Analysis on the Transportation Infrastructure Extension in Dubai

Author: Cesar Hanna

Date: March 3, 2021

Table of Contents

Introduction

Dubai is the most populous city in the United Arab Emirates (UAE) and the capital of the Emirate of Dubai.

Located in the eastern part of the Arabian Peninsula on the coast of the Persian Gulf, Dubai aims to be the business hub of Western Asia. It is also a major global transport hub for passengers and cargo. Oil revenue helped accelerate the development of the city, which was already a major mercantile hub. Dubai's oil output made up 2.1 percent of the Persian Gulf emirates economy in 2008. A centre for regional and international trade since the early 20th century, Dubai's economy relies on revenues from trade, tourism, aviation, real estate, and financial services. According to government data, the population of Dubai is estimated at around 3,400,800 as of 8 September 2020.

The Challenge

With the fast growing economy and the influx of expats to work in all sectors, Dubai has become quite populated. However, currently commuting to work, to school and/or to venues is heavily dependent on either cars or taxis. There is a metro that connects few areas, but it is not covering enough regions that people can rely on; same goes to the bus system that they have there but not enough coverage as well.

One of the major worldwide events is taking place in Dubai and that is Expo 2020; Dubai Expo 2020 is a hot topic, due to the increasing pressure that will be placed on Dubai's transport systems and existing infrastructure. The aftermath of the event is also talked about. in terms of how the construction work taking place to accommodate the event will impact Dubai in the ensuring years.

This is creating few problems to name few:

- Financial: Commuting by taxi can be expensive. A lot of families cannot afford that.
- Infrastructure: Infrastructure congestion is hindering the mobility which in turn can have an impact on businesses and on the long run the economy.
- Environmental: The United Arab Emirates are a contributor to greenhouse gas emissions, listed as having the 29th highest carbon dioxide emissions. Since the boom of the oil industry occurred in the early 21st century, the population and its consumption of energy have sharply increased.

Description of the Solution

One effective way to solve these problems is to improve the transportation system by extending the metro system and building obviously few other underground stations.

For the business decision to be effective on where to build those stations, I have considered couple of important features such as population, venues that also include schools and hospitals and nearest metro station. More details on those features will be explained in the next sections.

Description of Data

The data frame includes independent variables or features that will be used to cluster the communities accordingly. These features are:

- Community Number
- Community Name
- Population
- Number of venues
- Latitude
- Longitude
- Nearest walking distance MS from the centre of the community (Km)

Data References

The data will be extracted from:

- Wikipedia: to get all the community information
  - https://en.wikipedia.org/wiki/List_of_communities_in_Dubai
- Google Maps: to get the metro stations walking distances
  – https://www.google.com/maps
- Google: to get coordinates – https://www.google.de/
- Other site(s) to get
  coordinates: https://www.distancesfrom.com/ae and https://vymaps.com/AE
- Foursquare API: to get the venues

Decision Criterion

Based on the data at hand and for the sake of analysis only (of course there are a lot of variables to be considered to take a decision for such a project, such as cost, environmental impact, available area and many more), I am going to use a group of variables, to be able to take a decision on where to extend the metro infrastructure and build new stations.

It will be based on certain logical criterion to follow, taking into consideration that the metro stations will be built in the centre of communities that are conforming to the criterion. The reason for this consideration is that these communities have a fairly small surface area thus building a station in the centre should solve the problem.

In real life situations it is not always the case that all 3 features, population, number of venues and nearest walking distance metro station, are satisfying the criterion, therefore if 2 criterion are met it should be enough to decide.

Threshold defined in numbers:

- Population: >= 1000
- Number of venues: a ratio of 15 venues for every 1000 people (15:1000); that means the number of venues should be >=15 based on the aforementioned population criteria
- Nearest walking distance MS from the centre of the community: >=1.7 Km

Data Frame

- Step 1: Importing and installing all the necessary libraries and packages
- Step 2: Scraping the first table from wikipedia and putting it in a data frame.
- Step 3: Creating the second data frame with the coordinates and metro station distance features.
- Step 4: Merging the first and second data frames so we can prepare the resulted data frame to incorporate the venues later on.
- Step 5: Getting the data from Foursquare showing the venues of each community, and creating the third data frame.
- Step 6: Creating “Number of Venues” column, and name the new data frame "communities_count_venues_updated".
- Step 7: Merging all 3 data frames and dropping the un-needed columns

Step 1: Importing and installing all the necessary libraries and packages

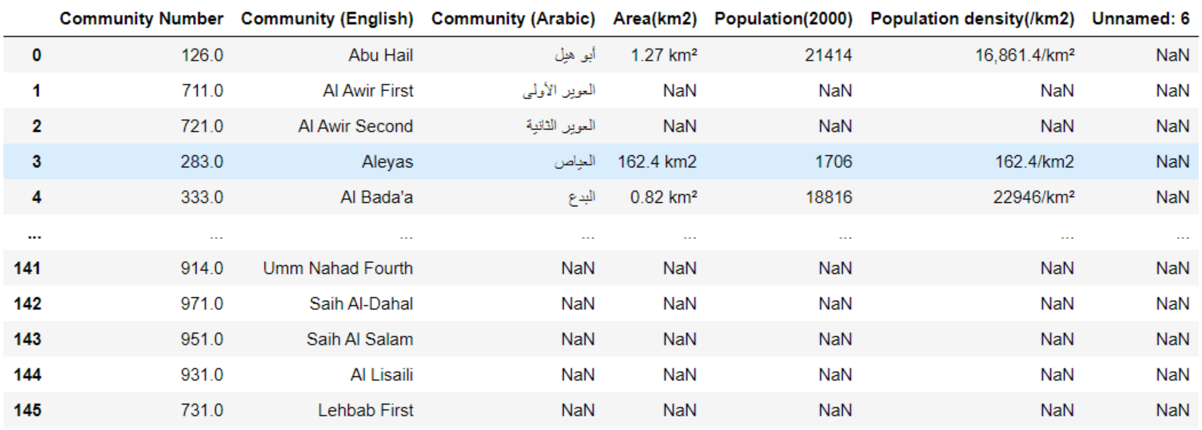
Here I imported the following packages and libraries that will be used in the exploratory data analysis and plotting:

- Numpy
- Pandas
- Geopy
- Requests
- JSON_Normalize
- Matplotlib
- Folium

Step 2: Scraping the first table from Wikipedia and putting it in a data frame.

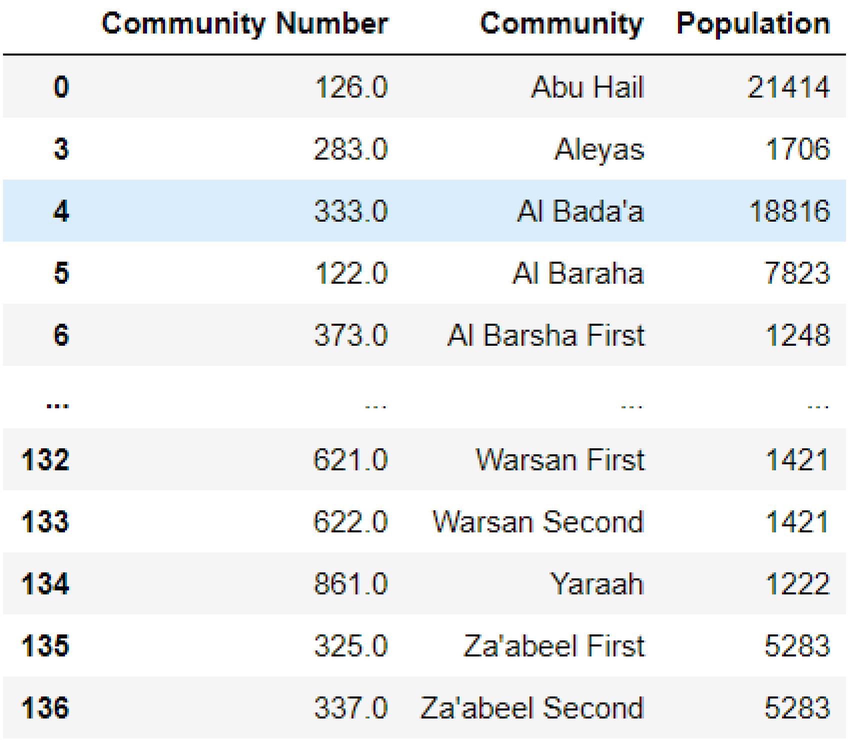
I have scraped the data/table of Dubai communities from the Wikipedia and converted it into a data frame so it can be worked on.

It needed few clean ups, which will be illustrated later on in the report; as you can below is an example of the raw data imported.



	Community Number	Community (English)	Community (Arabic)	Area(km2)	Population(2000)	Population density(/km2)	Unnamed: 6
0	126.0	Abu Hail	أبو هيل	1.27 km²	21414	16,861.4/km²	NaN
1	711.0	Al Awir First	العوير الأولى	NaN	NaN	NaN	NaN
2	721.0	Al Awir Second	العوير الثانية	NaN	NaN	NaN	NaN
3	283.0	Aleyas	العياص	162.4 km2	1706	162.4/km2	NaN
4	333.0	Al Bada'a	البدع	0.82 km²	18816	22946/km²	NaN
...	...	...	...	...	...	...	...
141	914.0	Umm Nahad Fourth	NaN	NaN	NaN	NaN	NaN
142	971.0	Saih Al-Dahal	NaN	NaN	NaN	NaN	NaN
143	951.0	Saih Al Salam	NaN	NaN	NaN	NaN	NaN
144	931.0	Al Lisaili	NaN	NaN	NaN	NaN	NaN
145	731.0	Lehbab First	NaN	NaN	NaN	NaN	NaN

I then did few clean-ups and renamed few columns that will be used later as Features. Here’s what the first data frame looks like:

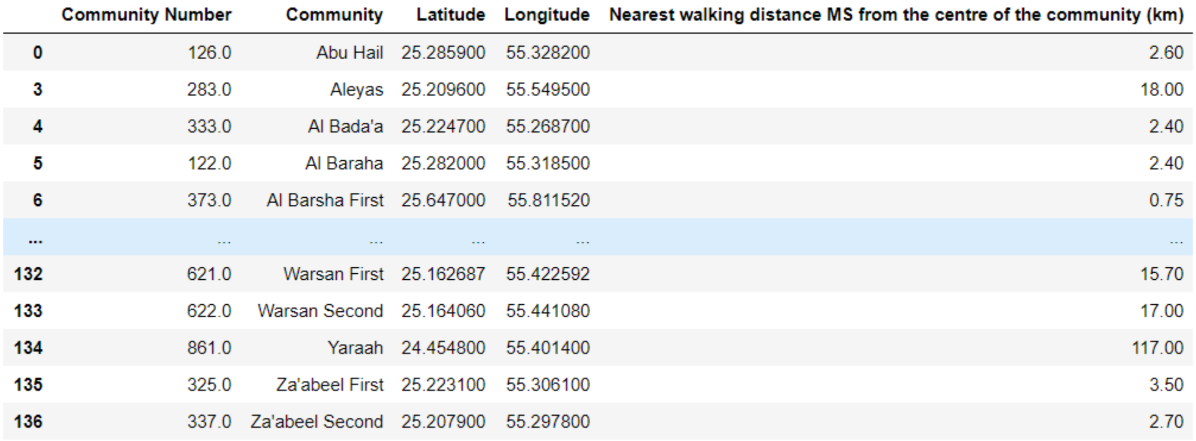


	Community Number	Community	Population
0	126.0	Abu Hail	21414
3	283.0	Aleyas	1706
4	333.0	Al Bada'a	18816
5	122.0	Al Baraha	7823
6	373.0	Al Barsha First	1248
...	...	...	...
132	621.0	Warsan First	1421
133	622.0	Warsan Second	1421
134	861.0	Yaraah	1222
135	325.0	Za'abeel First	5283
136	337.0	Za'abeel Second	5283

Step 3: Creating the second data frame with the coordinates and metro station distance features.

This step is done by googling each community coordinates separately and then measuring, using Google Maps, the distance from the centre of the community to the nearest metro station, since I couldn’t retrieve such data in bulk. The data is then saved into a .csv file that will be imported into a data frame.

Below is the second data frame, after I imported the data from the .csv file and cleaned it up. I basically removed all the rows where the Population is “NaN” (Not a Number) and renamed the Community column. As you can see the index is not reset but that doesn’t matter at this stage.

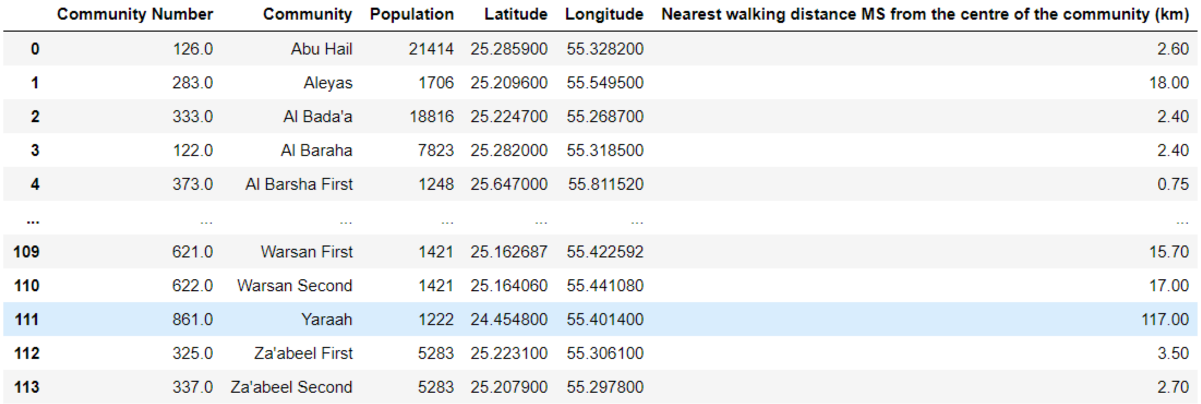


	Community Number	Community	Latitude	Longitude	Nearest walking distance MS from the centre of the community (km)
0	126.0	Abu Hail	25.285900	55.328200	2.60
3	283.0	Aleyas	25.209600	55.549500	18.00
4	333.0	Al Bada'a	25.224700	55.268700	2.40
5	122.0	Al Baraha	25.282000	55.318500	2.40
6	373.0	Al Barsha First	25.647000	55.811520	0.75
...	...	...	...	...	...
132	621.0	Warsan First	25.162687	55.422592	15.70
133	622.0	Warsan Second	25.164060	55.441080	17.00
134	861.0	Yaraah	24.454800	55.401400	117.00
135	325.0	Za'abeel First	25.223100	55.306100	3.50
136	337.0	Za'abeel Second	25.207900	55.297800	2.70

Step 4: Merging the first and second data frames so we can prepare the resulted data frame to incorporate the venues later on.

Now that I have the 2 needed data frames as shown previously, I merged them together to form the data frame which I will be using for the Foursquare API as explained in the next step.

The data in this step contain the Population column as shown below.

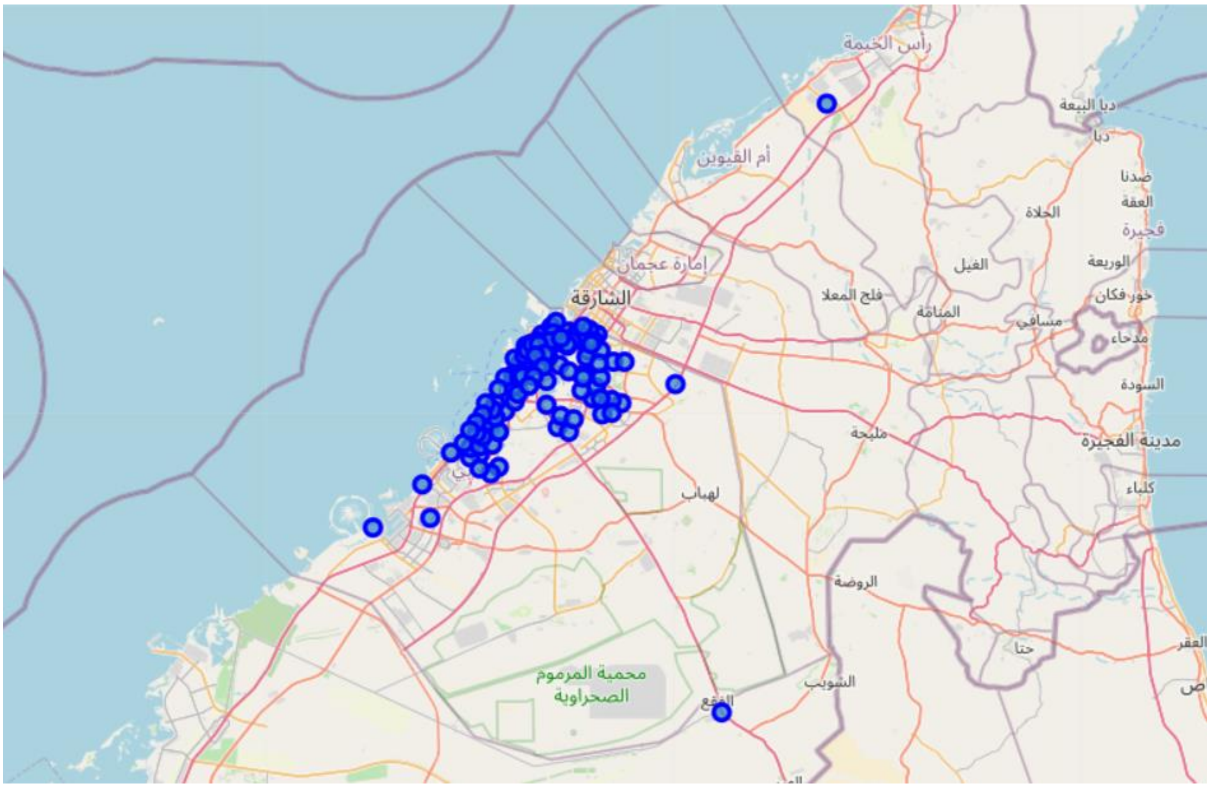


	Community Number	Community	Population	Latitude	Longitude	Nearest walking distance MS from the centre of the community (km)
0	126.0	Abu Hail	21414	25.285900	55.328200	2.60
1	283.0	Aleyas	1706	25.209600	55.549500	18.00
2	333.0	Al Bada'a	18816	25.224700	55.268700	2.40
3	122.0	Al Baraha	7823	25.282000	55.318500	2.40
4	373.0	Al Barsha First	1248	25.647000	55.811520	0.75
...	...	...	...	...	...	...
109	621.0	Warsan First	1421	25.162687	55.422592	15.70
110	622.0	Warsan Second	1421	25.164060	55.441080	17.00
111	861.0	Yaraah	1222	24.454800	55.401400	117.00
112	325.0	Za'abeel First	5283	25.223100	55.306100	3.50
113	337.0	Za'abeel Second	5283	25.207900	55.297800	2.70

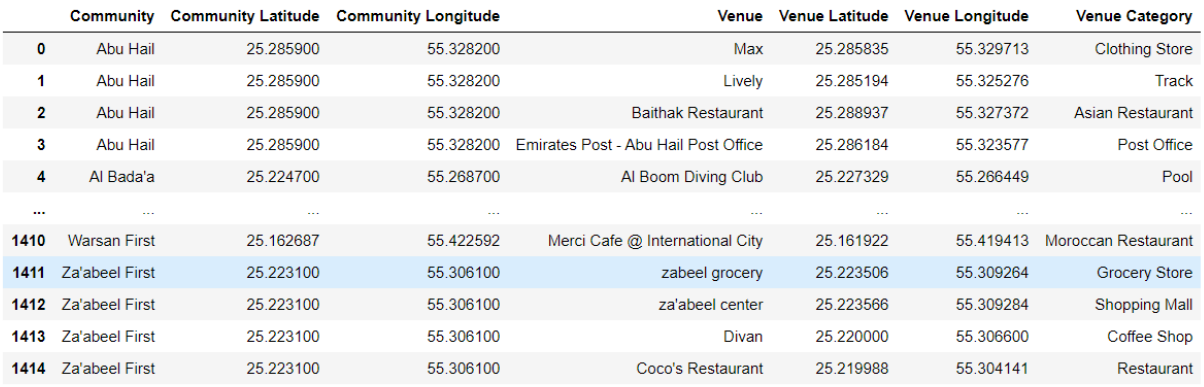
Step 5: Getting the data from Foursquare showing the venues of each community, and creating the third data frame

In this step I have written a script using geolocator to automatically search and return the coordinates of Dubai.

I then visualized the communities on Dubai map using Folium. Here is the result:



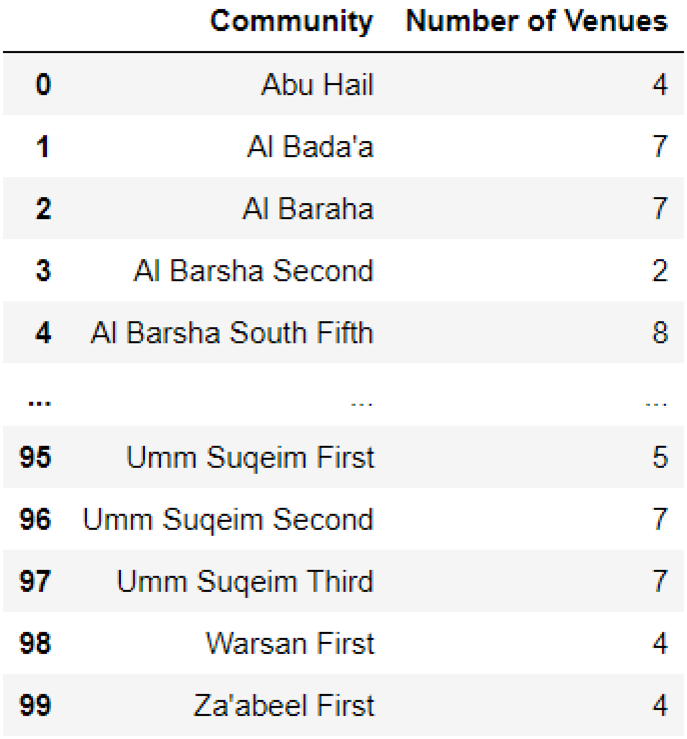
After, I have used a function that explores all the communities in Dubai and their venues. Below is the data frame as a result of that; as you can see we have 1415 venues all over Dubai as of 2020-06-05 per the Foursquare API.



	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abu Hail	25.285900	55.328200	Max	25.285835	55.329713	Clothing Store
1	Abu Hail	25.285900	55.328200	Lively	25.285194	55.325276	Track
2	Abu Hail	25.285900	55.328200	Baithak Restaurant	25.288937	55.327372	Asian Restaurant
3	Abu Hail	25.285900	55.328200	Emirates Post - Abu Hail Post Office	25.286184	55.323577	Post Office
4	Al Bada'a	25.224700	55.268700	Al Boom Diving Club	25.227329	55.266449	Pool
...	...	...	...	...	...	...	...
1410	Warsan First	25.162687	55.422592	Merci Cafe @ International City	25.161922	55.419413	Moroccan Restaurant
1411	Za'abeel First	25.223100	55.306100	zabeel grocery	25.223506	55.309264	Grocery Store
1412	Za'abeel First	25.223100	55.306100	za'abeel center	25.223566	55.309284	Shopping Mall
1413	Za'abeel First	25.223100	55.306100	Divan	25.220000	55.306600	Coffee Shop
1414	Za'abeel First	25.223100	55.306100	Coco's Restaurant	25.219988	55.304141	Restaurant

Step 6: Creating “Number of Venues” column and naming the new data frame "communities_count_venues_updated".

As mentioned earlier, I will be using few Features for my analysis and one of them is Number of Venues, therefore, I have grouped the above data frame by Community and used count() to get the number of venues per community.



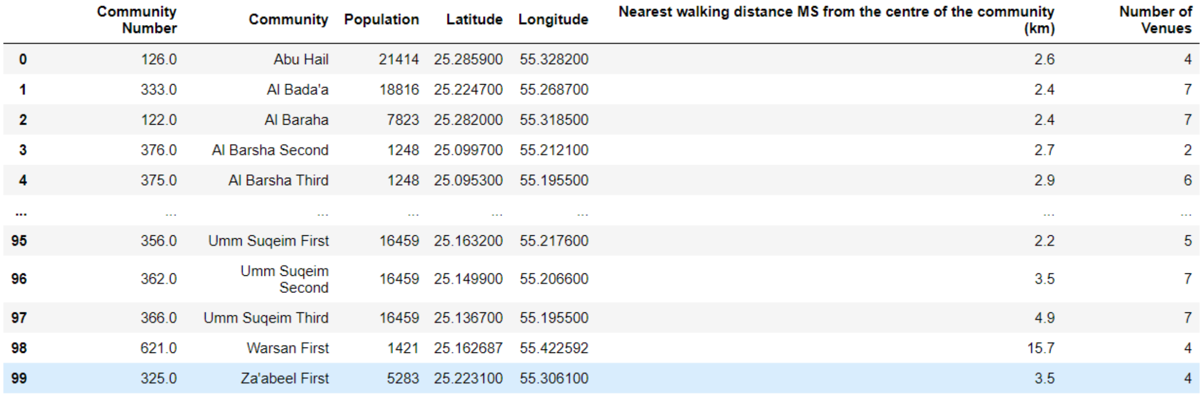
	Community	Number of Venues
0	Abu Hail	4
1	Al Bada'a	7
2	Al Baraha	7
3	Al Barsha Second	2
4	Al Barsha South Fifth	8
...	...	...
95	Umm Suqeim First	5
96	Umm Suqeim Second	7
97	Umm Suqeim Third	7
98	Warsan First	4
99	Za'abeel First	4

Step 7: Merging all 3 data frames and dropping the un-needed columns.

Last step in the Data Frame section is the merge of all the previous data frames by keeping only the needed columns that will be used later for clustering.

I have also converted the Community Number to float as the type was Object since this could create issues later.

This what the final data frame looks like before applying any normalization. You can see that the number of venues in each community is pretty low.



	Community Number	Community	Population	Latitude	Longitude	Nearest walking distance MS from the centre of the community (km)	Number of Venues
0	126.0	Abu Hail	21414	25.285900	55.328200	2.6	4
1	333.0	Al Bada'a	18816	25.224700	55.268700	2.4	7
2	122.0	Al Baraha	7823	25.282000	55.318500	2.4	7
3	376.0	Al Barsha Second	1248	25.099700	55.212100	2.7	2
4	375.0	Al Barsha Third	1248	25.095300	55.195500	2.9	6
...	...	...	...	...	...	...	...
95	356.0	Umm Suqeim First	16459	25.163200	55.217600	2.2	5
96	362.0	Umm Suqeim Second	16459	25.149900	55.206600	3.5	7
97	366.0	Umm Suqeim Third	16459	25.136700	55.195500	4.9	7
98	621.0	Warsan First	1421	25.162687	55.422592	15.7	4
99	325.0	Za'abeel First	5283	25.223100	55.306100	3.5	4

Methodology

The main focus will be on the **Population**, **Number of Venues** and the **Nearest walking distance MS from the centre of the community** features for the decision making analysis. The reason to choose these variables is because of their fundamental impact on the conception of the idea to extend the metro infrastructure and build new underground stations. Without having enough population, venues, reachability within an acceptable timeframe or all or some of them combined, the idea of extension will not be feasible.

The clustering methodology that will be used is K-Means, an un-supervised machine learning algorithm.

For that I will be normalizing my data and creating a piece of code that will generate a graph showing the best K to use; it is basically plotting the K numbers on the x axis and the mean distance of data points to cluster centroid on the y axis and determining the “Elbow”, the point where the line steepness starts to significantly decrease; that will be the best K to choose.

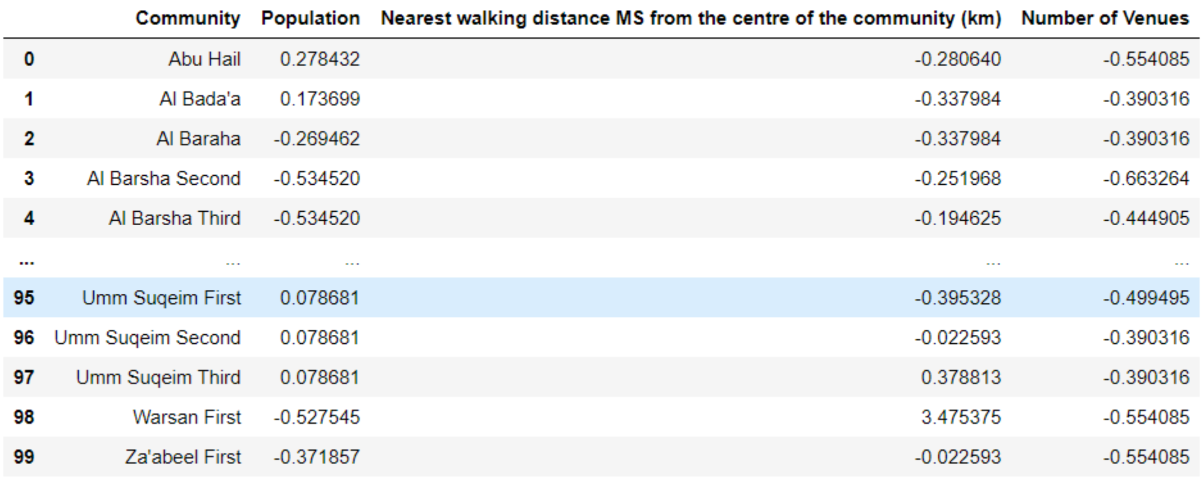
The last step in the methodology before starting the analysis phase is the use of the best K to cluster the communities based on the determined features as mentioned before, and then plotting those clusters for visualization and analysis.

Step 1: Data normalization

The data that was scrapped from Wikipedia as mentioned earlier was not optimal. I have realized that one of the Population values was showing an Object format which will certainly create an error when normalizing.

So in this step I have done some data wrangling by removing the comma from the value and changing the format to float.

I have used preprocessing.StandardScaler imported from sklearn to do the normalization by fitting and transforming the data. Here is what the final normalized ready to cluster data frame looks like:



	Community	Population	Nearest walking distance MS from the centre of the community (km)	Number of Venues
0	Abu Hail	0.278432	-0.280640	-0.554085
1	Al Bada'a	0.173699	-0.337984	-0.390316
2	Al Baraha	-0.269462	-0.337984	-0.390316
3	Al Barsha Second	-0.534520	-0.251968	-0.663264
4	Al Barsha Third	-0.534520	-0.194625	-0.444905
...	...	...	...	...
95	Umm Suqeim First	0.078681	-0.395328	-0.499495
96	Umm Suqeim Second	0.078681	-0.022593	-0.390316
97	Umm Suqeim Third	0.078681	0.378813	-0.390316
98	Warsan First	-0.527545	3.475375	-0.554085
99	Za'abeel First	-0.371857	-0.022593	-0.554085

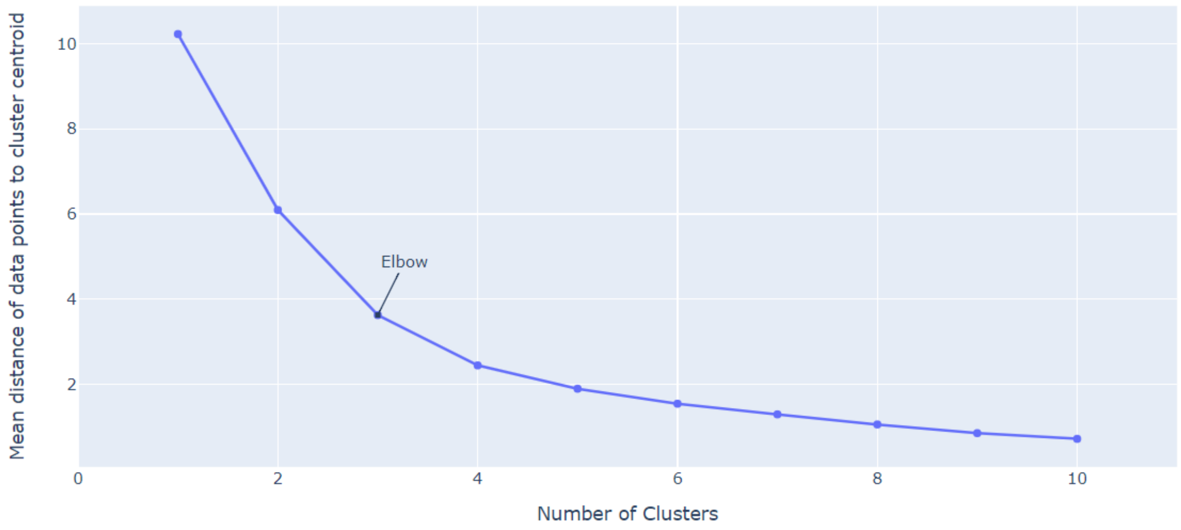
Step 2: Checking for the best K to be used in clustering

One of the important question when using k-means algorithm is what value of k to use. For that I have written a function that loops into a range of 20 k's, does the clustering for each value and plot these values into a graph that shows the "Elbow", a point with the best k.

I have installed and used plotly to be able to plot the following graph along with using MinMaxScaler imported from sklearn.preprocessing to fit and transform the data frame in order to acquire the mean distance of data or inertia.

Below is the graph illustrating the result:

Graphical Interpretation for Choosing the Best K

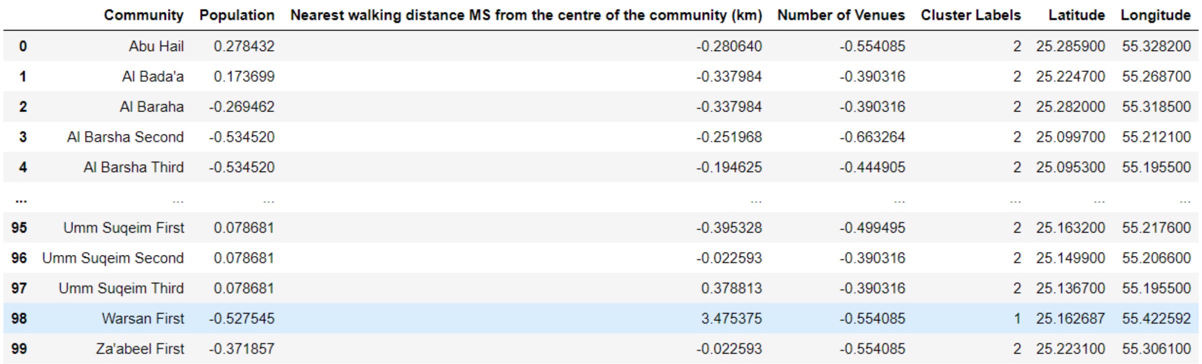


This graph shows that the 'Elbow' is at value 3, so we are going ahead with K=3

Step 3: Clustering

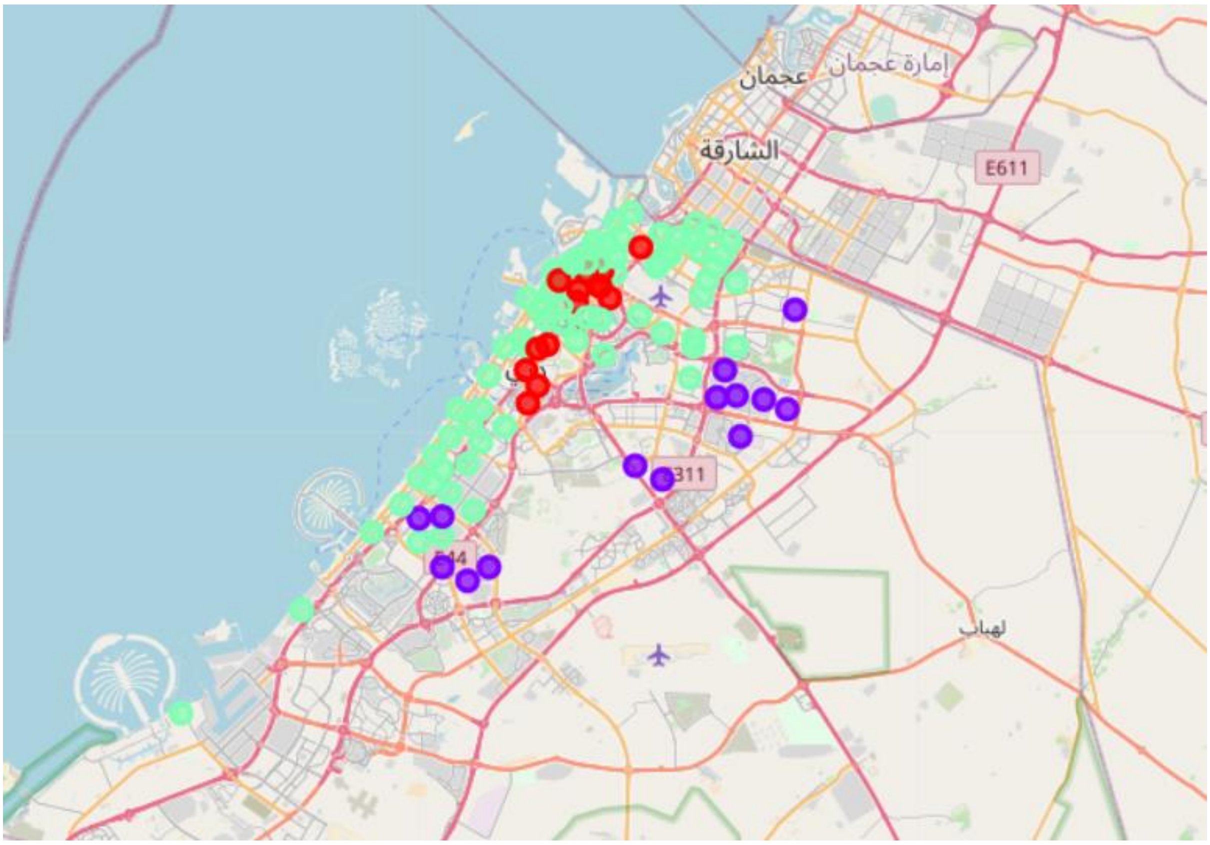
The reason I will be using **k-means++** as the algorithm is to seed the initial centers for k-means and avoiding the sometimes poor clustering found by the standard k-means algorithm. Market experiments show the augmentation of k-means, which is k-means++, improves both the speed and the accuracy of k-means, often quite dramatically.

As mentioned, I have clustered my data frame using k-means with **init='k-means++'** with **k=3** and **random_state=4** in order to always have the same clusters. Afterwards, I got the k-means labels and inserted them as a column called Cluster Labels into the normalized data frame. Below is the result of the data frame that will be used to visualize the clusters on the map:



	Community	Population	Nearest walking distance MS from the centre of the community (km)	Number of Venues	Cluster Labels	Latitude	Longitude
0	Abu Hail	0.278432	-0.280640	-0.554085	2	25.285900	55.328200
1	Al Bada'a	0.173699	-0.337984	-0.390316	2	25.224700	55.268700
2	Al Baraha	-0.269462	-0.337984	-0.390316	2	25.282000	55.318500
3	Al Barsha Second	-0.534520	-0.251968	-0.663264	2	25.099700	55.212100
4	Al Barsha Third	-0.534520	-0.194625	-0.444905	2	25.095300	55.195500
...	...	...	...	...	...	...	...
95	Umm Suqeim First	0.078681	-0.395328	-0.499495	2	25.163200	55.217600
96	Umm Suqeim Second	0.078681	-0.022593	-0.390316	2	25.149900	55.206600
97	Umm Suqeim Third	0.078681	0.378813	-0.390316	2	25.136700	55.195500
98	Warsan First	-0.527545	3.475375	-0.554085	1	25.162687	55.422592
99	Za'abeel First	-0.371857	-0.022593	-0.554085	2	25.223100	55.306100

Visualizing the clusters on the map:

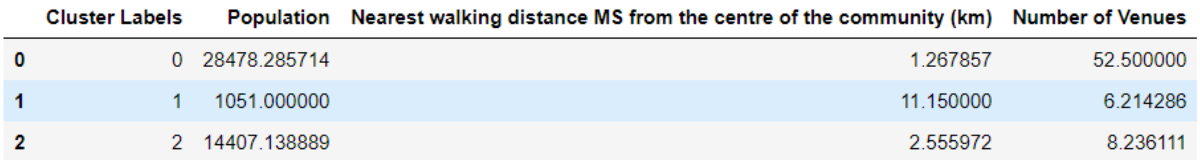


Results of Clustering

Further analysis on the clusters

In order to make sense of the clustering results, I needed to get the mean of the Population, the Nearest Walking Distance MS from the Center of the Community and the Number Venues for each cluster label. With that I can better analyse the difference between these clusters and reach a decision for the extension.

Here is the data frame showing the means:



	Cluster Labels	Population	Nearest walking distance MS from the centre of the community (km)	Number of Venues
0	0	28478.285714	1.267857	52.500000
1	1	1051.000000	11.150000	6.214286
2	2	14407.138889	2.555972	8.236111

I have categorized by colour each cluster to make things easier for reference:

- Label 0: **Red Cluster**
- Label 1: **Blue Cluster**
- Label 2: **Green Cluster**

Observations and Analysis

Compared to the defined decision criterion hypothesis, we see that the number of venues is way smaller than expected in all of the clusters; that is between 1 and 4 venues per 1000 people where the defined threshold ratio is 15:1000, therefore, it is fair to say that from the data exploration done the number of venues feature will not play a big role in influencing the decision procedure.

The focus will be on the population and nearest walking distance.

Both the Green Cluster and the Blue Cluster have a population higher than 1000 where the Green Cluster is around 14x higher. However, in regards to the nearest walking distance to the metro stations, the Blue Cluster mean distance is around 8.6km longer.

The Blue Cluster on the other hand has a population higher than the threshold (which is 1000) but with a quite long walking distance to the nearest metro station.

Another very important observation when looking at the cluster distribution on the map, the areas with the Blue Cluster are having almost no metro infrastructure.

Conclusions

Based on the observations and analysis done, we can conclude the following for the metro infrastructure extension and underground stations build-up:

- Blue Cluster should have priority 1 for extension.

- Green Cluster will have priority 2.

- Red Cluster no extension of infrastructure needed as the decision criterion is not met – As a citizen of Dubai, a real life observation shows indeed that most of the metro stations are located within these areas, which gives a good indicator of the accuracy of this model.

The reason of having the Green Cluster as priority 2 from my point of view, is because Dubai's weather is very hot for quite a long time of the year, and walking outside for more than 20-30 minutes might be inconvenient for a lot of people, therefore I have taken the walking distance to the metro station as a factor.

Notes

Couple of notes to consider in this analysis:

- This model could certainly use improvements, however, it shows quite good results which can be leverage in the decision making process.
- Feasibility studies of such a project rely definitely on few other factors such as cost/budget, resources, area capacity, environmental impact, etc.