Clustering Indian Cities into Tiers

based on

Availability of Facilities

❖ INTRODUCTION

Business Problem:

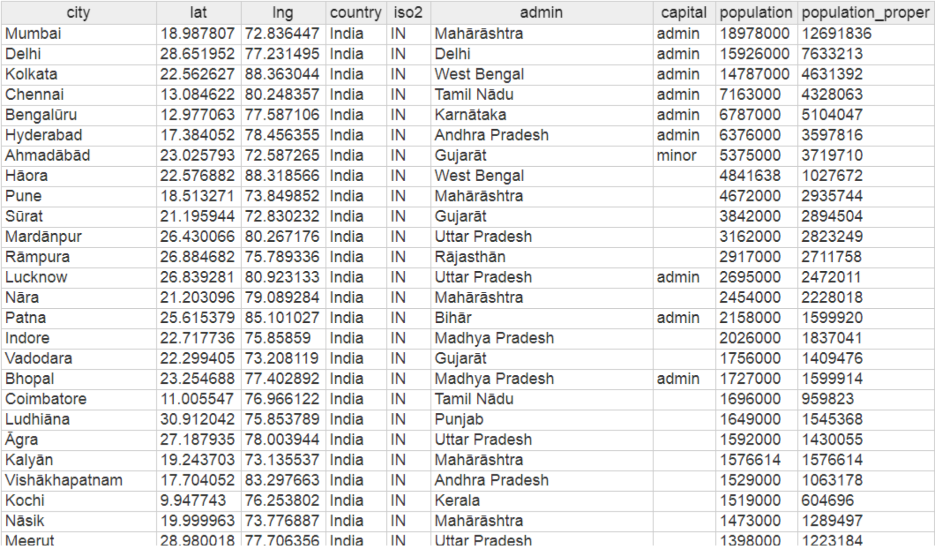
To group Indian cities into Tiers based on availability of facilities like healthcare, education, employment, transport, etc.

Stakeholders:

Private and public sector companies who may want to explore areas for setting up new branches in cities which are yet to experience major growth. This enables them to set up business with minimum cost and get first mover's advantage since lower tier cities would have lesser competitors.

❖ DATA

The data set used for this operation is taken from https://simplemaps.com/data/in-cities



city	lat	lng	country	iso2	admin	capital	population	population_proper
Mumbai	18.987807	72.836447	India	IN	Mahārāshtra	admin	18978000	12691836
Delhi	28.651952	77.231495	India	IN	Delhi	admin	15926000	7633213
Kolkata	22.562627	88.363044	India	IN	West Bengal	admin	14787000	4631392
Chennai	13.084622	80.248357	India	IN	Tamil Nādu	admin	7163000	4328063
Bengalūru	12.977063	77.587106	India	IN	Karnātaka	admin	6787000	5104047
Hyderabad	17.384052	78.456355	India	IN	Andhra Pradesh	admin	6376000	3597816
Ahmadābād	23.025793	72.587265	India	IN	Gujarāt	minor	5375000	3719710
Hāora	22.576882	88.318566	India	IN	West Bengal		4841638	1027672
Pune	18.513271	73.849852	India	IN	Mahārāshtra		4672000	2935744
Sūrat	21.195944	72.830232	India	IN	Gujarāt		3842000	2894504
Mardānpur	26.430066	80.267176	India	IN	Uttar Pradesh		3162000	2823249
Rāmpura	26.884682	75.789336	India	IN	Rājasthān		2917000	2711758
Lucknow	26.839281	80.923133	India	IN	Uttar Pradesh	admin	2695000	2472011
Nāra	21.203096	79.089284	India	IN	Mahārāshtra		2454000	2228018
Patna	25.615379	85.101027	India	IN	Bihār	admin	2158000	1599920
Indore	22.717736	75.85859	India	IN	Madhya Pradesh		2026000	1837041
Vadodara	22.299405	73.208119	India	IN	Gujarāt		1756000	1409476
Bhopal	23.254688	77.402892	India	IN	Madhya Pradesh	admin	1727000	1599914
Coimbatore	11.005547	76.966122	India	IN	Tamil Nādu		1696000	959823
Ludhiāna	30.912042	75.853789	India	IN	Punjab		1649000	1545368
Āgra	27.187935	78.003944	India	IN	Uttar Pradesh		1592000	1430055
Kalyān	19.243703	73.135537	India	IN	Mahārāshtra		1576614	1576614
Vishākhapatnam	17.704052	83.297663	India	IN	Andhra Pradesh		1529000	1063178
Kochi	9.947743	76.253802	India	IN	Kerala		1519000	604696
Nāsik	19.999963	73.776887	India	IN	Mahārāshtra		1473000	1289497
Meerut	28.980018	77.706356	India	IN	Uttar Pradesh		1398000	1223184

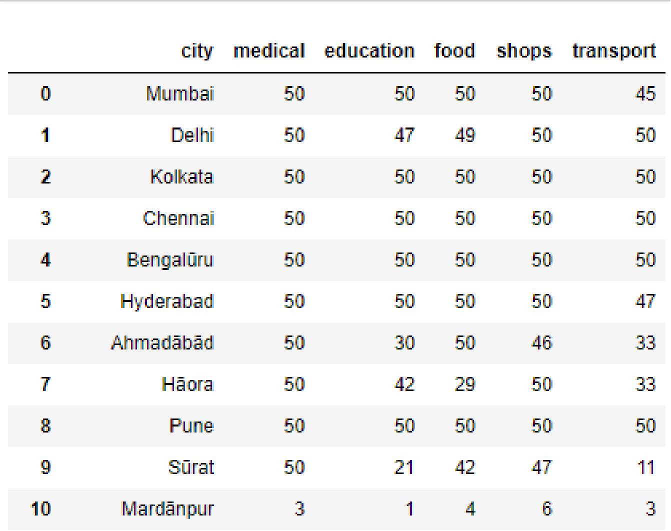
The dataset has 212 cities. Attributes which we will use for our application are:

1. City Name (city)
2. Latitude (lat)
3. Longitude (lng)
4. State (admin)
5. Population (population_proper)

Using Foursquare API we gather all the educational institutions, medical institutions, hotels & food joints, transport, shops & services for every city within a radius of 5km from its centre denoted by latitude & longitude values.

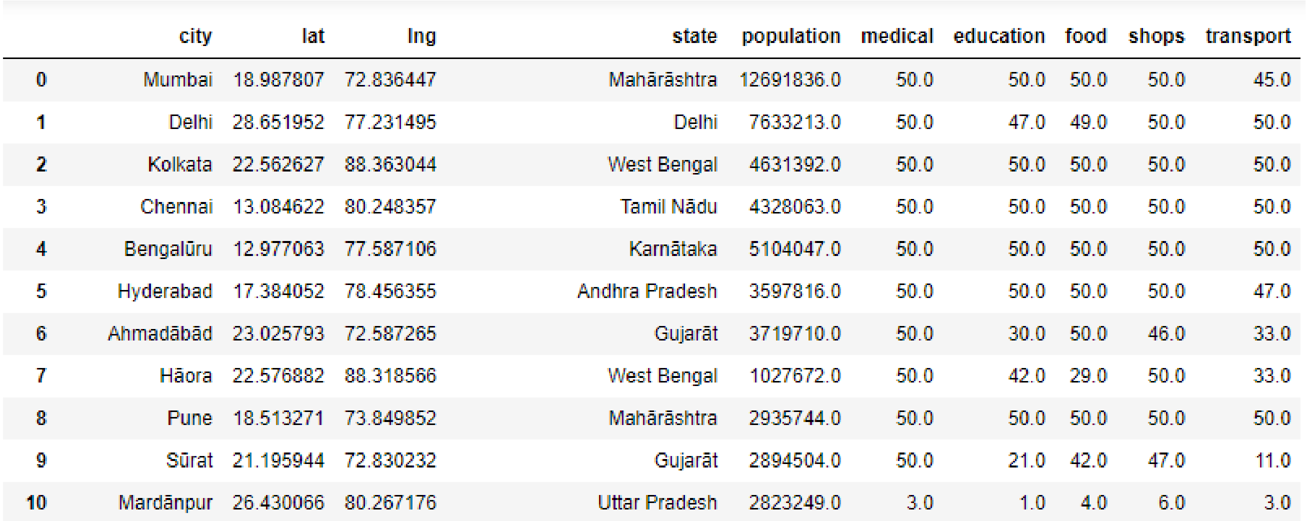
For each city a count of these venues is generated. Higher the count of these venues better the tier.

The dataset looks like this:



	city	medical	education	food	shops	transport
0	Mumbai	50	50	50	50	45
1	Delhi	50	47	49	50	50
2	Kolkata	50	50	50	50	50
3	Chennai	50	50	50	50	50
4	Bengalūru	50	50	50	50	50
5	Hyderabad	50	50	50	50	47
6	Ahmadābād	50	30	50	46	33
7	Hāora	50	42	29	50	33
8	Pune	50	50	50	50	50
9	Sūrat	50	21	42	47	11
10	Mardānpur	3	1	4	6	3

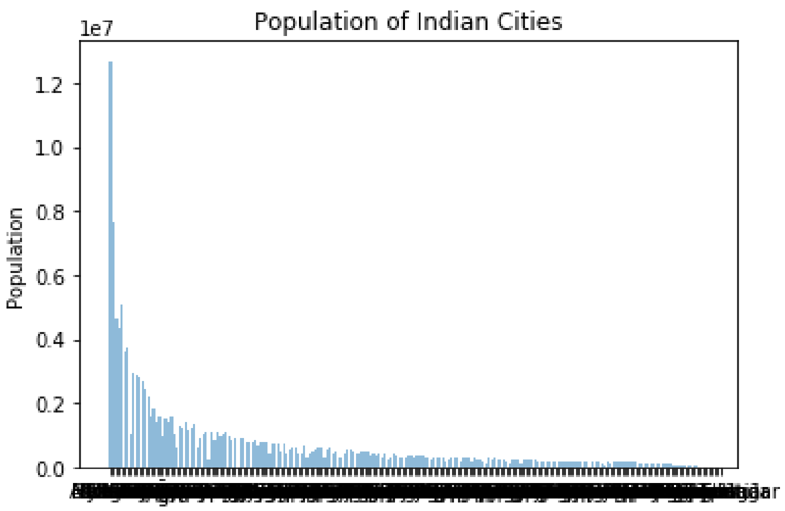
Finally we merge the two data frames using ‘city’ as index. We get the following dataset:

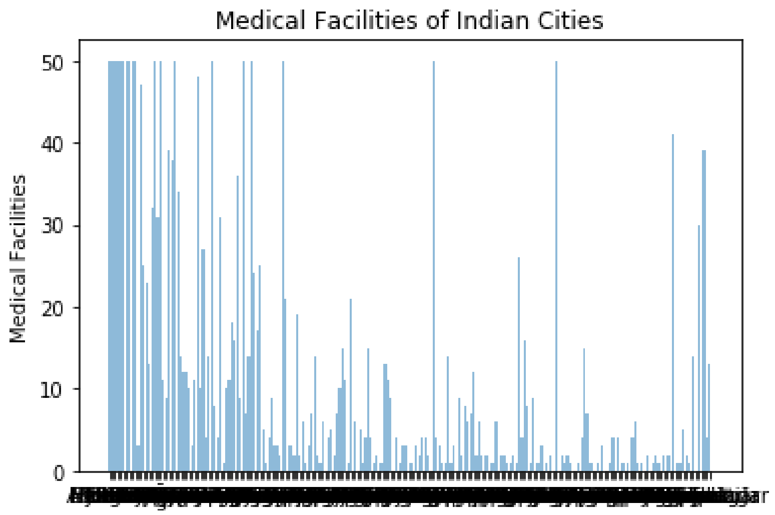


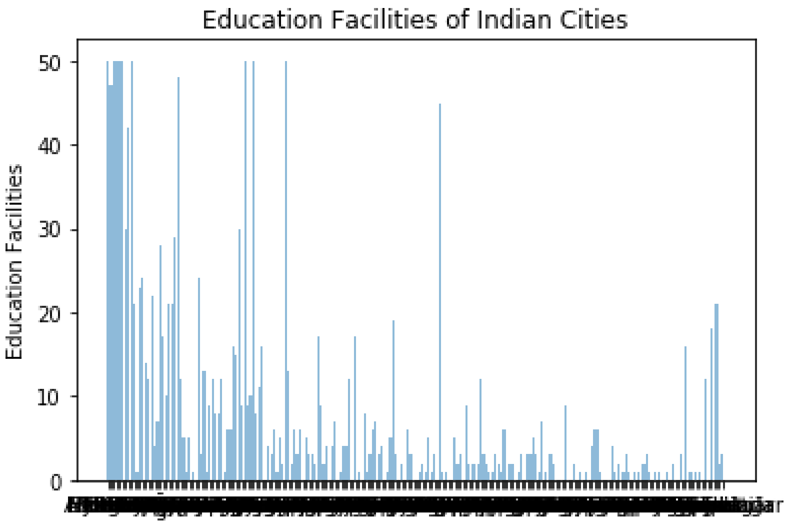
	city	lat	lng	state	population	medical	education	food	shops	transport
0	Mumbai	18.987807	72.836447	Mahārāshtra	12691836.0	50.0	50.0	50.0	50.0	45.0
1	Delhi	28.651952	77.231495	Delhi	7633213.0	50.0	47.0	49.0	50.0	50.0
2	Kolkata	22.562627	88.363044	West Bengal	4631392.0	50.0	50.0	50.0	50.0	50.0
3	Chennai	13.084622	80.248357	Tamil Nādu	4328063.0	50.0	50.0	50.0	50.0	50.0
4	Bengalūru	12.977063	77.587106	Karnātaka	5104047.0	50.0	50.0	50.0	50.0	50.0
5	Hyderabad	17.384052	78.456355	Andhra Pradesh	3597816.0	50.0	50.0	50.0	50.0	47.0
6	Ahmadābād	23.025793	72.587265	Gujarāt	3719710.0	50.0	30.0	50.0	46.0	33.0
7	Hāora	22.576882	88.318566	West Bengal	1027672.0	50.0	42.0	29.0	50.0	33.0
8	Pune	18.513271	73.849852	Mahārāshtra	2935744.0	50.0	50.0	50.0	50.0	50.0
9	Sūrat	21.195944	72.830232	Gujarāt	2894504.0	50.0	21.0	42.0	47.0	11.0
10	Mardānpur	26.430066	80.267176	Uttar Pradesh	2823249.0	3.0	1.0	4.0	6.0	3.0

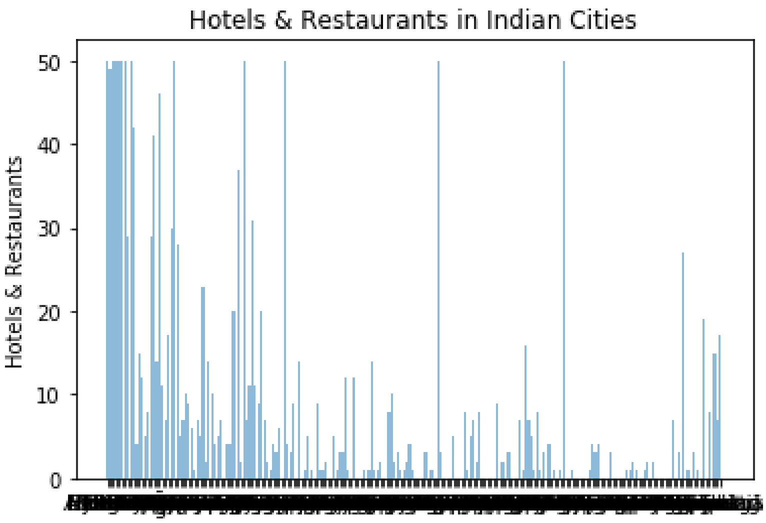
❖ METHODOLOGY:

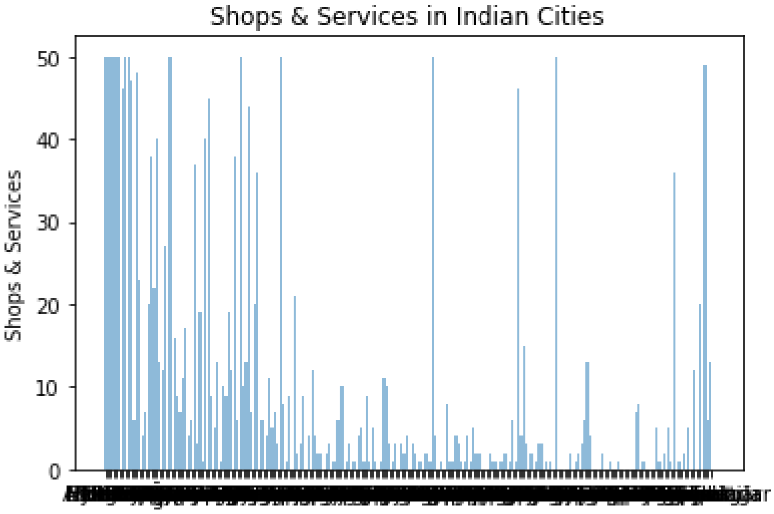
Data Visualization using Bar Graphs:

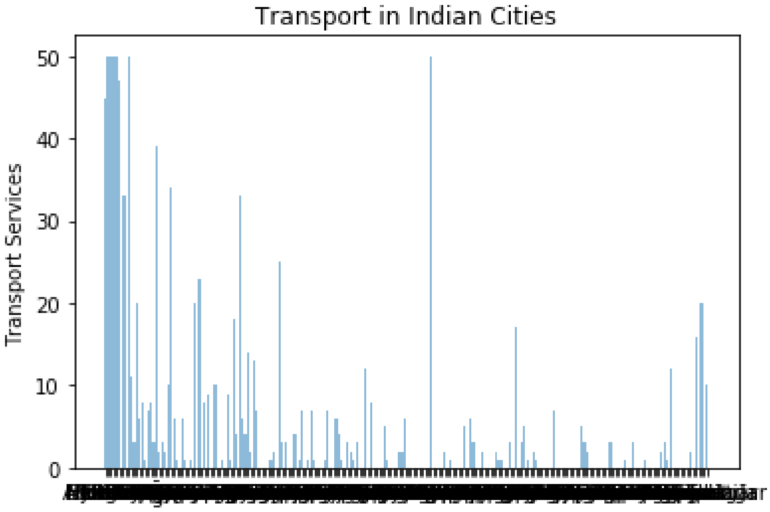






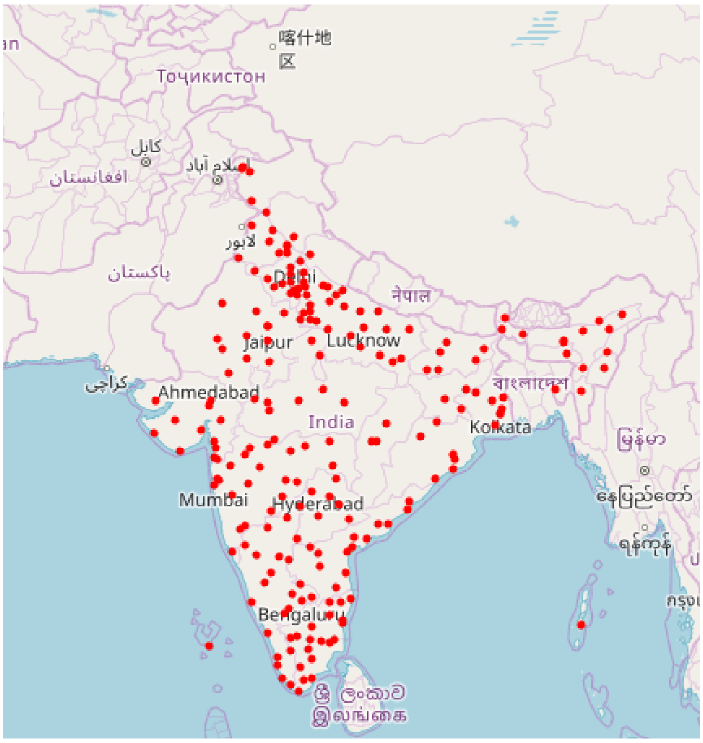






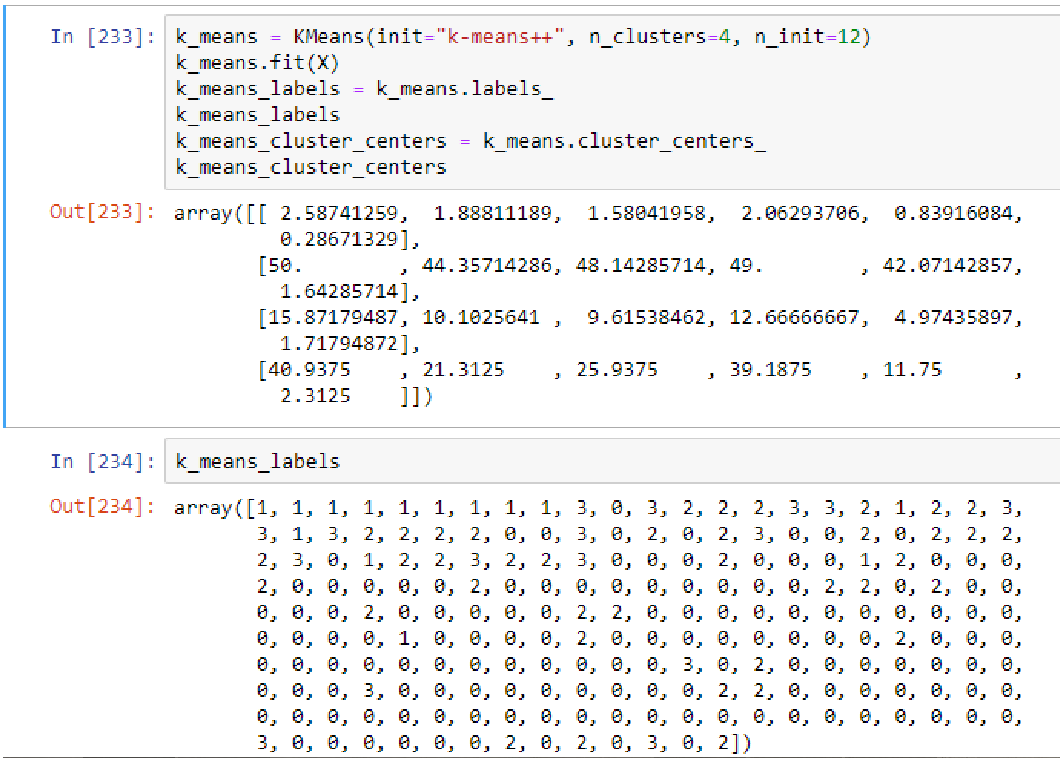
Based on these graphs I have decided to choose number of tiers (clusters) to be 4 for K-Means algorithm.

Plotting the cities on a Folium Map



❖ RESULTS:

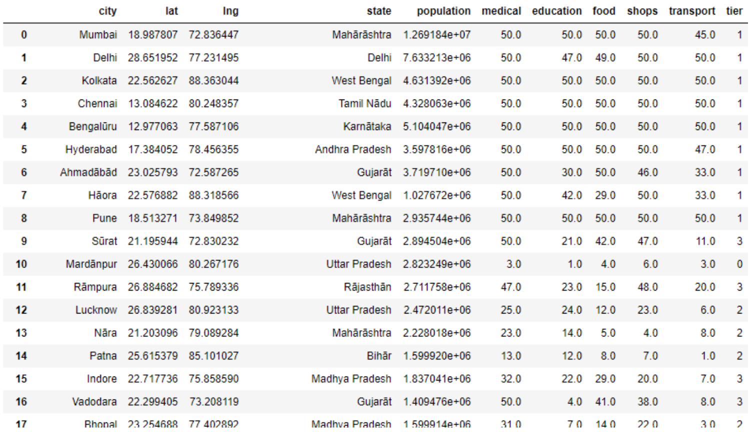
K-means centres and labels for the 4 clusters



```
In [233]: k_means = KMeans(init="k-means++", n_clusters=4, n_init=12)
          k_means.fit(X)
          k_means_labels = k_means.labels_
          k_means_labels
          k_means_cluster_centers = k_means.cluster_centers_
          k_means_cluster_centers

Out[233]: array([[ 2.58741259,  1.88811189,  1.58041958,  2.06293706,  0.83916084,
                   0.28671329],
                 [50.        , 44.35714286, 48.14285714, 49.        , 42.07142857,
                   1.64285714],
                 [15.87179487, 10.1025641 ,  9.61538462, 12.66666667,  4.97435897,
                   1.71794872],
                 [40.9375    , 21.3125    , 25.9375    , 39.1875    , 11.75      ,
                   2.3125    ]])
```

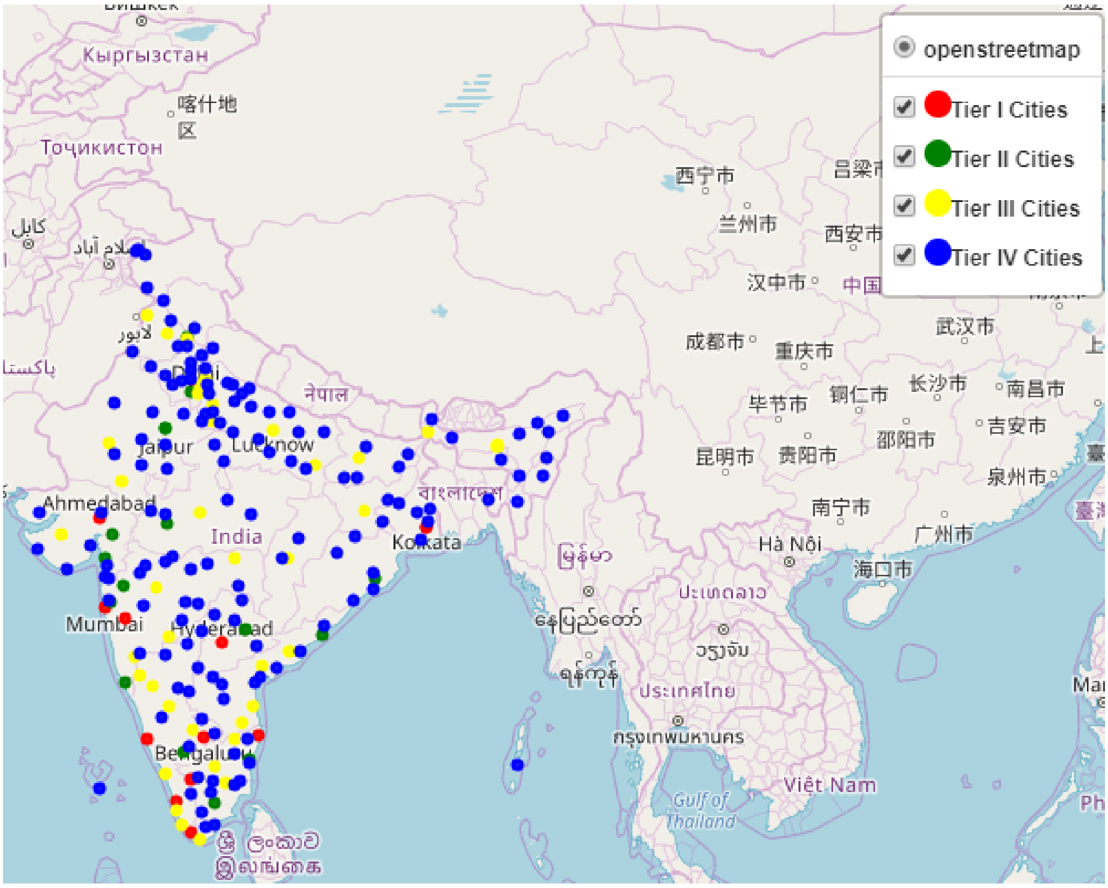
```
In [234]: k_means_labels

Out[234]: array([1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 0, 3, 2, 2, 2, 3, 3, 2, 1, 2, 2, 3,
                 3, 1, 3, 2, 2, 2, 2, 0, 0, 3, 0, 2, 0, 2, 3, 0, 0, 2, 0, 2, 2, 2,
                 2, 3, 0, 1, 2, 2, 3, 2, 2, 3, 0, 0, 0, 2, 0, 0, 0, 1, 2, 0, 0, 0,
                 2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 2, 0, 0,
                 0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 2, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 3, 0, 0, 0, 0, 0, 0, 2, 0, 2, 0, 3, 0, 2])
```

Snapshot of the cities and their predicted labels by K-Means



	city	lat	lng	state	population	medical	education	food	shops	transport	tier
0	Mumbai	18.987807	72.836447	Mahārāshtra	1.269184e+07	50.0	50.0	50.0	50.0	45.0	1
1	Delhi	28.651952	77.231495	Delhi	7.633213e+06	50.0	47.0	49.0	50.0	50.0	1
2	Kolkata	22.562627	88.363044	West Bengal	4.631392e+06	50.0	50.0	50.0	50.0	50.0	1
3	Chennai	13.084622	80.248357	Tamil Nādu	4.328063e+06	50.0	50.0	50.0	50.0	50.0	1
4	Bengalūru	12.977063	77.587106	Karnātaka	5.104047e+06	50.0	50.0	50.0	50.0	50.0	1
5	Hyderabad	17.384052	78.456355	Andhra Pradesh	3.597816e+06	50.0	50.0	50.0	50.0	47.0	1
6	Ahmadābād	23.025793	72.587265	Gujarāt	3.719710e+06	50.0	30.0	50.0	46.0	33.0	1
7	Hāora	22.576882	88.318566	West Bengal	1.027672e+06	50.0	42.0	29.0	50.0	33.0	1
8	Pune	18.513271	73.849852	Mahārāshtra	2.935744e+06	50.0	50.0	50.0	50.0	50.0	1
9	Sūrat	21.195944	72.830232	Gujarāt	2.894504e+06	50.0	21.0	42.0	47.0	11.0	3
10	Mardānpur	26.430066	80.267176	Uttar Pradesh	2.823249e+06	3.0	1.0	4.0	6.0	3.0	0
11	Rāmpura	26.884682	75.789336	Rājasthān	2.711758e+06	47.0	23.0	15.0	48.0	20.0	3
12	Lucknow	26.839281	80.923133	Uttar Pradesh	2.472011e+06	25.0	24.0	12.0	23.0	6.0	2
13	Nāra	21.203096	79.089284	Mahārāshtra	2.228018e+06	23.0	14.0	5.0	4.0	8.0	2
14	Patna	25.615379	85.101027	Bihār	1.599920e+06	13.0	12.0	8.0	7.0	1.0	2
15	Indore	22.717736	75.858590	Madhya Pradesh	1.837041e+06	32.0	22.0	29.0	20.0	7.0	3
16	Vadodara	22.299405	73.208119	Gujarāt	1.409476e+06	50.0	4.0	41.0	38.0	8.0	3
17	Bhopal	23.254688	77.402892	Madhya Pradesh	1.599914e+06	31.0	7.0	14.0	22.0	3.0	2

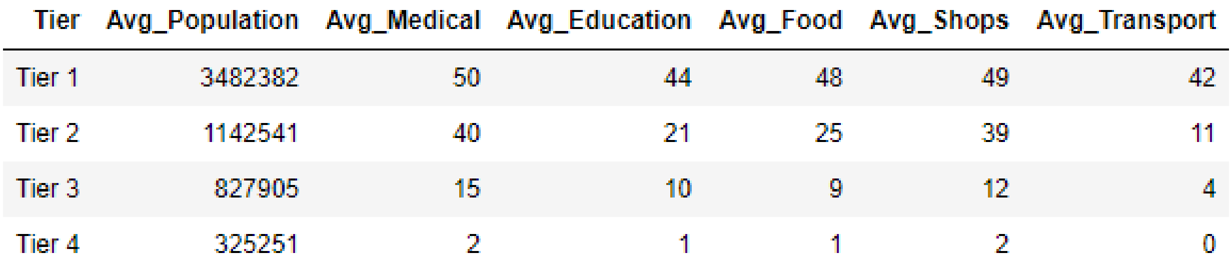
Colour Coded Cities on a Folium Map of India



The above map enables you to view cities tier-wise depending on the checked options in the Legend.

❖ DISCUSSION:

Let us see the average values for each tier



Tier	Avg_Population	Avg_Medical	Avg_Education	Avg_Food	Avg_Shops	Avg_Transport
Tier 1	3482382	50	44	48	49	42
Tier 2	1142541	40	21	25	39	11
Tier 3	827905	15	10	9	12	4
Tier 4	325251	2	1	1	2	0

Inferences:

1. Tier 1 & 2 cities have high population due to higher facilities
2. Tier 3 & 4 cities have low facilities and low population
3. Scope for improvement in facilities of tier 4 cities. Good opportunity for private & public corporations to invest in tier 4 cities.
4. Most tier 4 cities have very limited means of public transport. This situation can be improved.

❖ CONCLUSION:

I have implemented K-Means for clustering Indian Cities based on availability of facilities within a radius of 5 KM with the help of Foursquare API.