Clustering the different Neighborhoods in Metro Manila Using K-Means Clustering Algorithm

For completion of the requirements for the Applied Data Science Capstone Couse

John Wincel P. Marino

jwpmarino@gmail.com

April 16, 2020

I. Introduction

A. Description & Discussion of the Background

The National Capital Region of the Philippines (NCR) or more commonly known as Metro Manila is the largest of the three metropolitan areas in the Philippines where almost 13 million people live in a 619.57 square kilometer area.[1][2][3] These numbers make Metro Manila the most densely populated region in the Philippines and the 5th most populous urban area in the world.[4]

As a metropolitan area, it houses thousands of businesses, which means business owners needs to be smart where to place their businesses. Clustering the different areas in the region and identifying the characteristics of the customers in those areas would help businesses to efficiently target their customers. Clustering the different areas in Metro Manila would also help people who want to migrate to the different areas in the region by having more information about the area they want to settle in.

This paper aims to create a map and an information chart about the different types of areas/neighborhood in Metro Manila according to the venues and businesses established using Clustering.

II. Data

A. Data Sources

To produce the Clustering Model, the following datasets were used:

1. List of Areas in Metro Manila. From Wikipedia. [5]
2. Details and types of the venues in Metro Manila using FourSquare API
3. Geospatial data of the areas in Metro Manila using Nominatim API

The details and other information of the venues in Metro Manila gathered from FourSquare will be the main feature to describe the areas in Metro Manila. Only the hundred closest venues within a 500m radius around the given neighborhood was chosen for each neighborhood. The geospatial data is composed of the latitude and longitude coordinates of the areas in Metro Manila and will be used to locate where those areas are in the map.

B. Data Cleaning

Data that was downloaded and/or scrapped was contained in a pandas data frame. The initial source of data included all the neighborhoods in the Philippines, thus, only the neighborhoods in Metro Manila was selected and the rest of the data was discarded. The matching geospatial data of the areas was appended to the base data frame. (Figure 1) Neighborhoods without matching geospatial data was discarded.

	ZIP code	Neighborhood	City	Latitude	Longitude
0	401	Asian Development Bank	San Juan	14.588076	121.058301
1	550	Febias College of Bible	Valenzuela	14.687899	120.981408
2	702	Citibank	Makati	14.607314	121.078924
3	704	Producers Bank	Makati	14.529699	121.041247
4	707	Canadian Embassy	Makati	14.560645	121.016578

Figure 1. Neighborhood data set.which shows the ZIP code, Neighborhood Name, City it is part, and the latitude and longitude cooridnates

The venue data was retrieved using FourSquare API and was appended to the base data frame. The base data frame is composed of the name of the neighborhood, neighborhood latitude and longitude, the name of the venues, the venue latitude and longitude and the venue category. (Figure 2)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Asian Development Bank	14.588076	121.058301	Craft Coffee Revolution	14.585859	121.059212	Coffee Shop
1	Asian Development Bank	14.588076	121.058301	The Nostalgia Dining Lounge	14.587519	121.059753	Restaurant
2	Asian Development Bank	14.588076	121.058301	Gino's Brick Oven Pizza	14.585791	121.059640	Pizza Place
3	Asian Development Bank	14.588076	121.058301	The Café Mediterranean	14.585967	121.057122	Mediterranean Restaurant
4	Asian Development Bank	14.588076	121.058301	Wildflour Café + Bakery	14.585866	121.059573	Café

Figure 2. Base Dataset showing the neighborhood names, neighborhood latitude and longitudes, venue names, venue Latitudes and longitudes, and venue categoy

III. Methodology

Initial visualization of the different neighborhoods was done using Python Folium Library. Each neighborhood is marked with a circle marker according to its geospatial data. (Figure 3)

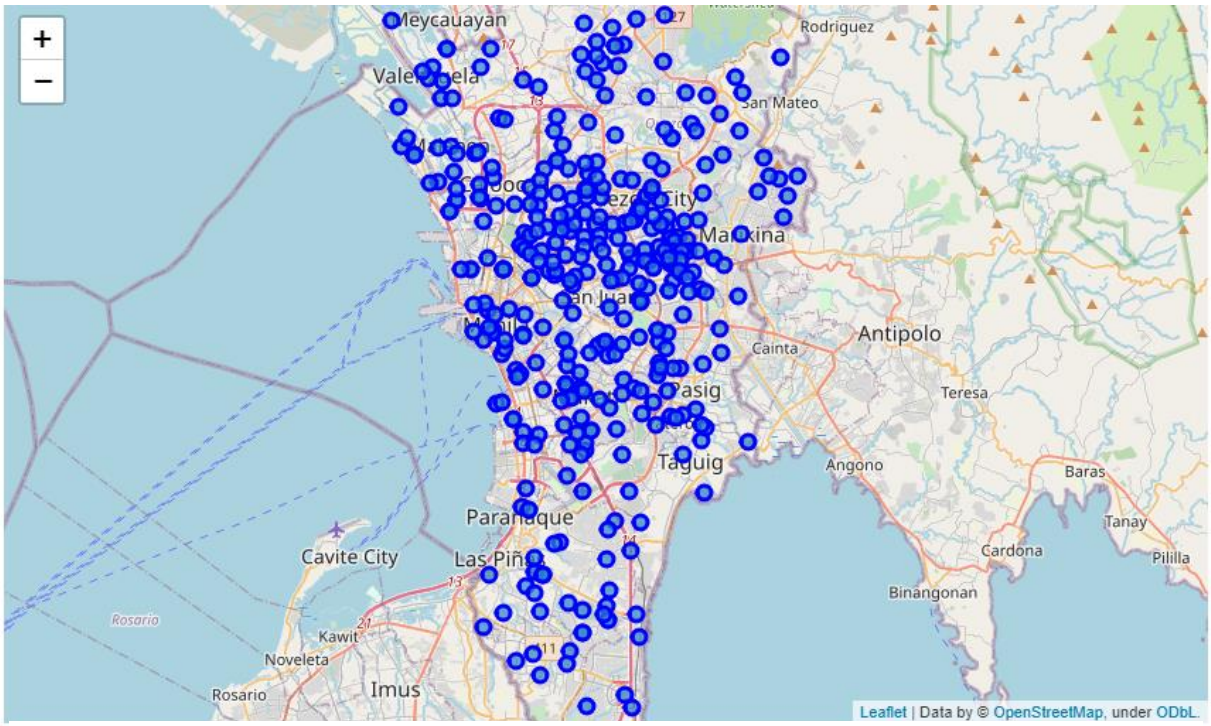

Figure 3. Map showing the location of the different neighborhoods in Metro Manila

There are a total of 337 types of venues from the FourSquare dataset. A lot of these types are similar and overlapping. Only the general types of venues are of concern and so, 16 general venue types were considered. (Figure 4.a) The 337 more specific types where mapped out to those 16 categories. Examples of the specific venue types that are included to a general venue type is shown in figure 4.b.

Venue Type	Venue Category
Accomodation	Hostel
	Hotel
	Hotel Pool
	Lounge
	Motel
	Rest Area
	Roof Deck
	Travel Lounge

Venue Type	Venue Category
Residential	Boarding House
	Residential Building
	(Apartment/Condo)

	General Venue Types
1	Accomodation
2	Art/Culture/Museum
3	Business Hub
4	Café/Tea/Dessert
5	City Essentials
6	Convenience stores and Food Essentials
7	Cosmetics, Clothes, Beauty and Health Services
8	Entertainment
9	Fast Food
10	Gadgets, Hardware, Toys, etc.
11	Night Life
12	Residential
13	Restaurants
14	Sports & Fitness
15	Tourist Destination
16	Vehicle/Transport Related

Figure 4. (a)(Left) Tables that show the general venue type and what specific venue types it comprises. (b) (Right) The 16 general venue types

A bar graph was plotted to visualize the number of venues per venue type. (Figure 5 & 6) It shows that there is a heavy bias towards venue types Restaurant and Café/Tea/ Dessert which is to be expected since Metro Manila is a metropolitan area but this would lead to a skewed model. In order to solve this, an Importance Factor was applied which emphasizes the low venue count types and deemphasize the high venue count types. (Figure 7) This will make low count venue types such as Business Hub or Residential have a more prominent impact towards areas with high densities of these types of venues.

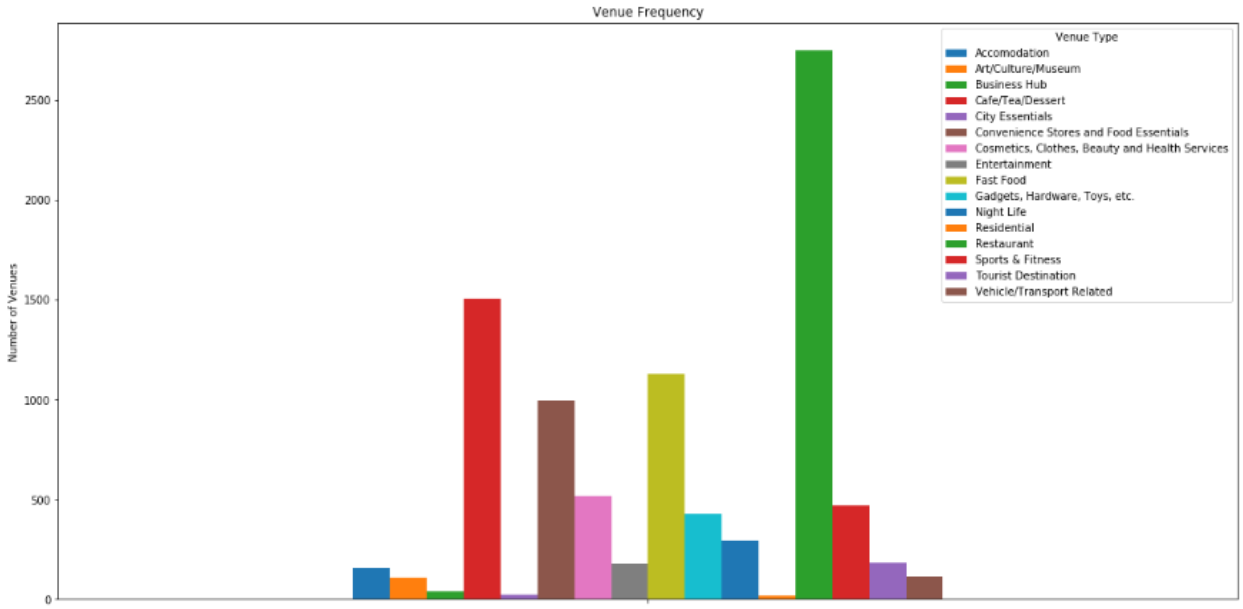

Figure 5. Bar graph that shows the number of venues per venue type

Venue Type	Venue Category
Accomodation	158
Art/Culture/Museum	109
Business Hub	41
Cafe/Tea/Dessert	1505
City Essentials	26
Convenience Stores and Food Essentials	996
Cosmetics, Clothes, Beauty and Health Services	520
Entertainment	176
Fast Food	1127
Gadgets, Hardware, Toys, etc.	425
Night Life	295
Residential	17
Restaurant	2750
Sports & Fitness	469
Tourist Destination	181
Vehicle/Transport Related	113

Figure 6. Table that shows the number of venues per venue type

Venue Type	Importance factor
Accomodation	0.107595
Art/Culture/Museum	0.155963
Business Hub	0.414634
Cafe/Tea/Dessert	0.011296
City Essentials	0.653846
Convenience Stores and Food Essentials	0.017068
Cosmetics, Clothes, Beauty and Health Services	0.032692
Entertainment	0.096591
Fast Food	0.015084
Gadgets, Hardware, Toys, etc.	0.040000
Night Life	0.057627
Residential	1.000000
Restaurant	0.006182
Sports & Fitness	0.036247
Tourist Destination	0.093923
Vehicle/Transport Related	0.150442

Figure 7. Table that shows the importance factor of each venue type.

Clustering Algorithms require integer type values to produce a model thus One-hot Encoding was used to assign dummy variables to each venue type with values 1 to signify that it is that venue type or 0 which signifies it is not that venue type. The importance factor was then applied.

K-Means Algorithm was used to cluster the neighborhoods. In order to find the appropriate number of clusters, The Elbow method was used. The algorithm was iterated using the number of clusters from 2 to 24. The sum of squared error was plotted against the number of clusters. It was determined that the most appropriate number of clusters is seven as the graph shows an 'elbow' or a drastic change in slope. (Figure 8)

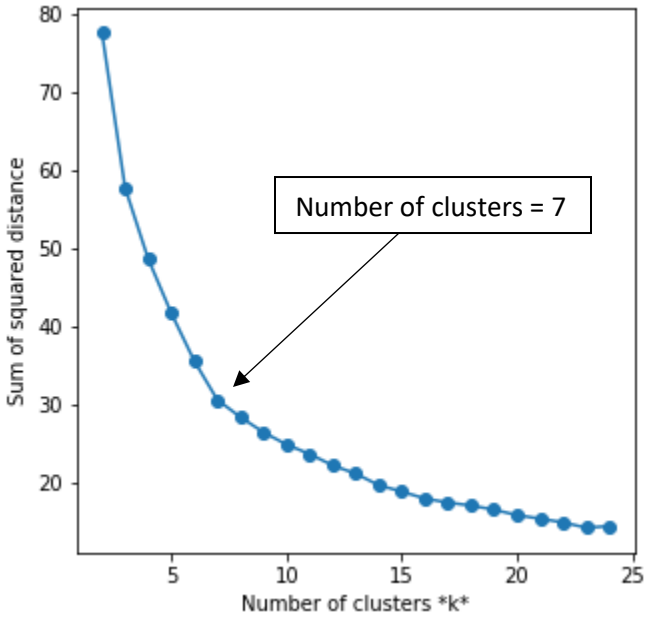

Figure 8. Plot of the K-means algorithm with different number of cluster parameters. The arrow points to the elbow or the most optimal number of cluster parameter.

Cluster labels that were produced from the model was appended to the data frame with its appropriate assignment.

IV. Results and Discussion.

A map using Python Folium was produced with markers that show the individual neighborhoods. The different colors of the markers signify what cluster that neighborhood is included. A popup label is also included to show more details about the neighborhood you clicked. (Figure 9)

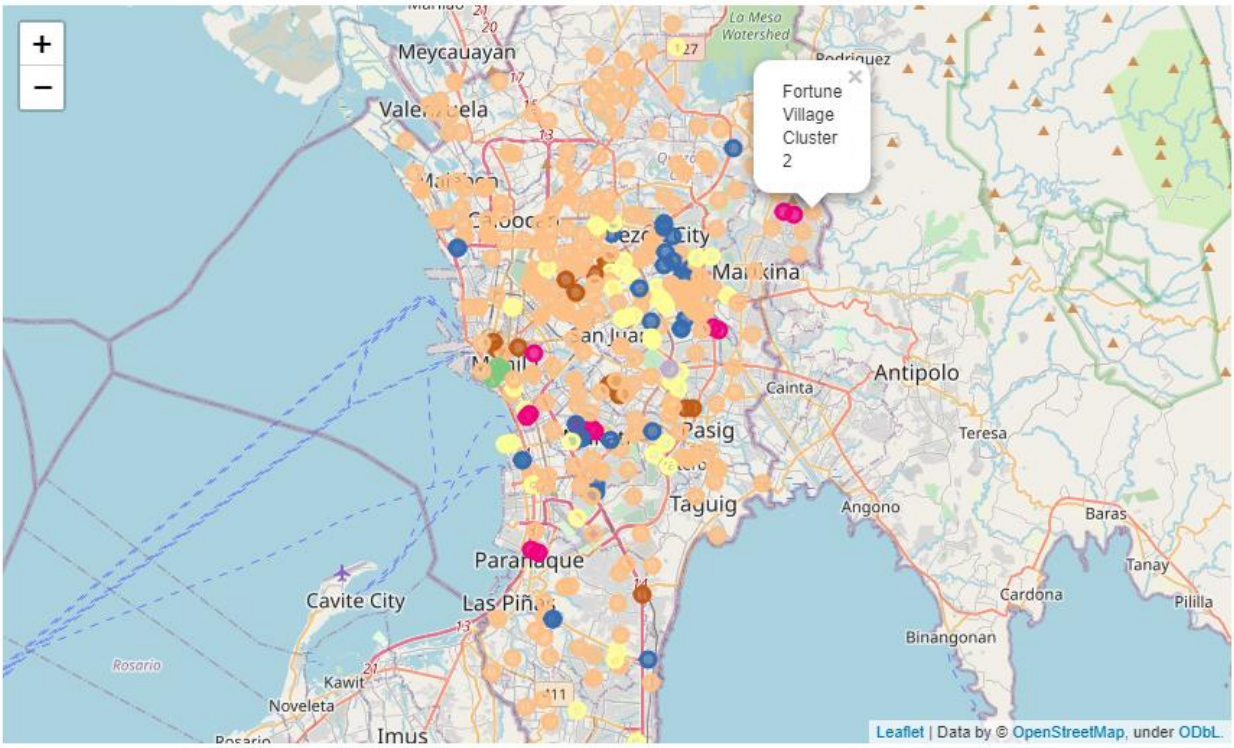

Figure 9. Map showing the different neighborhoods in Metro Manila. The different colors show the cluster those neighborhoods are part of. Notice the significant number of Cluster 2 neighborhoods.

Bar graphs of the different clusters were produced to show the features that are unique to the other clusters. Cluster 1 shows a significantly high count for Arts, Culture and Museums thus it was labeled “Arts, Culture and Science District”. (Figure 10) Cluster 2 shows high count for city essentials with relatively high count for the other venue types thus it was labeled “Downtown Church, Hospital or School Zone”.(Figure 11) Cluster 3 shows high count for vehicle/transport related, convenience stores and food essentials, tourist destinations, sports & fitness and fast food venue types and relatively high values for the rest of the venue types thus it was labeled “City Center”. (Figure 12) Cluster 4 shows high count for entertainment and gadgets, hardware, toys etc. venue types thus it was labeled “Entertainment and Shopping District”.(Figure 13) Cluster 5 shows high count for Business hub venue type thus it was labeled “Business District”. (Figure 14) Cluster 6 shows high count for residential venue types thus it was labeled “Residential District”. (Figure 15) Cluster 7 shows high count for City essentials venue type and lower count of the other venue types compared to cluster 2 thus it was labeled “Low profile Church, Hospital or School Zone”. (Figure 16) Figure 17 shows the bar graph with all the clusters plotted together.

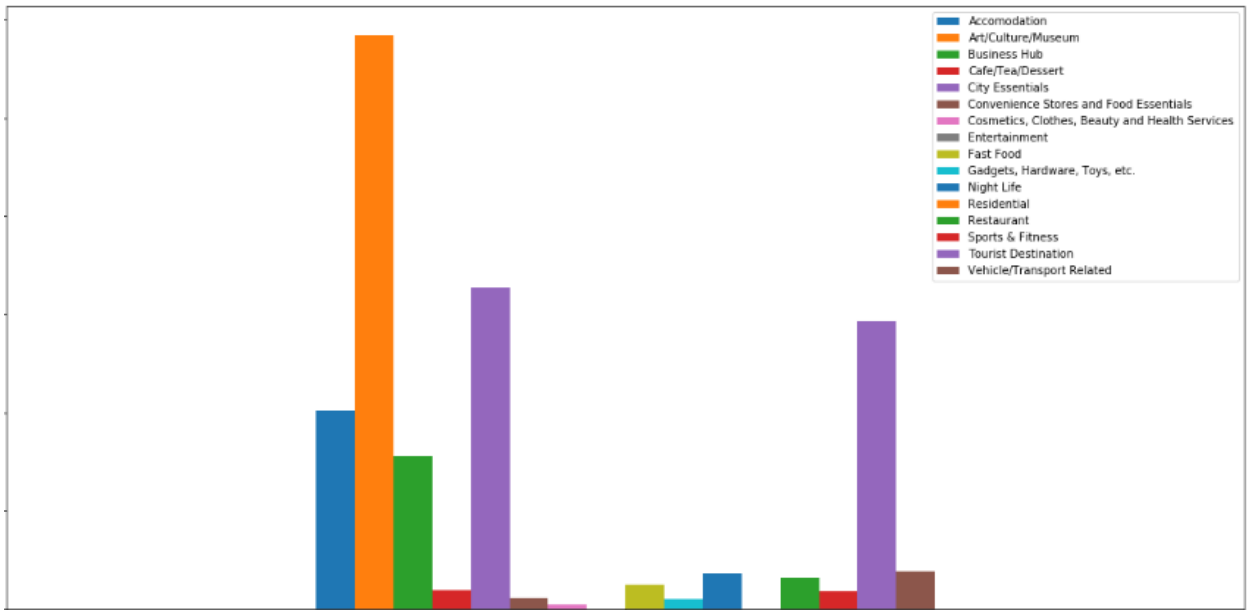

Figure 10. Cluster 1: Arts, Culture and Science District. Notice the high value for the Arts, Culture and Museum bar.

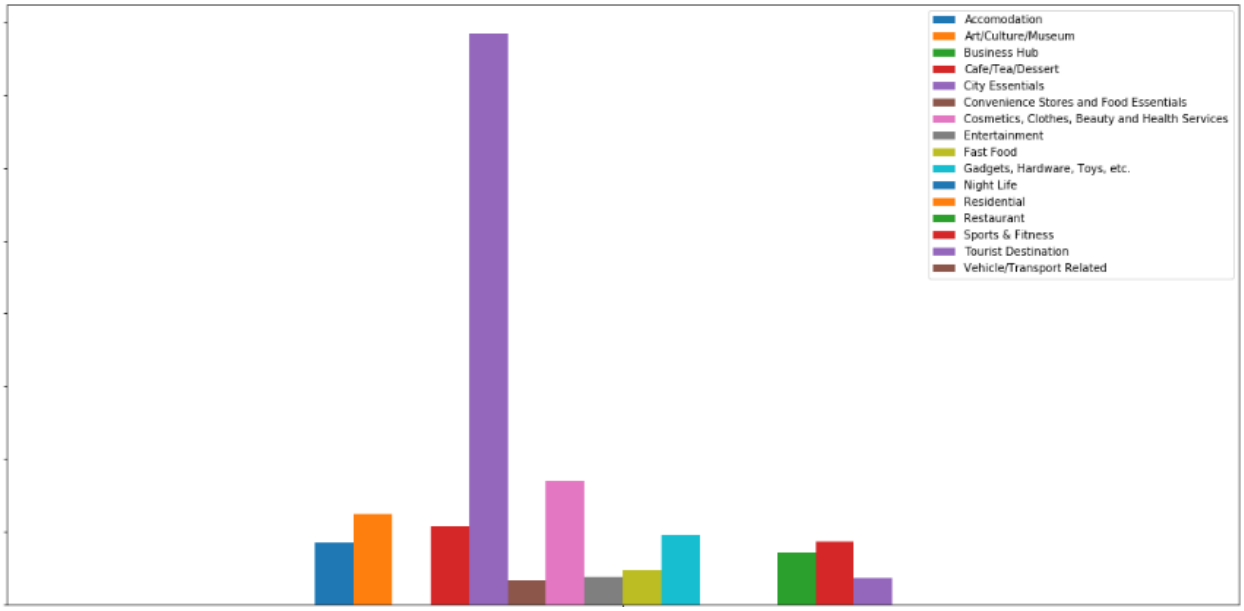

Figure 11. Cluster 2: Downtown Church, Hospital or School Zone. Notice the high value for the City Essential bar and higher bars for the other venue types compared to Cluster 7 (Figure 16)

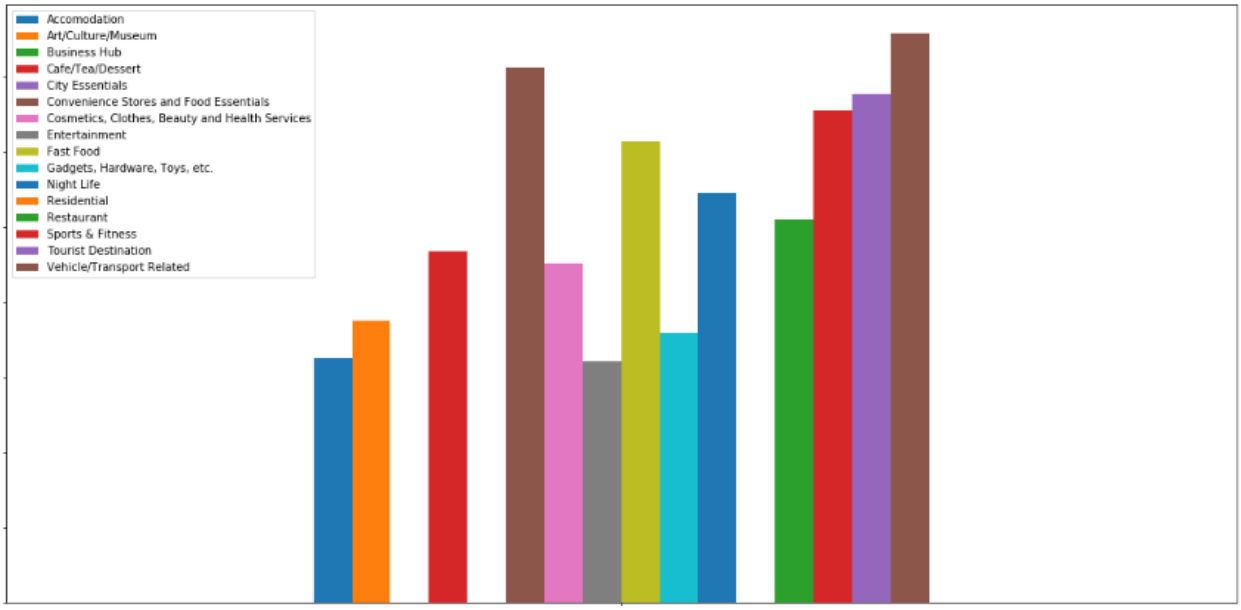

Figure 12. Cluster 3: City Center. Notice high count for most of the venue types but especially for the vehicle/transport related and convenience stores and food essential bars.

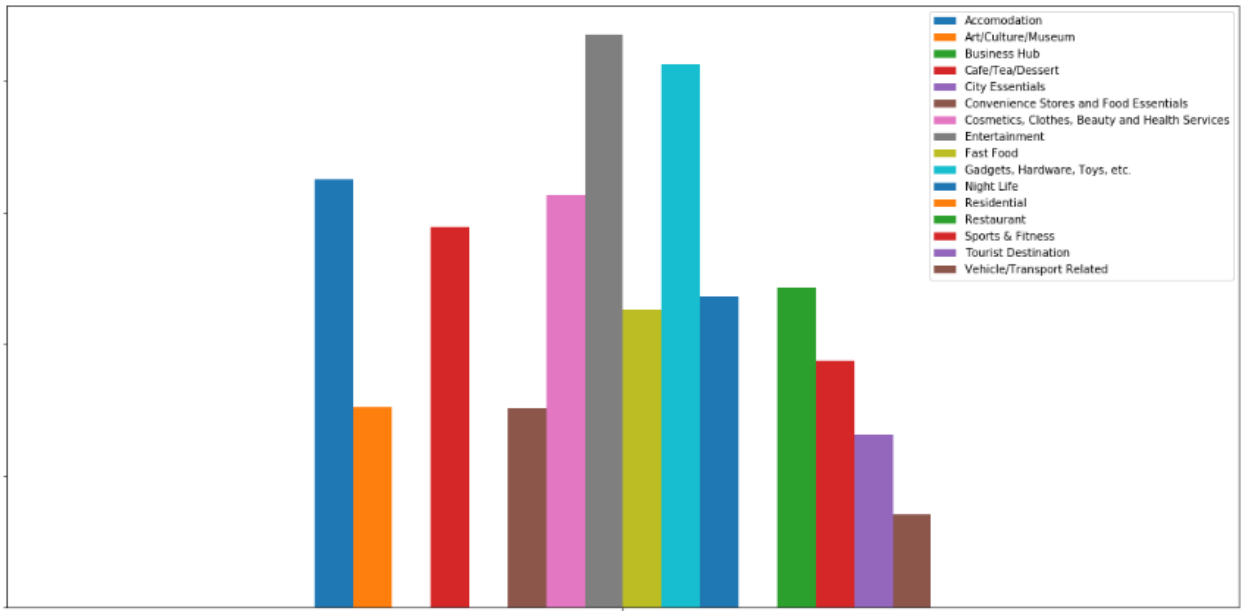

Figure 13. Cluster 4: Entertainment and Shopping District. Notice the high value for the Entertainment and the Gadgets, Hardware, Toys, etc. Bars.

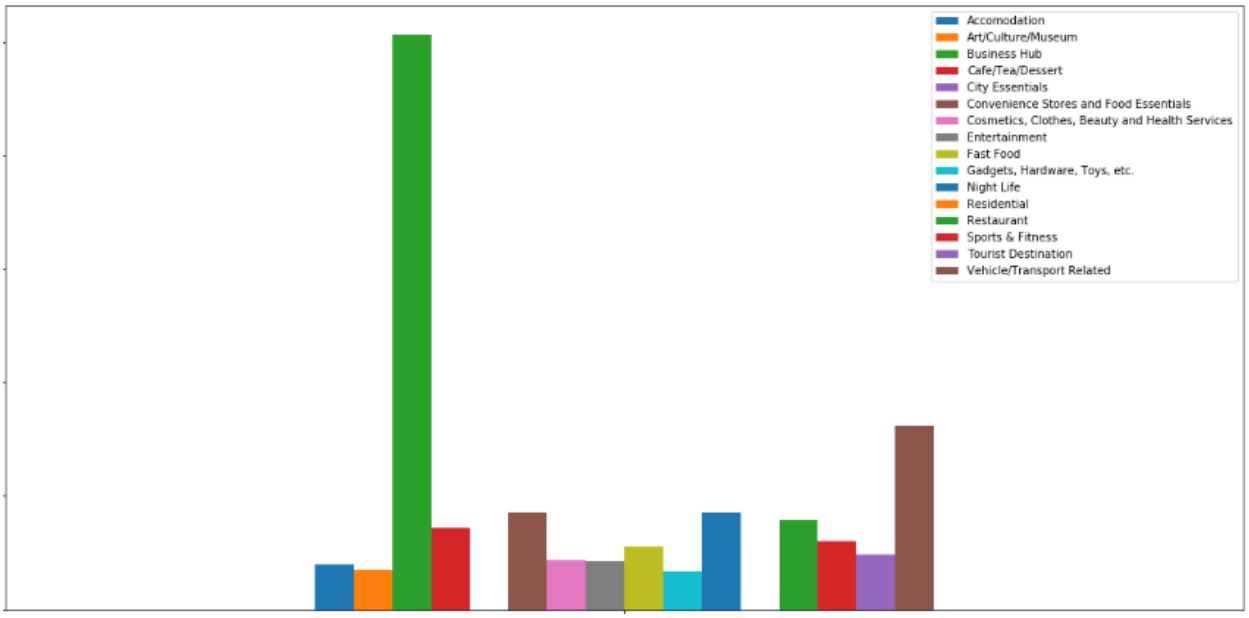

Figure 14. Cluster 5: Business District. Notice the high value for the Business Hub bar.

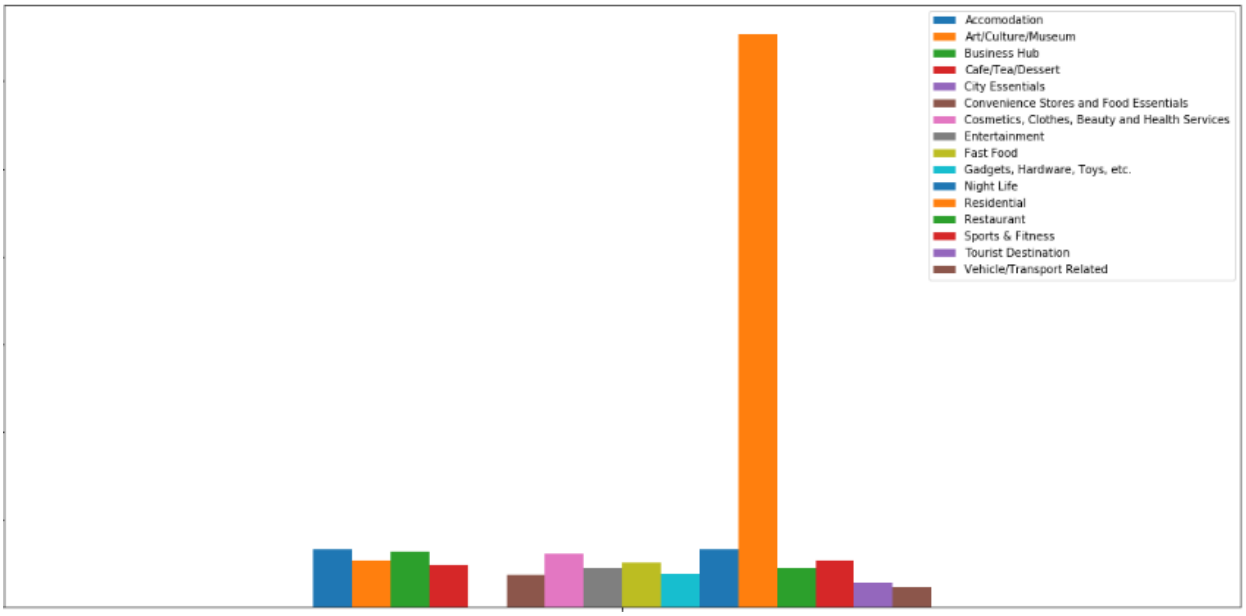

Figure 15. Cluster 6: Residential District. Notice the high value for the Residential bar.

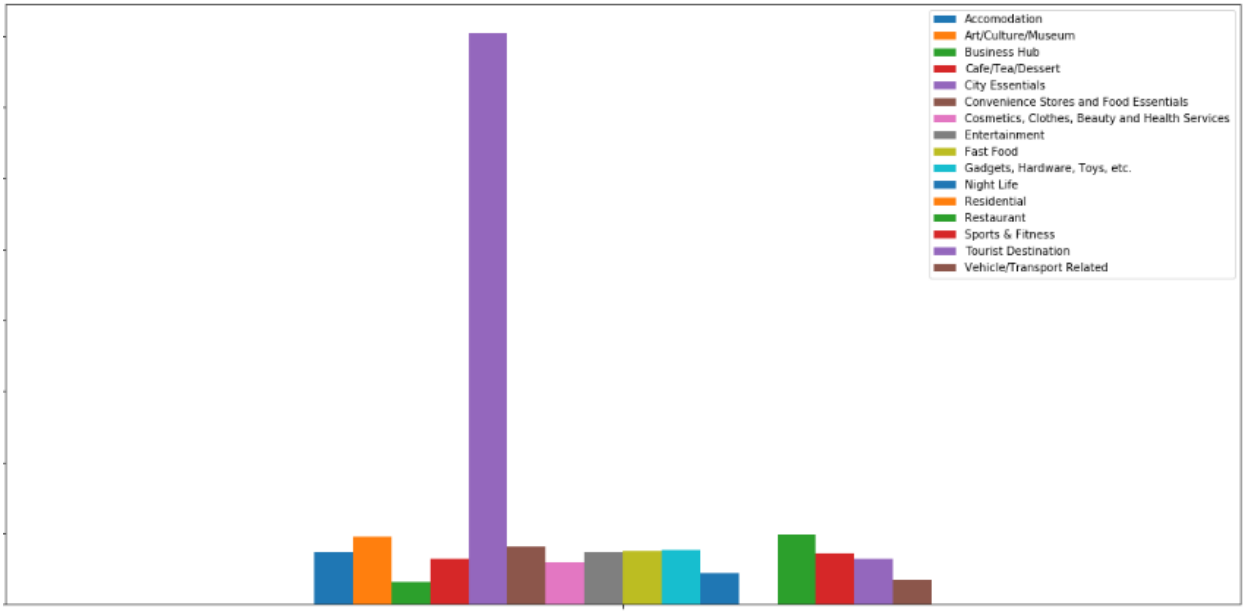

Figure 16. Cluster 7: Low Profile Church, Hospital or School Zone. Notice the high value for the City Essential bar and lower bars for the other venue types compared to Cluster 2 (Figure 11)

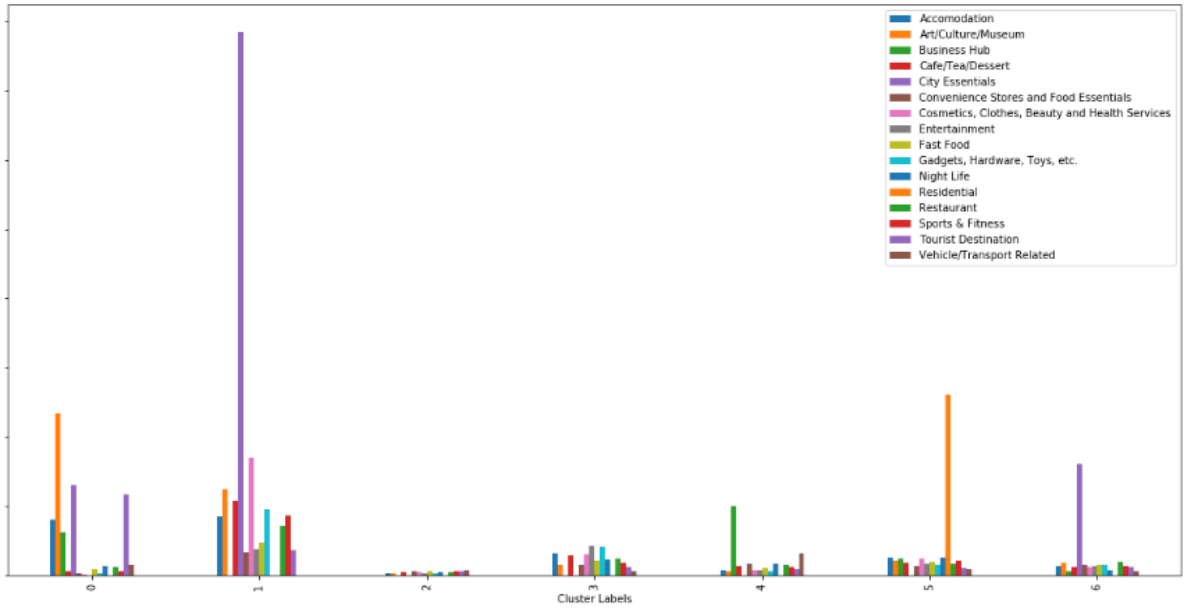

Figure 17. Bar graph that shows all the clusters compared to each other. Notice the difference of the heights of the bars of Cluster 7 (labeled 6) and Cluster 2 (labeled 1). Cluster 2 is categorized to be the Downtown Church, Hospital or School Zone due to its significantly higher bars compared to Cluster 7 which is categorized to be Low Profile Church, Hospital or School Zone

Given that Metro Manila is a metropolitan area, it can be expected that majority of the neighborhoods there would be urban areas with dense population of different businesses. This expectation is supported by the clustering model that was produced. Looking at the map of the different clusters, it is mostly dominated by cluster 3 which is the cluster we identified as the “City Center”. The summary of the seven clusters in Metro Manila is shown in Figure 18.

Cluster	Description
1	Arts, Culture and Science District
2	Downtown Church, Hospital or School Zone
3	City Center
4	Entertainment and Shopping District
5	Business District
6	Residential District
7	Low Profile Church, Hospital or School Zone

Figure 17. Summary of the seven clusters and their description

V. Conclusion and Recommendations

Businesses are competitive in metropolitan areas such as Metro Manila, thus information such as what type of neighborhood and what type of customer that neighborhood has is a valuable asset. Knowing what businesses and establishments are in a specific area helps business decisions of investors and business owners. Finding market inefficiencies in areas that have not been satisfied in terms of what products and services are provided there is supplemented by building a clustering model such as what was produced.

People who want to migrate to the different areas in Metro Manila can also be helped by using the information provided by the model. Knowing which areas match what they desire in terms of lifestyle is invaluable.

The Clustering model for Neighborhoods in Metro Manila can be further improved by changing the radius and limits of the venues in each neighborhood to its optimal values. A better dataset for the list of neighborhoods in Metro Manila as well as the venues found in those areas would show a more accurate visualization of the area. Lastly, addition of other features that describe the neighborhoods such as real estate prices or information of the people living there which would help produce a more accurate clustering model.

VI. References

[1] *"Presidential Decree No. 824 November 7, 1975". lawphil.net. Arellano Law Foundation. Retrieved April 15,2020.*

[2] *Census of Population (2015). "National Capital Region (NCR)". Total Population by Province, City, Municipality and Barangay. PSA. Retrieved April 15,2020.*

[3] "Presidential Decree No. 824". chanrobles.com. Chan Robles Virtual Law Library. Archived from the original on October 6, 2017. Retrieved *April 15,2020.*

[4] https://en.wikipedia.org/wiki/List_of_largest_cities

[5] https://en.wikipedia.org/wiki/List_of_ZIP_codes_in_the_Philippines