Thriving Communities of Dubai

Applied Data Science Specialization Capstone Project

Contents

1. Introduction to the opportunity

Dubai, an Emirate in the United Arab Emirates, is one of the world's most vibrant cities. Apart from being a tourist destination, it also acts as a financial and trading hub for middle east and even for the world.

To holiday or live or work in such a place, is both a privilege and a challenge posed by the plethora of choices that are available for both visitors and residents. It also poses a dynamic environment from a planning perspective for both, the city's administrators and investors intending to purchase or set-up businesses.

Data can play a key role in answering several such questions like:

"Which area has a good choice of UK curriculum schools"

"In which area of Dubai, can we find a high concentration of "Mediterranean Restaurants."

"Where should I open a New Mall"

"I have old parents, need to stay close to a good hospital."

"Which set of Communities do we need to focus on, for development."

We explore here a data driven approach to address 3 such business problems queries.

1. How are the **administrative areas** (known as **Communities**[1]) of Dubai compare to each other in terms of the social places in them, a common consideration for residents, investors of social establishments, advertisement placements etc.
2. How are the communities populated and what is the growth pattern?
3. What is the availability of Hospitals in the Communities especially with respect to the population in them? The target audience can be prospective residents, investors, and town planners.

---

[1] http://www.dubaifaqs.com/communities-in-dubai.php

Business Problem to Analyse :

Sl No.	Business Problem	Analyses	Extended Audience
BQ1	Which communities of Dubai should are most thriving for investing in residential properties and social establishments?	• Comparison of the no of venues of all categories in the Communities. • Population in the community	• Investors of social establishments (Restaurants, Gyms Café etc). • Media Companies for Advertisement placements. • Service Delivery Companies • Town planners.
BQ2	Are adequate hospitals available in all the communities?	• No of Hospitals in the Communities • Population in the community	• Prospective residents. • Healthcare investors • Town planners.
	Required for comparison in both analyses above.	• Comparison of Population density of the communities and their trend in population growth.	

2. Data

Dubai has a rich culture of open data made available by the Smart Dubai Office; a government entity entrusted to facilitate Dubai's citywide smart transformation. These datasets are available from their Dubai Pulse Website which acts as a clearing house of open data from various Dubai Government Departments. Apart from this we also used Foursquare APIs to get social or crowd sourced data for Dubai.

1. Dubai Admin Boundaries for Communities. Geospatial data (KML format) provided by **Dubai Municipality – Govt. of Dubai**. for the 2nd level second level of administrative divisions called Communities (**226** of them across Emirate of Dubai).

   https://www.dubaipulse.gov.ae/data/dm-location/dm_community-open

2. Dubai Population Estimate (2017, 2018 & 2019[2]) by Communities provided by **Dubai Statistics Center – Govt. of Dubai**.

   https://www.dsc.gov.ae/en-us/Themes/Pages/Population-and-Vital-Statistics.aspx?Theme=42

3. Foursquare Developers Access to venue data: provided by **Foursquare** through the **Places API** which offers real-time access to Foursquare's global database of rich venue data and user content to power location-based data requirements.
   - Details of top venues in and around a location.
   - Data of hospital locations identified by category – Hospitals, Medical Centers in Foursquare's Venues' Category hierarchy[3].

   https://developer.foursquare.com/docs/places-api/endpoints/

---

[2] 2019 : https://www.dsc.gov.ae/Report/DSC_SYB_2019_01%20_%2002.xlsx,
   2018 : https://www.dsc.gov.ae/Report/DSC_SYB_2018_01%20_%2002.xlsx.
   2017 : https://www.dsc.gov.ae/Report/DSC_SYB_2017_01%20_%2002.xlsx
[3] https://developer.foursquare.com/docs/build-with-foursquare/categories/

3. Methodology

a. Data retrieval, exploration, and wrangling

i. Communities of Dubai:

Geographic data of Communities the 2nd level administrative units of Dubai forms the basis of comparison for both BQ1 & BQ2. This data is utilized for:
a. Examining all the study parameters, community-wise
b. For generating the central coordinate for each community.
c. For calculating metrics like, population density, No of hospitals in a community etc.

The KML file downloaded from source is first converted to a GeoJSON file. Any common KML to GeoJSON converter (incl online[4]) can be used. Once the file is in GeoJSON format (which has the polygons for the communities), it is used for rendering in map and geometric calculations.

Unlike other cities / states, the Geo Coordinates (Lat, Lon) of each community was not available for scraping on the web. Hence the centroid function of GeoPandas was used to find the central point of each community from its polygon. Refer DataFrame below.

	COMM_NUM	SHAPE_AREA	SHAPE_LEN	lat	lon	Sector
0	333	1.984701e+06	7946.070175	25.225314	55.268768	Sector 3
1	121	8.707842e+05	5165.842443	25.287165	55.317958	Sector 1
2	282	1.269407e+07	14582.351508	25.230697	55.521222	Sector 2
3	353	2.430429e+06	6705.662879	25.175690	55.238309	Sector 3
4	736	1.305883e+07	15567.139691	25.024015	55.646574	Sector 7

Figure 1: Data frame of Community Centroids

The Communities marked at their Centroids can be visualized in the map below.

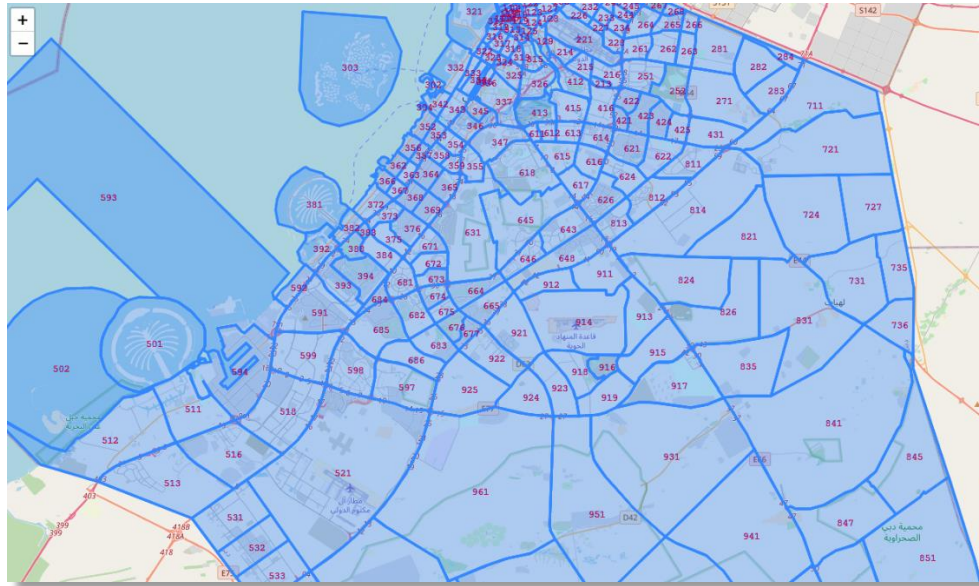[4] https://mygeodata.cloud/converter/kml-to-geojson

Figure 2:Map of Dubai Communities

ii.   Population Estimates:

Community-wise, population estimates from years 2017,18,19 were downloaded from source (Refer Section 2.2) as XL files. These are then combined to prepare a dataset showing the population over 3 years for each community.

The name of communities in the population data occasionally vary from those in the Communities Data. However, this is not a concern as the Community Nos are Unique and official IDs for the communities. Hence these can be used to relate, merge, or compare the datasets.

Based on exploratory data analysis and wrangling we determine that :
- There are a few communities (AL O'SHOOSH, AL LAYAN 2, HEFAIR) that show zero population even in the latest estimate year (2019). Based on our experience of residing in this city, we are aware that these areas are indeed low to no population, earmarked for future developments.
- In one community (AL YALAYIS 1) population in 2018 is nil which will be an issue for the population growth that we want to utilize. But it is clear that this community has been developed and settled in the year 2019 and is logically valid.
- There are several communities showing negative growth or decline which is examined and found to be logical based on real situation on the ground.

Above observations are kept in the purview of our results and conclusions.

Figure 3 : Communities (Shaded) with 2019 Population = 0 or Population Density 2019 less than 1/sqkm

The Statistics also show a dense cluster of communities where the population is concentrated (refer analysis on population density) with some outliers which are mostly areas that are being developed or planned for future (zero to very low population density).

The population data per community was also used as a basis for identifying the adequacy of healthcare facilities (Hospitals, Medical centers etc) in the communities.

iii.   **Venues Data (Social / Crowdsourced)**

Foursquare API is used to explore the no. of and types of venues in each community. We utilize the API for 2 purposes.

- **Details of top venues in and around a location**.
  Using the center of each community the API enumerates the most popular venues in different categories. Foursquare identifies over ?? categories grouped under 10 high level categories as follows.

Venue Category (High Level)	Foursquare ID
Arts & Entertainment	4d4b7104d754a06370d81259
College & University	4d4b7105d754a06372d81259
Event	4d4b7105d754a06373d81259
Food	4d4b7105d754a06374d81259
Nightlife Spot	4d4b7105d754a06376d81259
Outdoors & Recreation	4d4b7105d754a06377d81259
Professional & Other Places	4d4b7105d754a06375d81259
Residence	4e67e38e036454776db1fb3a
Shop & Service	4d4b7105d754a06378d81259
Travel & Transport	4d4b7105d754a06379d81259

Running the API for (api.foursquare.com/v2/venues) for each community (using 3.5 km [5]as the radius) we get the no. of popular High Level Venues for all 226 communities and save them in a dataframe as below.

	CNAME_E	COMM_NUM	lat	lon	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	AL BADA'	333	25.225314	55.268768	52	19	4	248	100	113	160	34	168	84
1	CORNICHE DEIRA	121	25.287165	55.317958	22	8	4	103	30	65	76	14	76	109
2	AL KHWANEEJ SECOND	282	25.230697	55.521222	0	10	0	9	0	7	5	1	5	5
3	AL SAFA FIRST	353	25.175690	55.238309	5	14	6	158	38	75	101	19	64	41
4	NAZWAH	736	25.024015	55.646574	0	0	0	3	0	6	1	0	4	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
221	JUMEIRA FIRST	332	25.229747	55.253889	23	13	5	190	107	93	121	27	76	87
222	MEREIYEEL	821	25.095226	55.512897	0	0	0	0	0	2	0	0	0	0
223	PORT SAEED	129	25.251006	55.332718	30	24	4	210	93	83	157	19	113	180
224	JABAL ALI INDUSTRIAL FIRST	599	24.999455	55.123519	0	2	0	18	3	3	47	4	11	6
225	WADI ALAMARDI	271	25.203261	55.491097	3	3	0	8	0	3	3	0	5	3

226 rows × 16 columns

Figure 4: No of High Level Venues in each Community

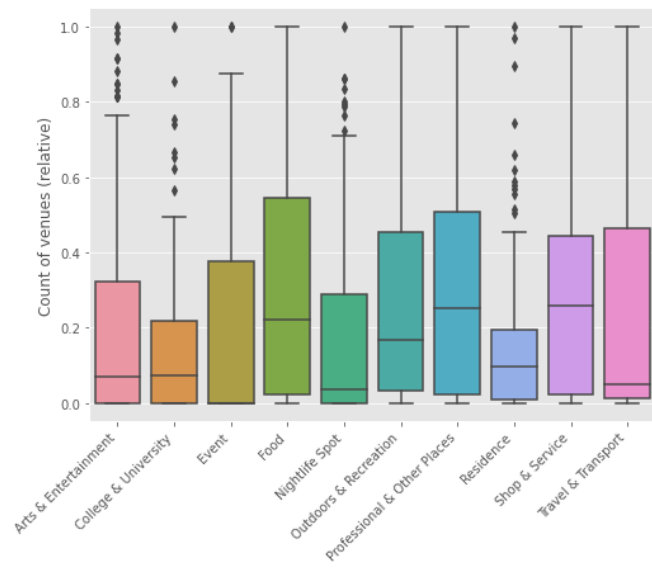The frequency each type of venue is visualized in the box plot below:

Figure 5: Distribution of Types of Venues in the Communities

---

[5] 3.5 km is taken based on approximate square root of average area of the community polygons.

- **Data of hospital locations identified by category.**

  Under the Hi-Level Venues category of Professional & Other Places, a 2nd level subcategory is available for Hospitals (Foursquare Venues Type IS 4bf58dd8d48988d196941735).

  We utilize this Venue type with the Foursquare API (api.foursquare.com/v2/venues/explore) for each community (using 8 km[6] as the radius around the centroid) to enumerate all the hospitals and its subcategory venues available in that area. Based on this broad search with a wide radius we get a large set of data which is then cleaned up for following issues.
  - ✓ Since we used a wide radius of 8 km, a number of venues are duplicated as the venues are picked up in the adjacent communities. This is eliminated by removing duplicates from the set using the Venue ID as the unique identifier.
  - ✓ Some of the subcategories (Veterinary, Buildings are not relevant for our study. Those are identified and dropped from the set.

Foursquare Venue ID	Name	Lat	Long	Actual Venue Type
570545e7498e955deec7a65d	Dubai Healthcare City	25.232924	55.32366	Medical Center (Locality)
5215c35011d2fdcad03d8e5a	Zabeel Veterinary hospital	25.210553	55.288277	Veterinarian (Vet facility)
4d38464dd60c6dcb3b5576bc	NMC	25.27490084	55.34970203	Office (Corp Office of Hospital Company)
4d449e267e2e5481545a628f	Al Shamsi Bldg	25.29202791	55.36881716	Building (Commercial Bldg)
5107bf93e4b09b1bbfa06164	Habtoor Business Tower	25.08562246	55.14167716	Building (Commercial Bldg)

  - ✓ Some of the healthcare facilities of adjacent Sharjah Emirate are also picked up due to the large radius search. They are dropped based on identifying the Community under which the healthcare venue falls as explained in next part.

  Using the set of healthcare venues identified, we do a point in polygon analysis to accurately identify which community the healthcare facility belongs to, by using the healthcare venue (point) in polygon (communities) method of geopandas . This way we are also able to eliminate the healthcare venues of Sharjah which were picked up earlier.

  Based on above steps, we finally arrive at a dataset of all healthcare venues (132) in Dubai communities, with their community identified. Refer dataframe below.

---

[6] 8 km is found to ensure that entire Dubai is covered by the scan and no gaps are left out between the circles.

	Venue ID	Venue	Latitude	Longitude	Venue Category	geometry	COMM_NUM
0	4dd3feece4cd6154148c0509	Iranian Hospital (المستشفى الإيراني)	25.230393	55.269144	Hospital	POINT (55.26914 25.23039)	333
1	5e2006c3e7117a0008c8038e	Fakih IVF - Dubai	25.221963	55.263826	Hospital	POINT (55.26383 25.22196)	333
2	514705f3e4b0da934ca8ef83	Al Noor Polyclinic	25.229366	55.273964	Hospital	POINT (55.27396 25.22937)	333
3	4bbc246bafe1b713fc21304b	Belhoul European Hospital	25.237432	55.275028	Hospital	POINT (55.27503 25.23743)	333
4	4bbc29d93de8c9b654b79aad	Medcare Hospital (مستشفى ميدكير)	25.183387	55.242802	Hospital	POINT (55.24280 25.18339)	353
...	...	...	...	...	...	...	...
127	5cb2d03ac97f2800396be0e9	King's College Hospital - Dubai Hills	25.112392	55.254775	Hospital	POINT (55.25478 25.11239)	631
128	51a1df79501987a82027628e	burjeel hospital for advance surgeries-tarabic...	25.146749	55.223495	Hospital	POINT (55.22349 25.14675)	364
129	4f9bf32be4b0de6267893b8a	Mercato Clinic	25.217103	55.253267	Hospital	POINT (55.25327 25.21710)	332
130	5bf4e4b51f8ed6002c643ec3	Mediclinic Deira	25.251147	55.335197	Hospital	POINT (55.33520 25.25115)	129
131	50fe8907e4b03c6cc88da036	Al Tadawi Medical Center مركز التداوي الطبي	25.257071	55.333701	Hospital	POINT (55.33370 25.25707)	129

Figure 6: All Healthcare facilities with Community No.

b. Analyzing population data based on communities.

Apart from analyzing the population data itself, it was cleaned up and structured to form the basis for other queries under study.

- The population data was merged with the community data (using the community ID as the linking field) to get community area, location etc information and population data in one dataframe.
- The calculated column for population density (community population / community area) of the latest year (2019) was added. (Pop2019*1000000/SHAPE_AREA Pop per SqKm)
- The calculated column for population growth of 2019 from 2018 was added. (Pop2019/Pop2018 – 1).
- We add in a dummy Value of 9999 to the population growth at row for community AL YALAYIS 1, as that community had zero population in 2018 and 1114 in 2019 and was showing Infinite population growth.
- Other Communities which had zero/zero growth are left with values of population growth = zero.

Finally, the resulting dataframe is **describe**d as below.

	COMM_NUM	SHAPE_AREA	SHAPE_LEN	lat	lon	Pop2017	Pop2018	Pop2019	Pop2019density	PopGrowth19by18
count	226.000000	2.260000e+02	226.000000	226.000000	226.000000	226.000000	226.000000	226.000000	226.000000	226.000000
mean	490.694690	2.105540e+07	15928.212597	25.119813	55.327693	13170.154867	14125.110619	14849.115044	6712.612029	603.813035
std	261.874728	4.042519e+07	14509.579155	0.146898	0.146786	24155.353167	24297.850367	24883.566435	15412.114960	6923.760207
min	101.000000	7.308970e+04	1397.451420	24.660479	54.904850	0.000000	0.000000	0.000000	0.000000	-42.164019
25%	283.250000	2.366359e+06	6779.850625	25.032402	55.239103	314.000000	419.250000	637.250000	23.334105	0.000000
50%	392.500000	6.224066e+06	10883.227338	25.153611	55.319806	4751.500000	5860.500000	6586.000000	1119.787975	5.303471
75%	680.000000	1.921986e+07	19299.326017	25.242531	55.403908	13406.000000	16723.500000	17811.500000	4560.645629	9.090909
max	991.000000	3.650142e+08	122334.492899	25.338912	56.139180	202918.000000	197838.000000	196316.000000	111380.654928	99999.000000

Figure 7:Dataframe of Population Data

The following key insights were developed based on analysis of the Population data.

i. Population Growth:

Considering the wide span of the population growth amongst the communities (from -42% to 29,250% *(next lower value below the infinite growth)*). Population growth was binned into 6 groups with corresponding no of communities falling in them.

Bin Name	No of Communities	Description
> 100	6	Greater than 100%, Very High growth
20 < x <= 100	8	20% to 100%, High Growth
2 < x <= 20	131	2% to 20%, Medium Growth
-2 < x <=2	77	-2% to 2% , Negligible growth/decline
-20< x <= -2	3	-20 % to - 2% Medium Decline
x <= -20	1	Less than -20% High Decline.
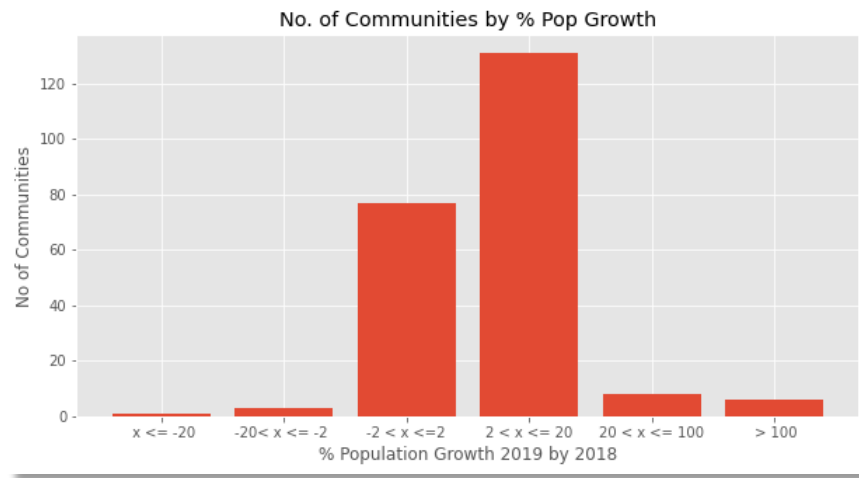
Figure 8: Community wise Population growth in Dubai
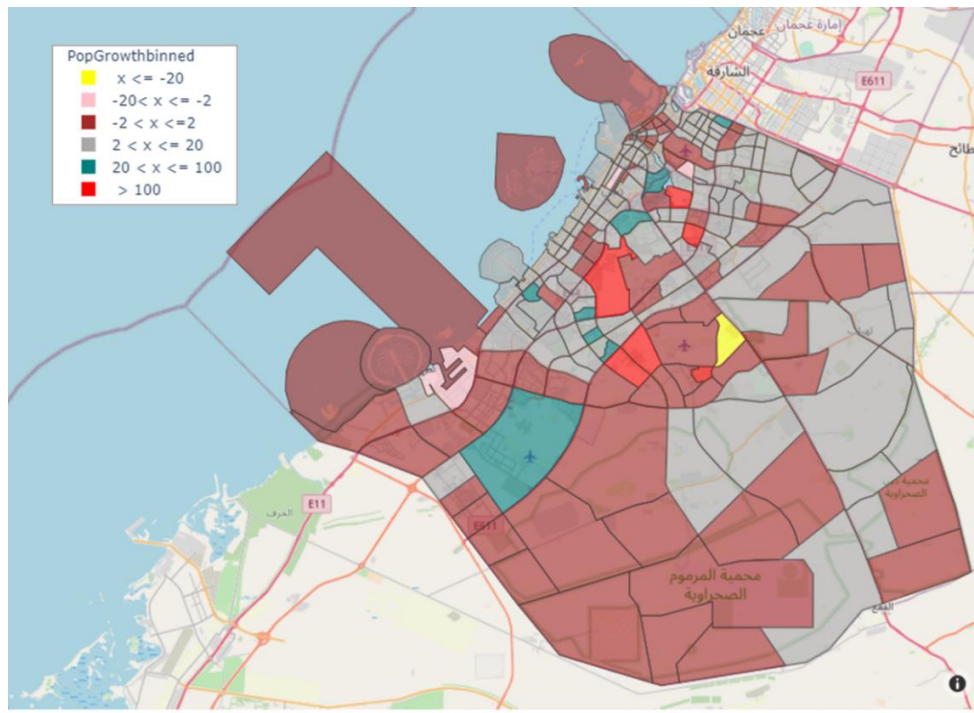
Figure 9: Choropleth of Population Growth categories

ii.   Population Density:

Considering the wide span of the population density amongst communities (from 0 to 111k /SqKm), Population density was binned into 6 groups with corresponding no of communities falling in them.

Bin Name	No of Communities	Description
> 50k	6	Greater than 50k/sqkm, Very High Density
10k < x <= 50k	30	10k to 50k / sqkm, High Density
1000 < x <= 10k	79	1k to 10k / sqkm, Medium Density
10 < x <1000	61	10 to 1000 /sqkm, Low Density
0< x <= 10	47	0 to 10 / sqkm. Very Low Density
x = 0	3	0, Not Populated.
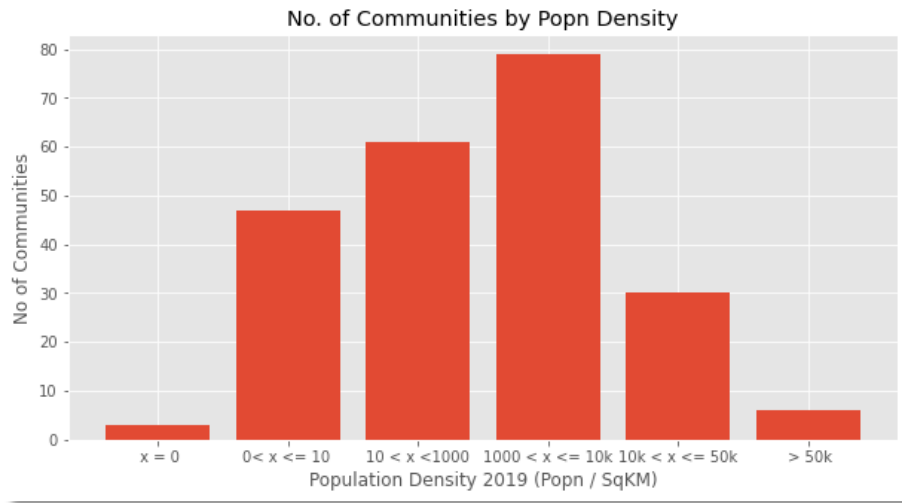
Figure 10 : Community wise Population Density of Dubai
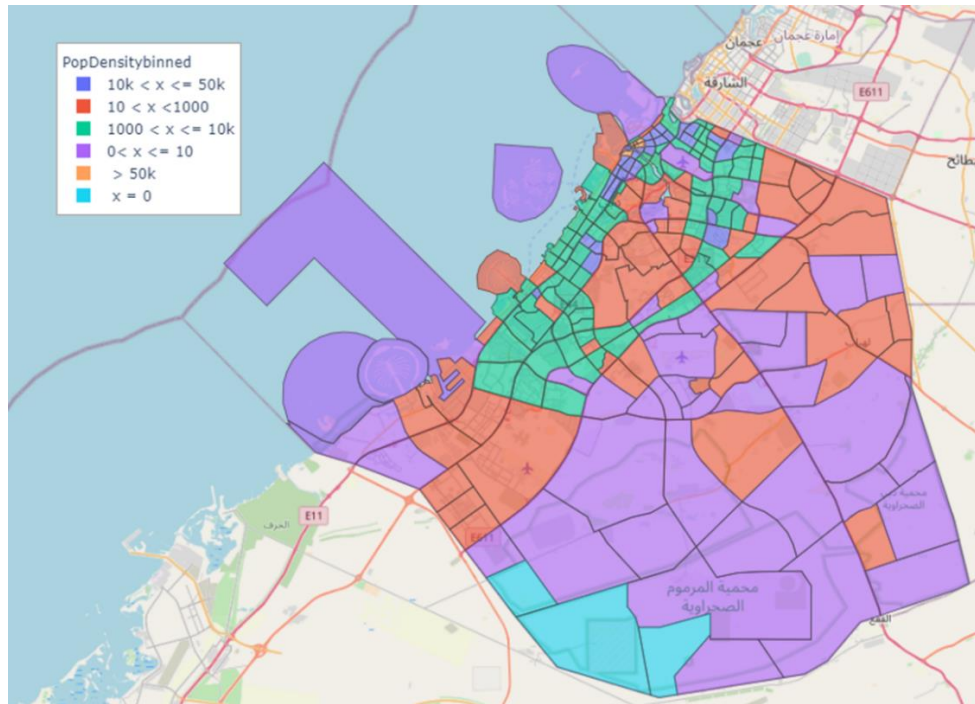
Figure 11: Choropleth of Population Density categories

iii. Population Growth vs Population Density:

A heatmap of the Population Growth vs Population Density categories was prepared to identify various types of communities and to examine how many communities are there in various combinations of growth and density. It was observed that the maximum no of communities ie 67 of 226 communities (~30% of them) are Medium Population Density (1k to 10k / sqkm) & Medium Population Growth (2% to 20%). This heat map can be further utilized to identify other target community groups like High Growth – Low Density (6 communities) which will be of interest from a futuristic investment. The Heatmap prepared is shown below:
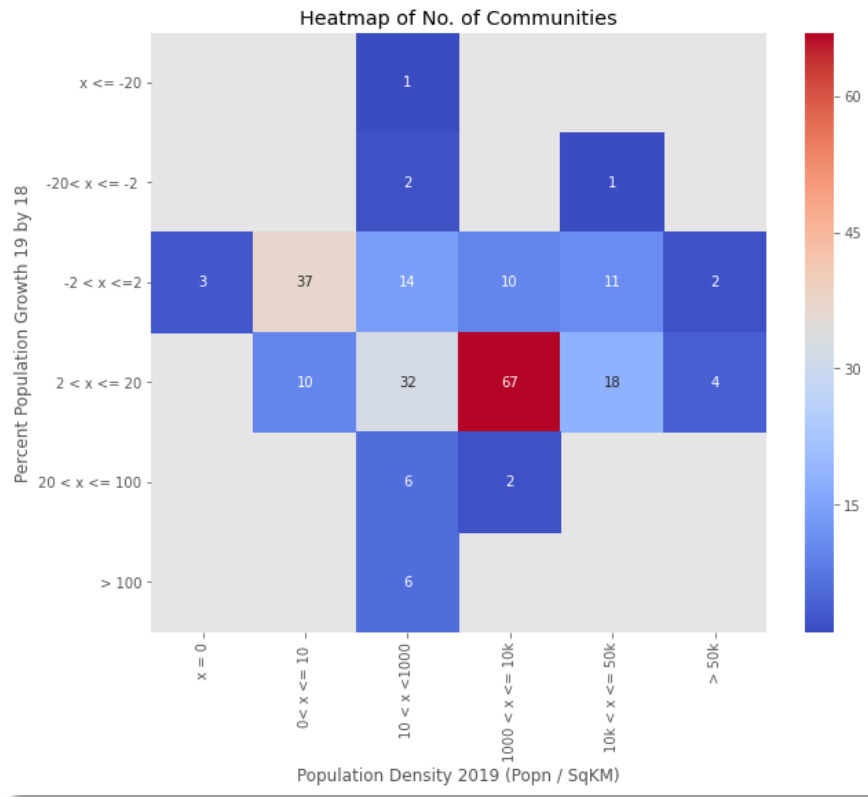
Figure 12: Heatmap of Communities' Population Growth & Density charateristics

c. *K-means clustering to segment communities.*

Performing K-means clustering algorithm to segment the communities based on the type of venues found on Foursquare.

To group the communities to answer our business problem, we utilize K-means clustering ML algorithm. It essentially segregates the overall set of communities (as seen in Figure 4: No of High Level Venues in each Community) into separate clusters which have a distinctly different pattern of distribution of the venues. The algorithm is unsupervised and hence we first need to arrive at the right number of desired clusters.

Selection of number of clusters:

We iterate the K-Means clustering algorithm for number of desired clusters (going upto 6) to observe (using boxplots) the clustering results. NOTE : **we could have utilized the Elbow Method or Silhouette Method too**, but the iteration gives a easy to interpret visualization for optimal K.
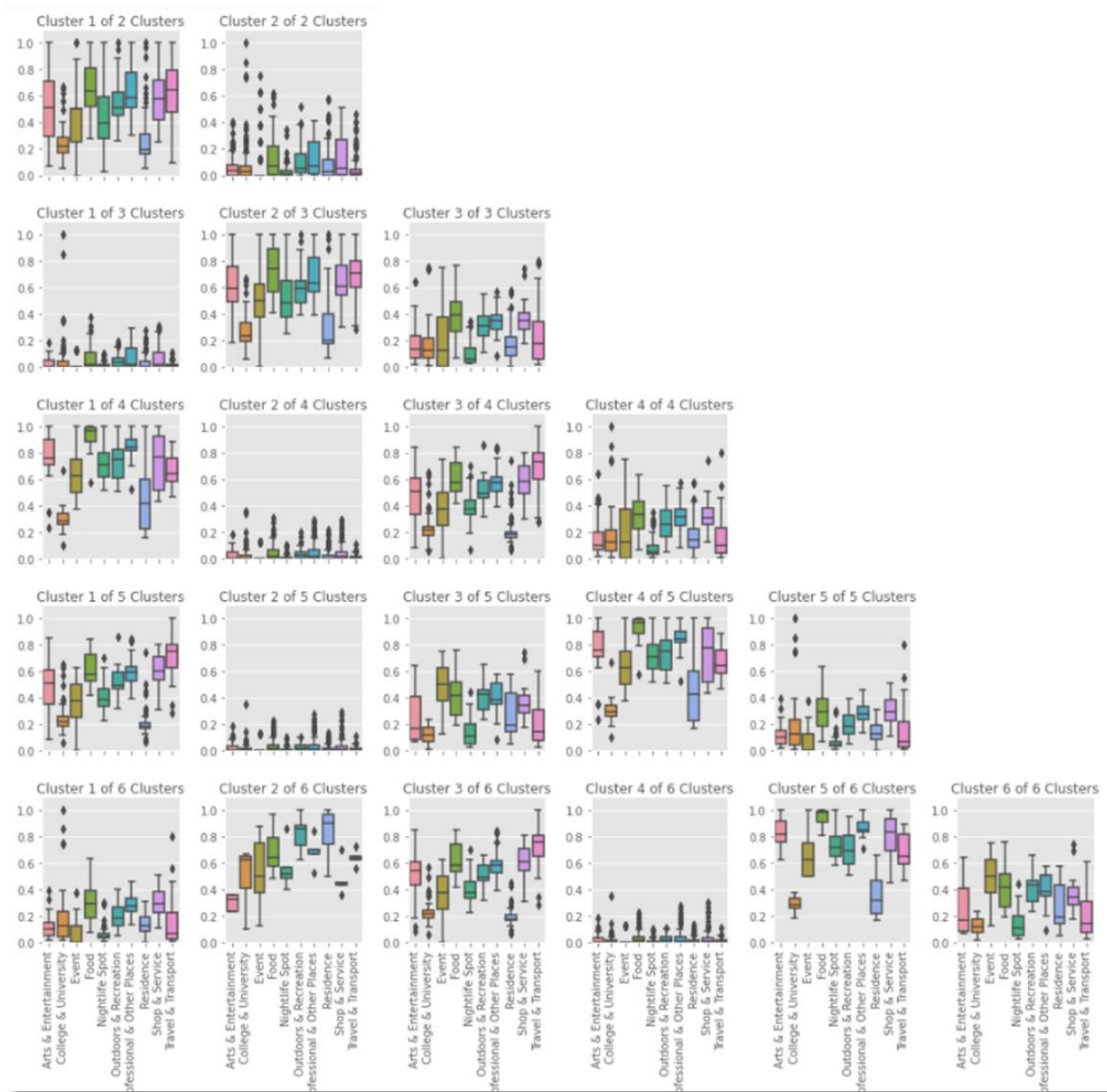

Figure 13 : K-Means Clustering Iteration from 2 to 6 Clusters.

From the iteration we observe that we get a good resolution upto a max of 4 clusters beyond which differentiating between 2-3 clusters within the set has no meaningful interpretation. Hence, we chose 4 clusters to study (3rd Iteration set in above figure).

ii.   Clusters and their interpretation:

Clustering Results with 4 clusters are as below.
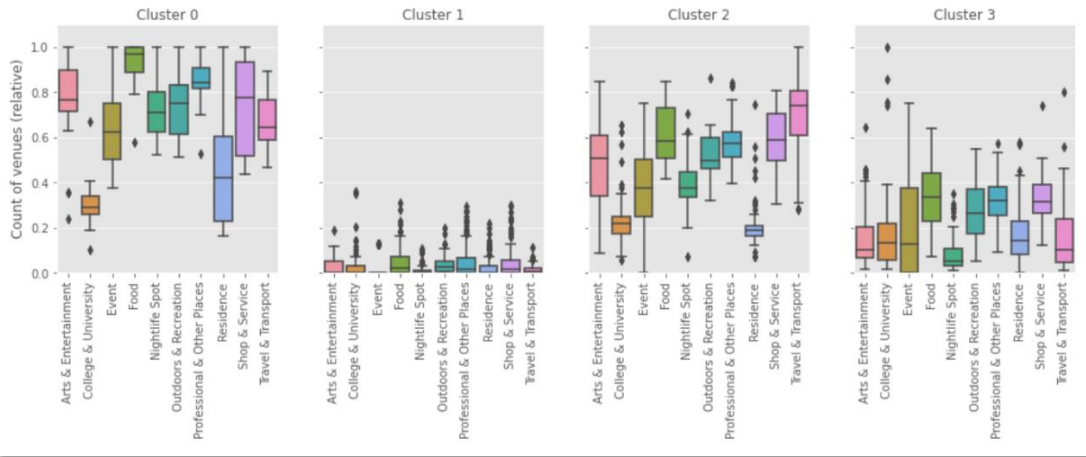
Figure 14 : Dubai 226 segregated into 4 clusters.

The 4 Clusters are be interpreted as below:

Cluster ID	No of Communities	Characteristics	Interpretation
Cluster 0	19	Has high frequencies for most venue categories except College & Univ. and Residence.	These communities are the most diversely developed part of Dubai.
Cluster 1	101	Has low frequencies for all venue categories.	These communities are the most under-developed part of Dubai.
Cluster 2	41	Has high to moderate frequencies with high occurrence of Art & Entertainment, Travel & Transport and Shops & Services	These are the areas involved in tourism, markets and travel.
Cluster 3	65	Has moderate frequencies with Food, Professional & Other places and Shops & Services being the most popular.	These are medium developed areas a little outside the central parts of the city.

Geographic distribution of the clusters:
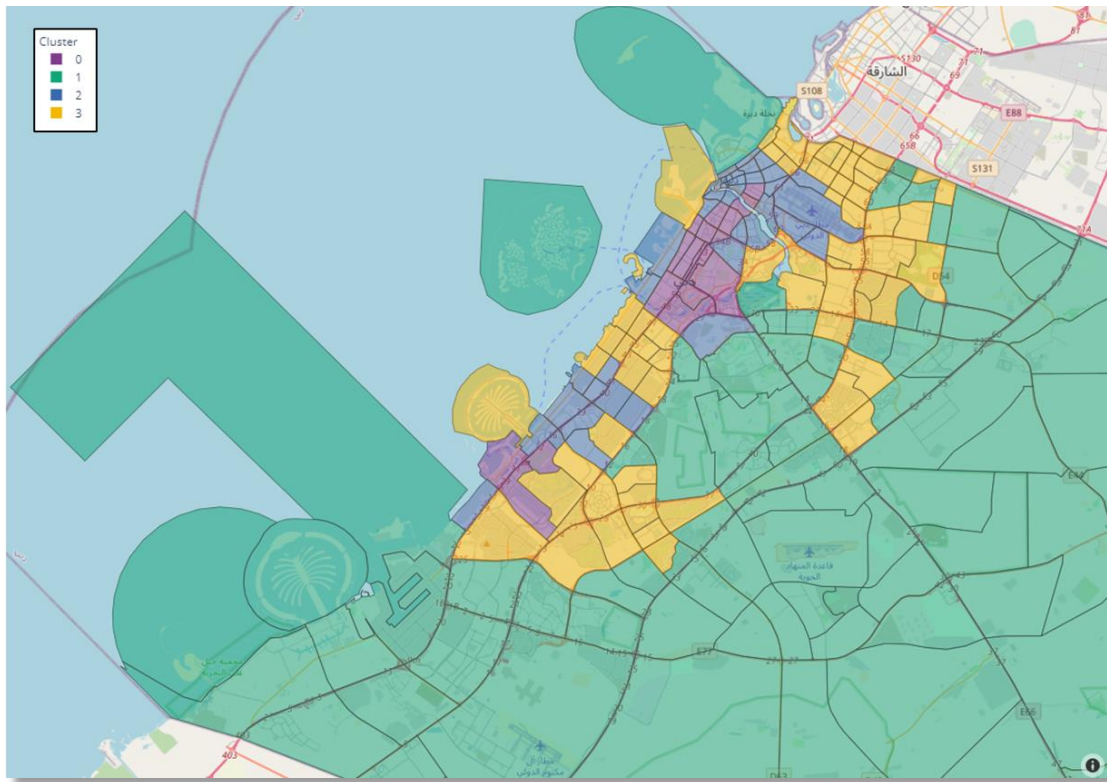
Figure 15: Clusters of Dubai Communities based on Venues.

The choropleth of the communities marked for the clusters clearly shows that :

- Development has been mostly focused in an around a few closely located areas.
- There are a lot of areas outside the main city which have very few venues or are not developed.
- Correlating with population growth data, we can establish areas which are poised for further growth.

d. *Analyzing Healthcare facilities with respect to population of the communities.*

We derive a data set of **132 healthcare venues** across the Dubai communities and identify which communities these fall under. A location map of these facilities are shown below with the communities demarcated.

This gives a clear view of how well the healthcare facilities are distributed among the communities.

i. No of Healthcare venues in a Community:

We analyze the no. of healthcare venues that are present in each community of Dubai. It shows 2 relevant points.

- There are a large number of communities where there are no healthcare venues. It is to be noted that most of them are also communities with zero or very low population which is validated based on the population study done in section 3.b.

- There is also one community where there are 21 healthcare venues. Based on local knowledge we can validate that this is the Umm Hurair 2 Community which houses the Dubai Healthcare City a dedicated economic free zone for healthcare facilities.
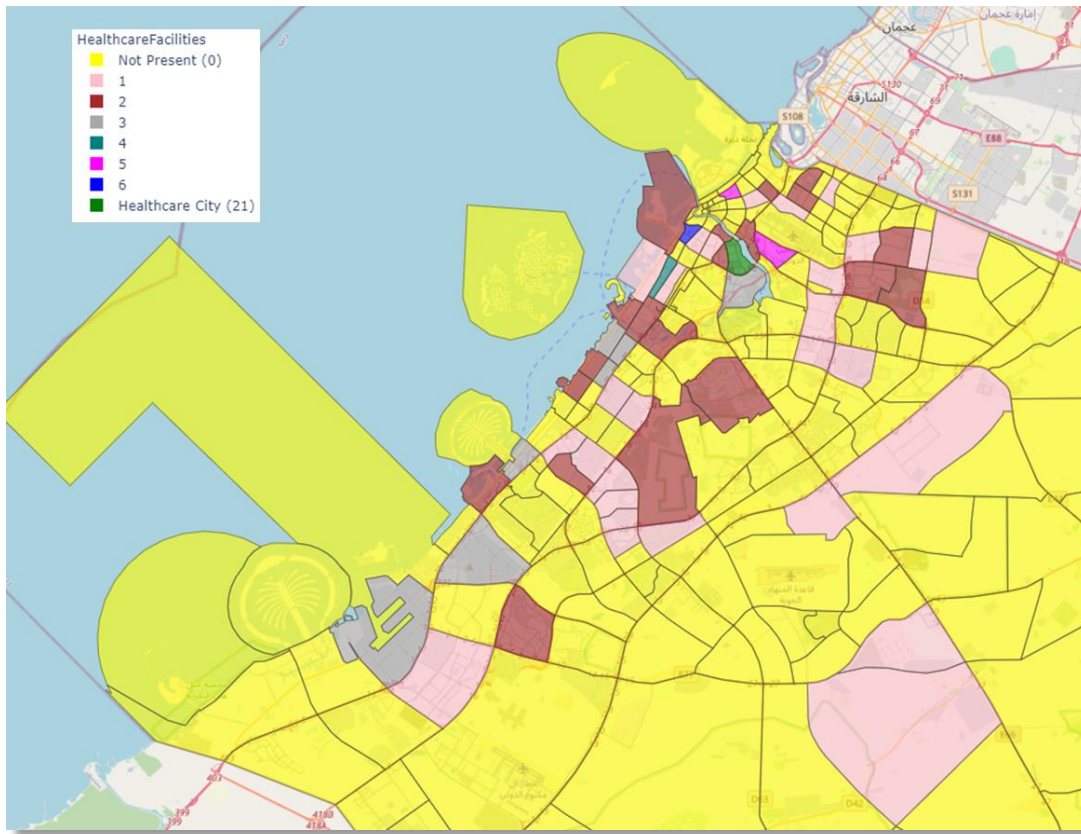

Figure 16: No of Healthcare Venues in Each Community

ii. Population being served by a Healthcare facility for each community:

We analyzed the Population in each community vis a vis the no of healthcare venues in it by checking for Population (in thousands) per healthcare Venue. Prior to this we remove all areas which have no healthcare venues as found in earlier section. We also remove the Umm Hurair 2 community as it is an outlier with a large no of healthcare venues.
- The resulting choropleth of Population(x1000) per Healthcare Venue gives a clear picture of the communities which are well served or underserved.
- It is to be noted that on an average basis, Dubai shows about 25k / healthcare venue (3.5 M Population in 2019 / 132 Healthcare Venues)
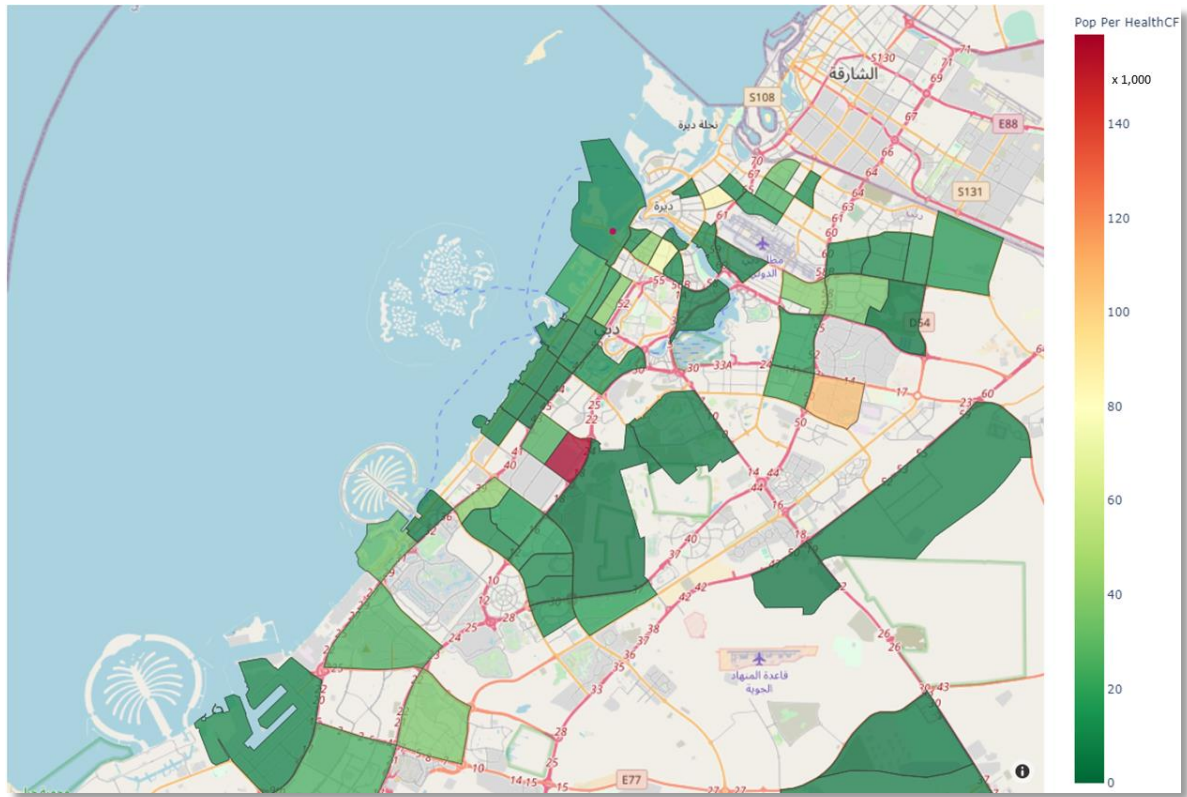
Figure 17: Population (x 1000) served per Healthcare Venue

4. Results

The results of the study have been derived under relevant points in the methodology section and are summarized below.

a. Population:

- Dubai is a has a potential for growth with a many communities with very low population which are earmarked for future developments.
- The Statistics also show close groups of communities where the population is concentrated.
- Population growth in the communities is varied but over 64% (145/226) communities show 2% to 100% yearly growth, an open opportunity for anyone looking at planning in the city.
- Population density is very varied and shows that in about 48% of communities (108/226) the population density is below 1000 persons /sqkm and population can be expected to go up in the future.
- It was observed that when both dimensions of density and growth are considered together, the maximum no of communities ie 67 of 226 communities (~30% of them) are Medium Population Density (1k to 10k / sqkm) & Medium Population Growth (2% to 20%).

b. High Level Venues

- 4 distinct type of communities can be identified based on the clustering of similar types of venues identified on Foursquare.
- These clusters show some amount of geographical contiguity, ie groups of communities adjacent to each other show similar venue characteristics.
- Development has been mostly focused in an around a few closely located areas.
- There are a lot of areas outside the main city which have very few venues or are not developed yet.

c. Healthcare Facilities

- On an average basis, Dubai shows about 25k people / healthcare venue (3.5 M Population in 2019 / 132 Healthcare Venues)
- There are many communities where there are no healthcare venues but most of them are of zero or very low population, hence not a reason for concern at the moment.
- The Umm Hurair-2 Community houses the Dubai Healthcare City a dedicated economic free zone for healthcare facilities which has a large no. of healthcare venues, which caters to the broader needs of entire Dubai.
- There are 2 communities (Al Quoz Industrial Area-2 and Warsan-1 aka International City) both known to be densely populated areas and have one healthcare venue each, thus showing high number of people / venue (over 80k).

Bringing it all together to answer the business questions we observe that

Sl No.	Broad Business Problem	Results used.	Rationale
BQ1	Which communities of Dubai are most thriving for investing in residential properties and social establishments?	• Communities in Cluster 3 • Medium Population Density (1k to 10k / sqkm) & High Population Growth (20% to 100%).	Cluster 3 communities are medium developed areas a little outside the central parts of the city and have less venues, thus providing opportunity for growth at relatively less competition. Medium population density and high growth also mean that the future prospects have a very high potential.
BQ2	Are adequate hospitals available in all the communities?	• 2 communities showing high population served per healthcare facility	The communities have a high population. They provide housing for a lot of workers from the service and industrial sectors. The communities will be well served by additional healthcare facilities.

5. Discussion and Recommendations

Our goal was to answer the broad business problems by establishing a data driven approach. Based on the publicly available free data, the methodology was developed and executed to arrive at a fairly well-organized response.
This approach can be further refined with additional data being made available through commercial means thus increasing the confidence level by a greater extent.

Some of the areas where we felt better data would have been useful are :

• Foursquare data showed some inaccuracies at venue level ie some known sub category level were found to be missing and location of some of the venues were wrong.
• The population data was only available till 2019 and 2020 data would have given a more current picture.
• For healthcare venues an estimate of the capacity of the venues was not available and hence we had to assume that the venues on an average cater to a similar number of patients which may not be true thus needing more refinement.
• Additional data like employment, real estate prices, transportation etc can play a major role in driving the answers to the business problems being addressed. These data should also be examined for improving the confidence of the conclusions.

Having said that, the approach laid out in the study still remains valid for any additional data which can made available in the future to update and refine the results.

6. Conclusion

The communities of Dubai are analyzed utilizing a combination of datasets from Dubai Municipality, Dubai Statistics Center and Foursquare to answer business problems that are of key interest to this thriving Emirate of Dubai. We discover, process and statistically lay out the basis for our recommendations to the business queries using data driven tools like statistical analysis, machine learning, data visualization etc. A well structured process is laid out which can lend itself to additional, more current and further complete datasets to refine / update the results in the future that can minimize the risk to investors, residents, planners, businesses etc who can utilize this approach.

7. Table of Figures