″E / 14.5870°N 121.1758°E	Antipolo	776386	306.10 km2(118.19 sq mi)	2,500/km2(6,500/sq mi)	Rizal	IV-A
3	10°40′34″N 122°57′05″E / 10.6762°N 122.9513°E	Bacolod	561875	162.67 km2(62.81 sq mi)	3,500/km2(9,100/sq mi)	Negros Occidental	VI
4	14°27′34″N 120°56′24″E / 14.4594°N 120.9401°E	Bacoor	600609	46.17 km2(17.83 sq mi)	13,000/km2(34,000/sq mi)	Cavite	IV-A
...	...	...	...	...	...	...	...
143	10°53′46″N 123°04′21″E / 10.8962°N 123.0726°E	Victorias	87933	133.92 km2(51.71 sq mi)	660/km2(1,700/sq mi)	Negros Occidental	VI
144	17°34′22″N 120°23′12″E / 17.5729°N 120.3867°E	Vigan	53879	25.12 km2(9.70 sq mi)	2,100/km2(5,400/sq mi)	Ilocos Sur	I
145	6°54′17″N 122°04′35″E / 6.9046°N 122.0763°E	Zamboanga City	861799	1,414.70 km2(546.22 sq mi)	580/km2(1,500/sq mi)	Zamboanga del Sur [ii]	IX
146	Table notes ^ Land area figures are taken from...	Table notes ^ Land area figures are taken from...	Table notes ^ Land area figures are taken from...	Table notes ^ Land area figures are taken from...	Table notes ^ Land area figures are taken from...	Table notes ^ Land area figures are taken from...	Table notes ^ Land area figures are taken from...	Table n figures
147	Dates of inauguration/organization Many pre-19...	Dates of inauguration/organization Many pre-19...	Dates of inauguration/organization Many pre-19...	Dates of inauguration/organization Many pre-19...	Dates of inauguration/organization Many pre-19...	Dates of inauguration/organization Many pre-19...	Dates of inauguration/organization Many pre-19...	inaugurat

148 rows × 11 columns

The following steps were taken to make the dataset usable for subsequent analysis.

- Remove last two rows that do not represent data.
- Split Area and Density, retaining the value in metric units which is the standard in the Philippines.
- Split Coordinates field into Longitude and Latitude.
- Process numeric fields to so they can be converted to numeric datatypes. This includes removing spaces and non-numeric characters or symbols.
- Remove columns that are not needed such as Region, Legal Class, Charter, Approval and Ratification Dates.
- Append Province to City name to differentiate cities with the same name but located in different provinces, such as Naga in Cebu and Naga in Camarines Sur.

- Resolve missing Province names. Upon checking cities with unassigned Province names, they were all located in Metro Manila. Although Metro Manila is not technically a province, this will be used for tagging.

After all these steps, the final dataset has 146 rows of Philippine cities and 6 columns.

	City	Population	Area_sqkm	Density_sqkm	Latitude	Longitude
1	Manila, Metro Manila	1780148	42.88	42000	14.5906	120.9799
2	Pasay, Metro Manila	416522	13.97	30000	14.5437	120.9954
3	Caloocan, Metro Manila	1583978	53.33	30000	14.6571	120.9841
4	Navotas, Metro Manila	249463	10.77	28000	14.6569	120.9478
5	Makati, Metro Manila	582602	21.57	27000	14.5568	121.0235
...	...	...	...	...	...	...
142	Tabuk, Kalinga	110642	700.25	160	17.4136	121.4440
143	Bayugan, Agusan del Sur	103202	688.77	150	8.7143	125.7481
144	Borongan, Eastern Samar	69297	475.00	150	11.6085	125.4353
145	Ilagan, Isabela	145568	1166.26	120	17.1442	121.8889
146	Puerto Princesa, Palawan	255116	2381.02	110	9.7376	118.7350

Exploratory Data Analysis - Manila

Foursquare API was used to extract the venues present in each city. Data exploration was first done on Manila, which is the most densely populated city in the Philippines. The parameters Limit=200 and Radius=1000 were used for the API.

A total of 100 venues were pulled for Manila, with 45 unique categories. Below shows the first 5 venues.

	name	categories	lat	lng
0	Adobo Connection	Filipino Restaurant	14.590246	120.982349
1	The French Baker	Bakery	14.590214	120.983140
2	9 Spoons	Restaurant	14.589810	120.978785
3	Coco Bango Cafe	Asian Restaurant	14.589654	120.975609
4	The Bayleaf Hotel	Hotel	14.589945	120.978808

There are 3 restaurants, 1 bakery and 1 hotel. The list of 45 venue categories for Manila is shown below.

```
array(['Filipino Restaurant', 'Bakery', 'Restaurant', 'Asian Restaurant',
       'Hotel', 'Antique Shop', 'Roof Deck', 'Historic Site', 'Church',
       'Ice Cream Shop', 'Spanish Restaurant', 'History Museum',
       'Japanese Restaurant', 'Korean Restaurant', 'Café', 'Pizza Place',
       'Arts & Crafts Store', 'Monument / Landmark', 'Italian Restaurant',
       'Chinese Restaurant', 'Coffee Shop', 'Museum', 'Public Art',
       'Beer Garden', 'Fast Food Restaurant', 'Park', 'Deli / Bodega',
       'Gift Shop', 'Donut Shop', 'Shoe Store', 'Ramen Restaurant',
       'Department Store', 'Camera Store', 'Government Building', 'Food',
       'Bubble Tea Shop', 'Fountain', 'Jewelry Store', 'Breakfast Spot',
       'Burger Joint', 'Electronics Store', 'Convenience Store',
       'BBQ Joint', 'Fried Chicken Joint', 'Plaza'], dtype=object)
```

The categories are very specific so it would be useful to group them into larger buckets. For example, Filipino Restaurant, Restaurant and Asian Restaurant can all be grouped into Restaurant. This makes sense because we use venue categories as indicators of quality of life. The presence of restaurants is more relevant than the specific cuisine served.

Processing All Cities

The next step is to extract venues for all 146 cities using Foursquare API. A total of 2207 rows were pulled coming from 203 unique categories. These categories were extracted to Excel, mapped to larger buckets of categories and imported back to pandas for re-categorization. After this step, there are 18 venue categories.

```
array(['Restaurant', 'Lodging', 'Entertainment', 'Food shop',
       'Specialty Shops', 'Museum and Landmarks', 'Pub or Bar',
       'Nature and Outdoor', 'Shopping', 'Office', 'Personal Services',
       'Athletics or Sports', 'Grocery or Market', 'Transportation',
       'Utilities', 'Hospital', 'School', 'Church'], dtype=object)
```

The final dataset for clustering is in the format below.

	City	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
0	Alaminos, Pangasinan	Restaurant	Athletics or Sports	Specialty Shops	Shopping	Transportation	Lodging	Church	Entertainment	Food shop	Grocery or Market
1	Angeles, Pampanga	Restaurant	Specialty Shops	Food shop	Utilities	Museum and Landmarks	Church	Entertainment	Grocery or Market	Hospital	Lodging
2	Antipolo, Rizal	Restaurant	Food shop	Grocery or Market	Specialty Shops	Utilities	Museum and Landmarks	Church	Entertainment	Hospital	Lodging
3	Bacolod, Negros Occidental	Restaurant	Lodging	Personal Services	Food shop	Nature and Outdoor	Entertainment	Museum and Landmarks	Athletics or Sports	Pub or Bar	Office
4	Bacoor, Cavite	Food shop	Athletics or Sports	Grocery or Market	Transportation	Church	Entertainment	Hospital	Lodging	Museum and Landmarks	Utilities

K-Means Clustering

K-means Clustering was done on the 146 cities with five as the number of clusters. There are 5 cities that were not assigned a cluster because they did not have data from Foursquare. These cities were dropped in the rest of the analysis, resulting in 141 final cities.

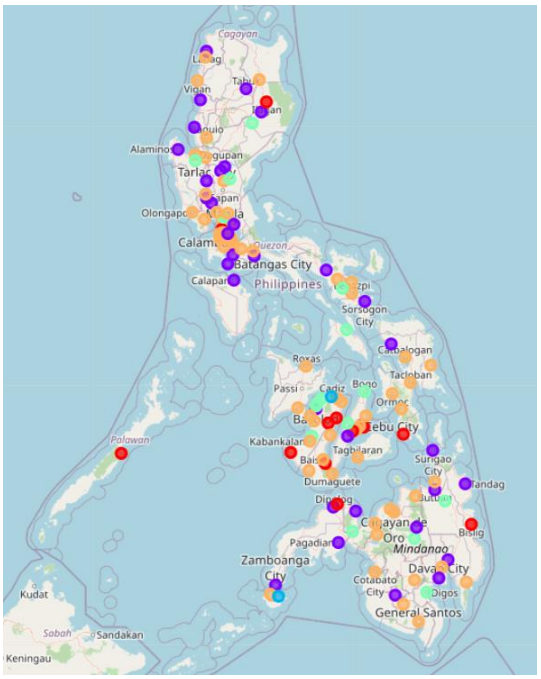
	City	Population	Area_sqkm	Density_sqkm	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
49	Santo Tomas, Batangas	179844	95.41	1900	14.0505	121.1100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75	Ozamiz, Misamis Occidental	141828	169.95	830	8.1691	123.8454	NaN	NaN	NaN	NaN	NaN	NaN	NaN
82	Gapan, Nueva Ecija	110303	164.44	670	15.2977	120.9566	NaN	NaN	NaN	NaN	NaN	NaN	NaN
115	Samal, Davao del Norte	104123	301.30	350	7.0878	125.7400	NaN	NaN	NaN	NaN	NaN	NaN	NaN
121	Passi, Iloilo	80544	251.39	320	11.1174	122.6432	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Results

The clustering exercise produced the following distribution of cities. We are interested in where the Metro Manila cities are since the *Balik Probinsya* program targets the residents from these cities for outbound relocation.

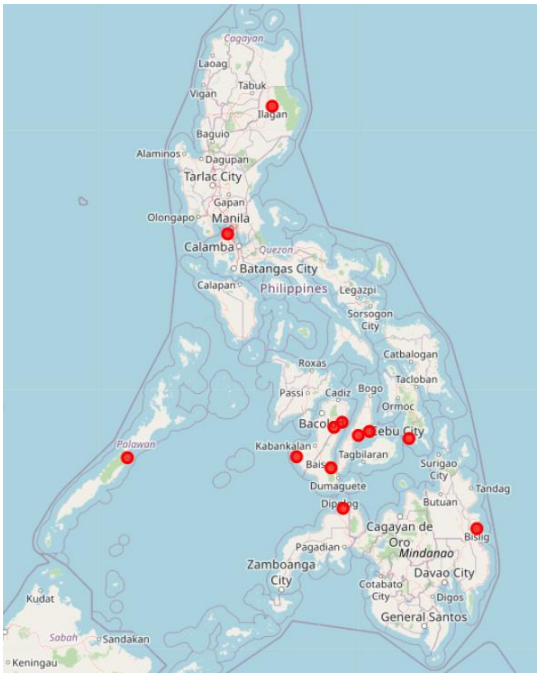
Cluster	Number of Cities	Metro Manila Cities
1	12	0
2	37	4
3	2	0
4	19	2
5	71	10
Total	141	16

The map below shows the location of the cities by cluster. Each of the clusters will be examined in more detail in the following section.

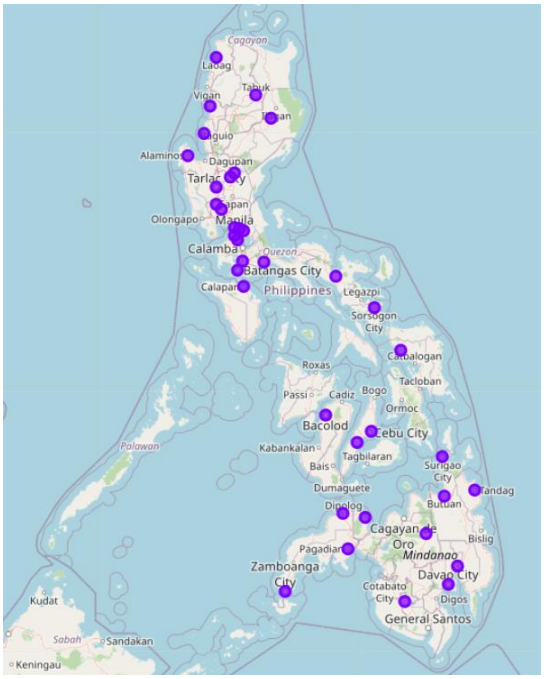


Cluster 1

Cluster 1 has a total of 12 cities. Most of the cities are located in the Visayas region. The cities are 'Cavite City, Cavite', 'Lapu-Lapu, Cebu', 'Naga, Cebu', 'Maasin, Southern Leyte', 'Canlaon, Negros Oriental', 'Tanjay, Negros Oriental', 'San Carlos, Negros Occidental', 'Bislig, Surigao del Sur', 'Dapitan, Zamboanga del Norte', 'Sipalay, Negros Occidental', 'Ilagan, Isabela', 'Puerto Princesa, Palawan'.



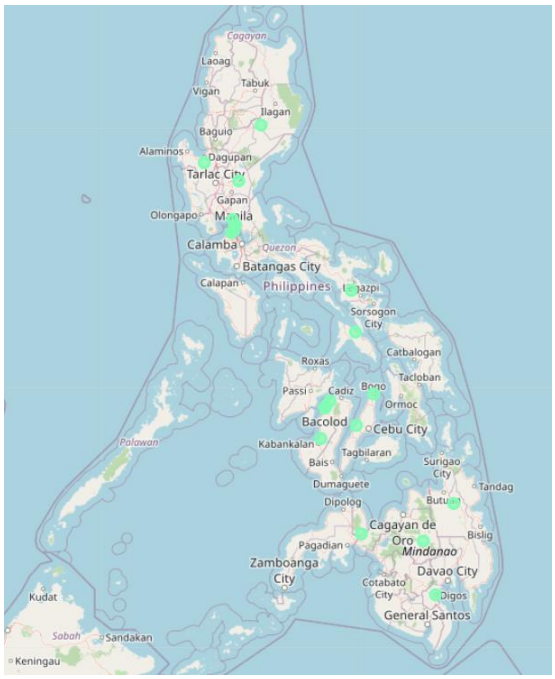
Cluster 2
Cluster 2 has a total of 37 cities and four are in Metro Manila, highlighted below. The cities are well-distributed throughout the entire country. The cities in this cluster are **'Caloocan, Metro Manila'**, **'Marikina, Metro Manila'**, **'Las Piñas, Metro Manila'**, 'Mandaue, Cebu', **'Muntinlupa, Metro Manila'**, 'Angeles, Pampanga', 'San Fernando, Pampanga', 'Bacolod, Negros Occidental', 'Lucena, Quezon', 'Antipolo, Rizal', 'Naga, Camarines Sur', 'Lipa, Batangas', 'Tagum, Davao del Norte', 'San Fernando, La Union', 'Tarlac City, Tarlac', 'Batangas City, Batangas', 'Carcar, Cebu', 'Laoag, Ilocos Norte', 'San Jose, Nueva Ecija', 'Davao City, Davao del Sur', 'Tacurong, Sultan Kudarat', 'Surigao City, Surigao del Norte', 'Sorsogon City, Sorsogon', 'Candon, Ilocos Sur', 'Zamboanga City, Zamboanga del Sur', 'Alaminos, Pangasinan', 'Dipolog, Zamboanga del Norte', 'Calapan, Oriental Mindoro', 'Pagadian, Zamboanga del Sur', 'Butuan, Agusan del Norte', 'Cauayan, Isabela', 'Oroquieta, Misamis Occidental', 'Calbayog, Samar', 'Tandag, Surigao del Sur', 'Malaybalay, Bukidnon', 'Muñoz, Nueva Ecija', 'Tabuk, Kalinga'.



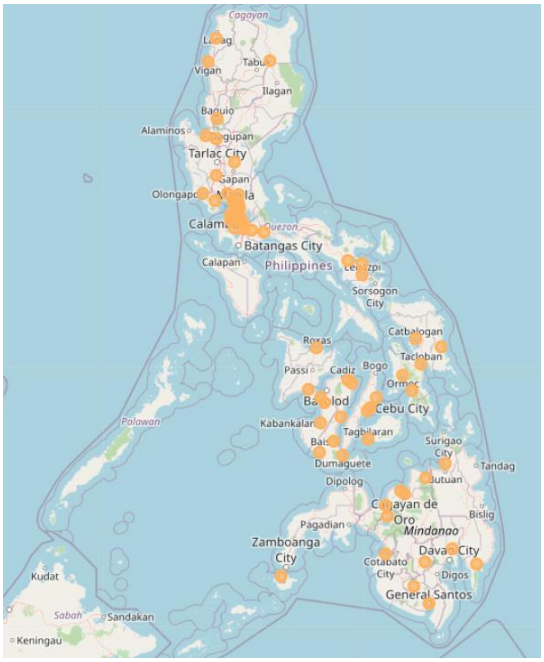
Cluster 3
Cluster 3 is the smallest and it contains two cities, none in Metro Manila. The cities are 'Cadiz, Negros Occidental', 'Lamitan, Basilan'.

Cluster 4

Cluster 4 has 19 cities and two are in Metro Manila. The cities are **'San Juan, Metro Manila'**, **'Parañaque, Metro Manila'**, 'Bacoor, Cavite', 'Meycauayan, Bulacan', 'San Carlos, Pangasinan', 'Toledo, Cebu', 'Bogo, Cebu', 'Victorias, Negros Occidental', 'Digos, Davao del Sur', 'Silay, Negros Occidental', 'Santiago, Isabela', 'Talisay, Negros Occidental', 'Masbate City, Masbate', 'Ligao, Albay', 'Palayan, Nueva Ecija', 'Tangub, Misamis Occidental', 'Valencia, Bukidnon', 'Himamaylan, Negros Occidental', 'Bayugan, Agusan del Sur'



Cluster 5
Cluster 5 is the largest cluster, with 71 cities and 10 located in Metro Manila. As with Clusters 2 and 4, the cities are well-distributed throughout the country. The cities in this cluster are **'Manila, Metro Manila'**, **'Pasay, Metro Manila'**, **'Navotas, Metro Manila'**, **'Makati, Metro Manila'**, **'Malabon, Metro Manila'**, **'Mandaluyong, Metro Manila'**, **'Quezon City, Metro Manila'**, **'Pasig, Metro Manila'**, **'Taguig, Metro Manila'**, 'San Pedro, Laguna', **'Valenzuela, Metro Manila'**, 'Biñan, Laguna', 'Dasmariñas, Cavite', 'Cabuyao, Laguna', 'Santa Rosa, Laguna', 'Imus, Cavite', 'Baguio, Benguet', 'Iloilo City, Iloilo', 'Talisay, Cebu', 'San Jose del Monte, Bulacan', 'Dagupan, Pangasinan', 'Trece Martires, Cavite', 'General Trias, Cavite', 'Dumaguete, Negros Oriental', 'Malolos, Bulacan', 'Calamba, Laguna', 'Mabalacat, Pampanga', 'Cebu City, Cebu', 'Tagbilaran, Bohol', 'Marawi, Lanao del Sur', 'Vigan, Ilocos Sur', 'Roxas, Capiz', 'Cotabato City, Maguindanao', 'Tanauan, Batangas', 'Cabanatuan, Nueva Ecija', 'Cagayan de Oro, Misamis Oriental', 'Legazpi, Albay', 'Olongapo, Zambales', 'Urdaneta, Pangasinan', 'San Pablo, Laguna', 'Danao, Cebu', 'Tacloban, Leyte', 'General Santos, South Cotabato', 'Tabaco, Albay', 'Tuguegarao, Cagayan', 'Tagaytay, Cavite', 'Balanga, Bataan', 'Iriga, Camarines Sur', 'Panabo, Davao del Norte', 'Koronadal, South Cotabato', 'Isabela, Basilan', 'Escalante, Negros Occidental', 'La Carlota, Negros Occidental', 'El Salvador, Misamis Oriental', 'Sagay, Negros Occidental', 'Tayabas, Quezon', 'Bago, Negros Occidental', 'Iligan, Lanao del Norte', 'Kidapawan, Cotabato', 'Catbalogan, Samar', 'Ormoc, Leyte', 'Cabadbaran, Agusan del Norte', 'Batac, Ilocos Norte', 'Kabankalan, Negros Occidental', 'Guihulngan, Negros Oriental', 'Baybay, Leyte', 'Bais, Negros Oriental', 'Mati, Davao Oriental', 'Gingoog, Misamis Oriental', 'Bayawan, Negros Oriental' and 'Borongan, Eastern Samar'.



Discussion

A total of 146 cities were in the initial dataset and 16 are in Metro Manila. Five cities were not assigned clusters due to unavailable Foursquare data. A total of 5 clusters were created and only 3 contained Metro Manila cities.

Cluster 2 has 4 Metro Manila cities -- Caloocan, Marikina, Las Piñas and Muntinlupa. Residents in these areas can look into the 33 other cities in the same cluster as relocation options. Cities in

Cluster 2 are also well-distributed across the Philippines, which gives urban residents added flexibility especially if they have other considerations such as proximity to their home provinces or employment.

Cluster 4 has 2 Metro Manila cities -- San Juan and Parañaque. Residents might consider the 17 other cities in the same cluster for outbound relocation. As with Cluster 2, Cluster 4 cities are also scattered across the Philippines.

Cluster 5 is the largest cluster and contains the most number of Metro Manila cities at 10. These are Manila, Pasay, Navotas, Makati, Malabon, Mandaluyong, Quezon City, Pasig, Taguig and Valenzuela. Residents in these cities can choose to relocate to 61 other cities across the Philippines.

From the perspective of local governments, the grouping of cities can be used as reference for common venues in Metro Manila cities in order to determine what types of establishments to prioritize or incentivize for future development.

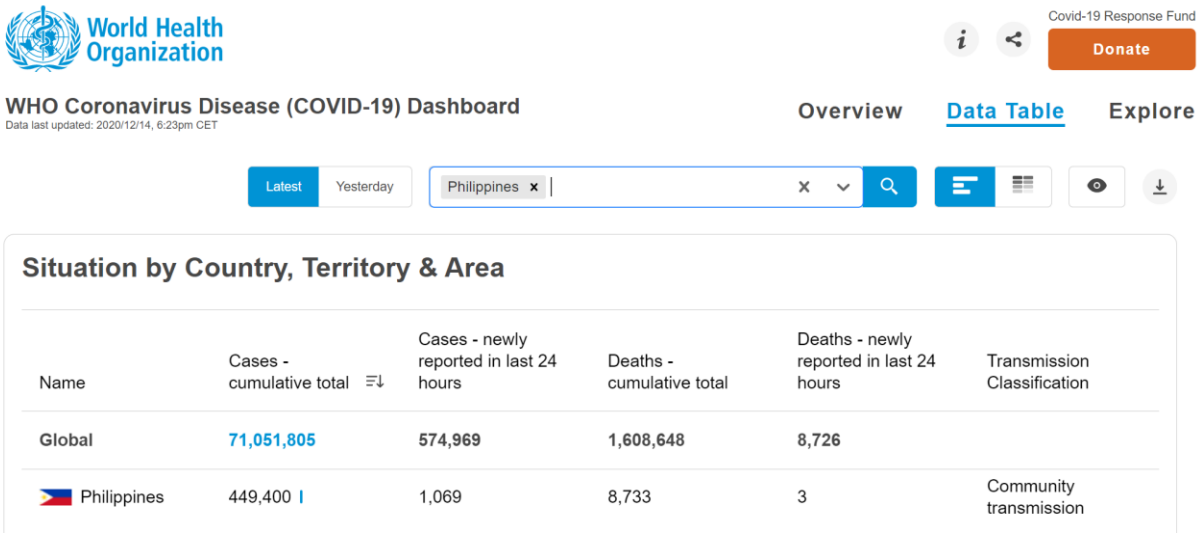
Conclusion

The goal of the 'Balik Probinsya' program is to encourage urban area residents of Metro Manila to relocate back to the provinces. Quality of life is a key factor for relocation and for this exercise, the presence of different venues/facilities is used as an indicator for quality of life. It is not a complete measure but was sufficient to demonstrate and interpret clustering Philippines cities using k-means algorithm.

Five clusters were generated and three contained Metro Manila cities. The other cities in the same cluster, which were well distributed across the country, are the potential relocation options for Metro Manila residents. Cities in the provinces can also look at the clusters to identify what features or venues they can prioritize to become more attractive as relocation options.

References

1. Coronavirus cases and deaths from WHO Coronavirus Disease (COVID-19) Dashboard (https://covid19.who.int/)



2. List of Philippine cities snapshot from Wikipedia - https://en.wikipedia.org/wiki/List_of_cities_in_the_Philippines

3. Balik Probinsya program from Wikipedia - https://en.wikipedia.org/wiki/Balik_Probinsya

4. Quality of Life indicators - https://www.bfs.admin.ch/bfs/en/home/statistics/cross-sectional-topics/city-statistics/indicators-quality-life.html