

***CLUSTERING APPROACH TO DATA MINING
FOR IDENTIFYING LOCALITIES IN
PUNE, INDIA WHERE THE HEALTH CARE
INFRASTRUCTURE NEEDS TO BE IMPROVED***



By-Vedant Asawale

Contents



1.BACKGROUND



2.PROBLEM
STATEMENT



3.DATA



4.METHODOLOGY



5.RESULTS



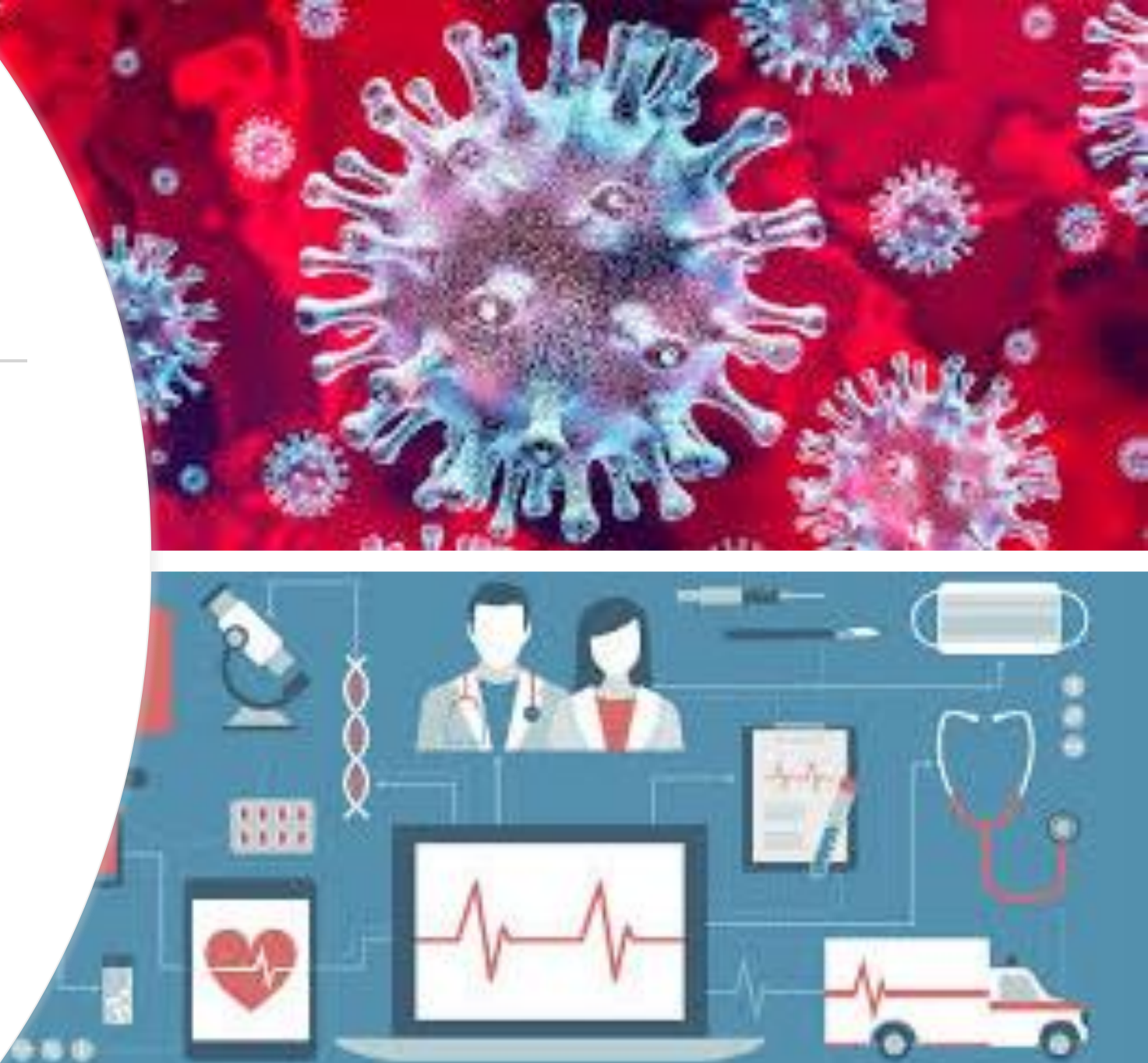
6.FUTURE SCOPE



7.CONCLUSION

Background

- 1.Covid-19 and its effect on the Humanity
- 2.Lack of adequate Health Care Infrastructure.
- 3.Prepare for future virus crisis.
- 4.Need to identify regions with severe need for new Hospitals.

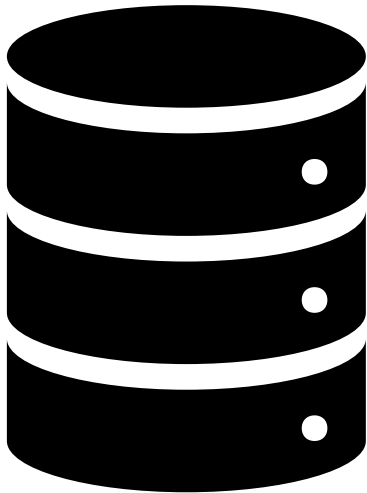


Problem Statement



- Identify the localities or regions in Pune city, India where the number of hospitals are less and hence there is urgent need to improve the Health care infrastructure to prepare for any epidemic or disaster crisis which might arise in future.

Data

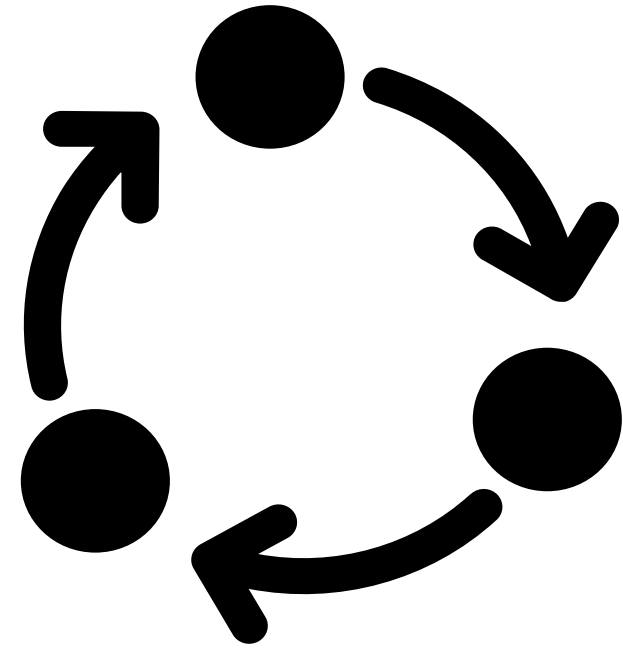


The following data was used in the analysis

- 1. List of all localities present in Pune.
- 2. Geographical coordinates of all localities
- 3. Details of all hospitals present in each locality within 3km radius.

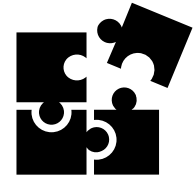
Methodology

- 1.Importing relevant libraries
- 2.Web-Scraping.
- 3.Extracting Geographical coordinates.
- 4.Preliminary Visualization.
- 5.Fetching Location Data.
- 6.Extracting Relevant Feature.
- 7.Finding the optimal number of Clusters.
- 8.Clustering.
- 9.Final Visualization.
- 10.Find clusters with highest need for new Health Care infrastructure.



Importing relevant libraries

- The following libraries were used in this Project.
- 1.Pandas
- 2.Numpy
- 3.BeautifulSoup
- 4.Folium
- 5.Geopy
- 6.Json
- 7.Requests
- 8.Sklearn
- 9.Yellowbrick



Web-Scraping.

- The next step was to create a list of all localities in the Pune city.
- This was done through web scraping.
- The URL ["https://www.mapsofindia.com/pune/localities/"](https://www.mapsofindia.com/pune/localities/) was used to get the list of all localities in Pune using the technique of web scraping via the [Beautiful Soup API](#).

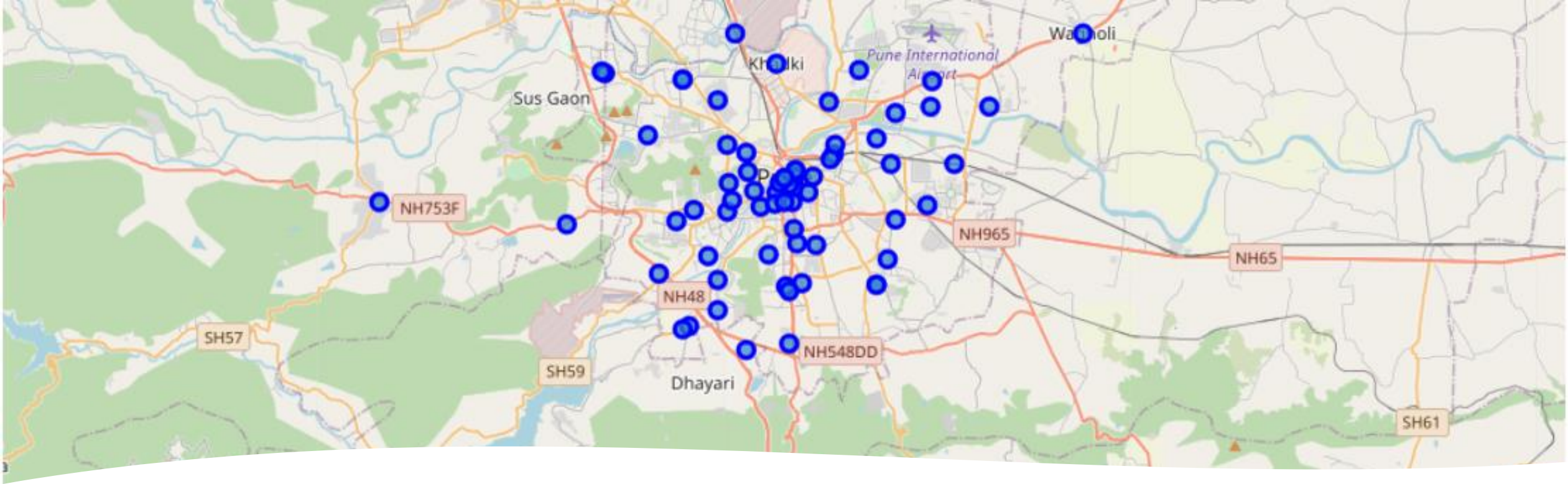
Pune Locality maps		
▶ Alandi Road	▶ Hadapsar	▶ Navi Peth
▶ Ambegaon Budruk	▶ Hadapsar Industrial Estate	▶ Padmavati
▶ Anandnagar	▶ Hingne Khurd	▶ Parvati Darshan
▶ Aundh	▶ Jangali Maharaj Road	▶ Pashan
▶ Aundh Road	▶ Kalyani Nagar	▶ Paud Road
▶ Balaji Nagar	▶ Karve Nagar	▶ Pirangut
▶ Baner	▶ Karve Road	▶ Prabhat Road
▶ Baner road	▶ Kasba Peth	▶ Pune Railway Station
▶ Bhandarkar Road	▶ Katraj	▶ Rasta Peth
▶ Bhavani Peth	▶ Khadaki	▶ Raviwar Peth
▶ Bibvewadi	▶ Khadki	▶ Sadashiv Peth
▶ Bopodi	▶ Kharadi	▶ Sahakar Nagar
▶ Budhwar Peth	▶ Kondhwa	▶ Salunke Vihar
▶ Bund Garden Road	▶ Kondhwa Budruk	▶ Sasson Road
▶ Camp	▶ Kondhwa Khurd	▶ Satara Road
▶ Chandan Nagar	▶ Koregaon Park	▶ Senapati Bapat Road
▶ Dapodi	▶ Kothrud	▶ Shaniwar Peth
▶ Deccan Gymkhana	▶ Law College Road	▶ Shivaji Nagar
▶ Dehu Road	▶ Laxmi Road	▶ Shukrawar Peth
▶ Dhankawadi	▶ Lulla Nagar	▶ Sinhadgad Road
▶ Dhayari Phata	▶ Mahatma Gandhi Road	▶ Somwar Peth
▶ Dhole Patil Road	▶ Mangalwar peth	▶ Swargate

Extracting Geographical coordinates

	Locality	Latitude	Longitude
0	Alandi Road	18.5523	73.8733
1	Ambegaon Budruk	18.4511	73.8376
2	Anandnagar	18.5083	73.8152
3	Aundh	18.5619	73.8102
4	Aundh Road	18.5619	73.8102
...
76	Viman Nagar	18.5214	73.8545
77	Wagholi	18.5806	73.9833
78	Wanowrie	18.4884	73.8987
79	Warje	18.482	73.8002
80	Yerawada	18.5656	73.8866

81 rows × 3 columns

- Thereafter the geographical coordinates ie latitude and longitude of all localities were obtained.
- The Geopy library along with Nominatim geocoder was used for this purpose



Preliminary Visualization

- Getting a sense of the data before performing any analysis is always a good idea.
- A map of all our localities in Pune was plotted using the **Folium** library.

Fetching Location Data

- We needed information of all the hospitals present within 3km radius of each locality present in Pune city
- The FourSquare API was used to fetch the venue data for our localities.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alandi Road	18.552349	73.873259	Ruby Hall Clinic	18.533768	73.876809	Hospital
1	Alandi Road	18.552349	73.873259	Serene Hospital	18.572227	73.879019	Hospital
2	Alandi Road	18.552349	73.873259	Inlaks & Budhrani Hospital	18.535327	73.887502	Hospital
3	Alandi Road	18.552349	73.873259	Jehangir Hospital	18.530495	73.876567	Hospital
4	Alandi Road	18.552349	73.873259	Sahaydri Hospital	18.554218	73.897066	Hospital
...
667	Warje	18.482044	73.800170	Deoyani hospital	18.494410	73.812720	Hospital
668	Warje	18.482044	73.800170	Shashwat hospital	18.495182	73.813535	Hospital
669	Warje	18.482044	73.800170	Sahyadri Hospital	18.507457	73.805752	Hospital
670	Yerawada	18.565632	73.886576	Serene Hospital	18.572227	73.879019	Hospital

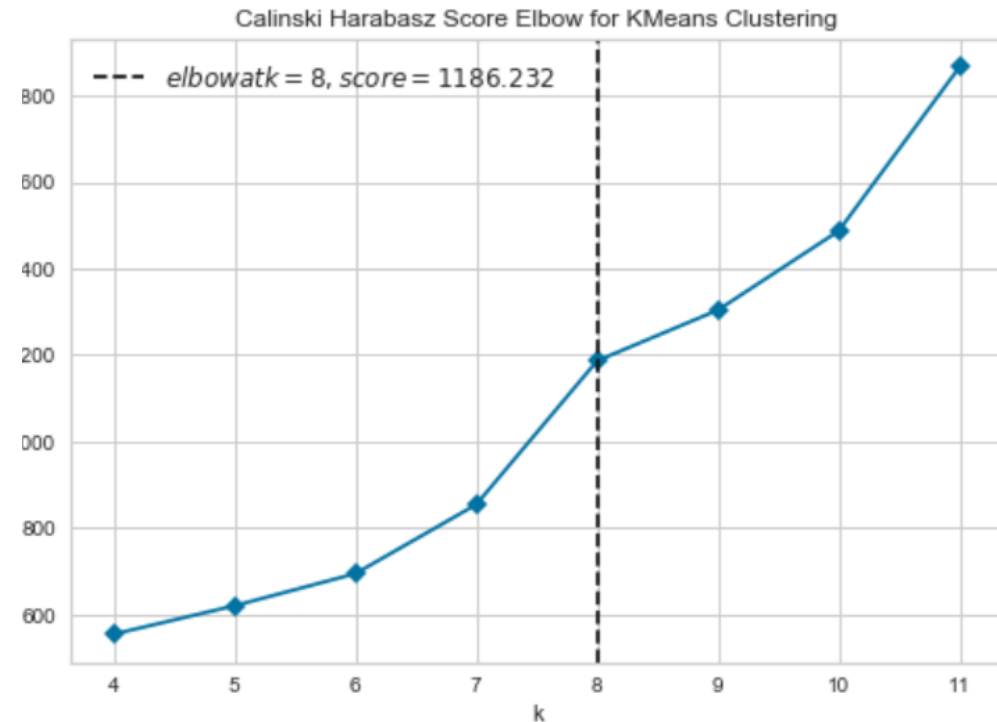
Extracting Relevant Feature

- In our project, the clustering of localities was performed on the basis of number of hospitals present in those localities.
- Thus, the relevant feature for our project was the count of hospitals.
- The count of hospitals for each locality was calculated by aggregating the data for each locality from the data which was returned by FourSquare API.

	Locality	Locality Latitude	Locality Longitude	Hospital Count
0	Alandi Road	18.552349	73.873259	6
1	Ambegaon Budruk	18.451118	73.837555	3
2	Anandnagar	18.508338	73.815208	10
3	Aundh	18.561883	73.810196	8
4	Aundh Road	18.561883	73.810196	8
...
74	Wadgaon Sheri	18.550441	73.917167	4
75	Wagholi	18.580630	73.983310	3
76	Wanowrie	18.488368	73.898667	6
77	Warje	18.482044	73.800170	4
78	Yerawada	18.565632	73.886576	2

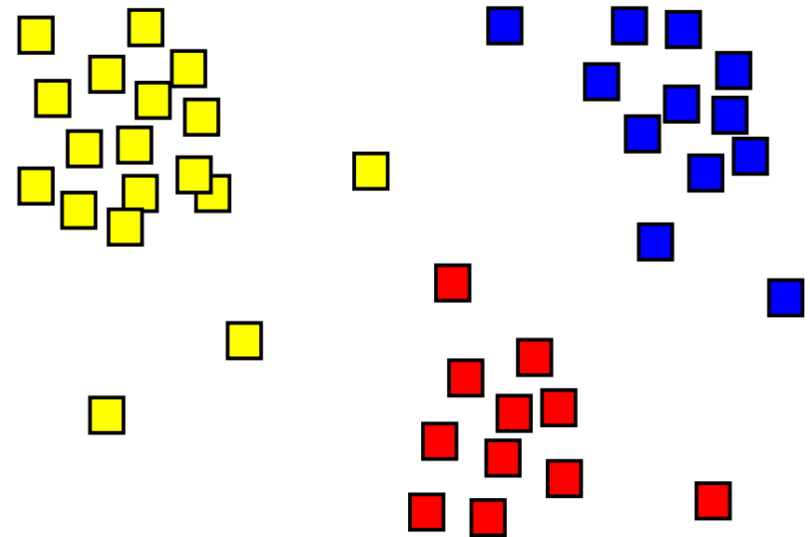
Finding the optimal number of Clusters.

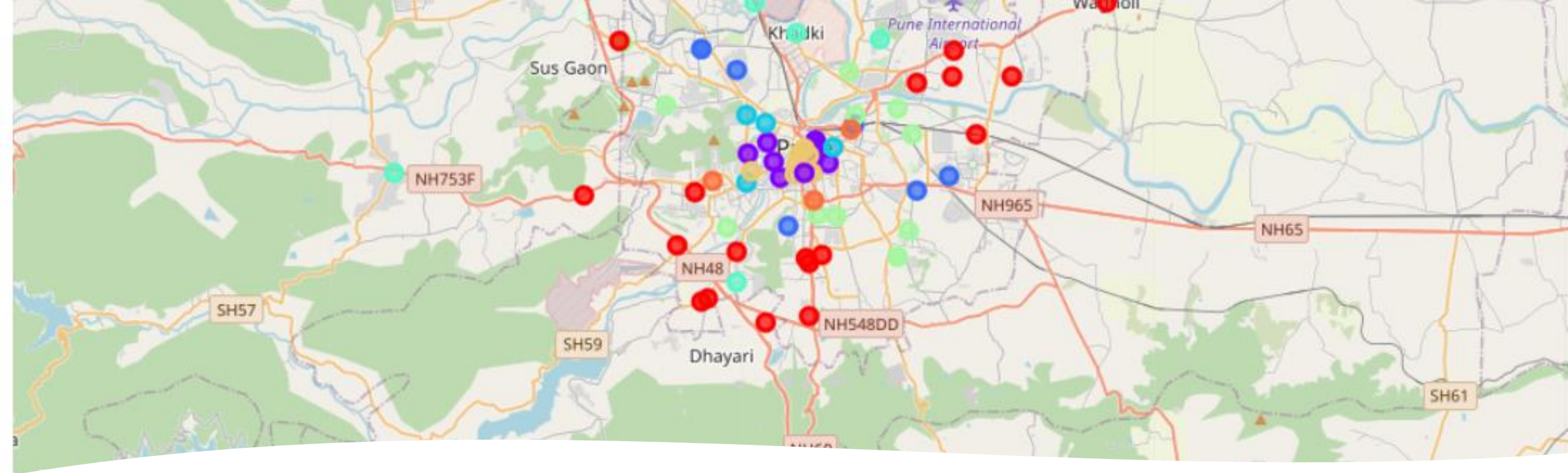
- We used the elbow method along with calinski harabasz score to find the optimal number of cluster(8)
- The score is defined as ratio between the within-cluster dispersion and the between-cluster dispersion.



Clustering

- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)
- **Kmeans algorithm** is easy to perform and most popular clustering algorithm and hence the same algorithm was used.
- Number of hospitals was used as a feature.





Final Visualization

- The Localities were visualized post clustering with unique color assigned to each cluster.
- The Folium library was used to create this map of clustered localities.

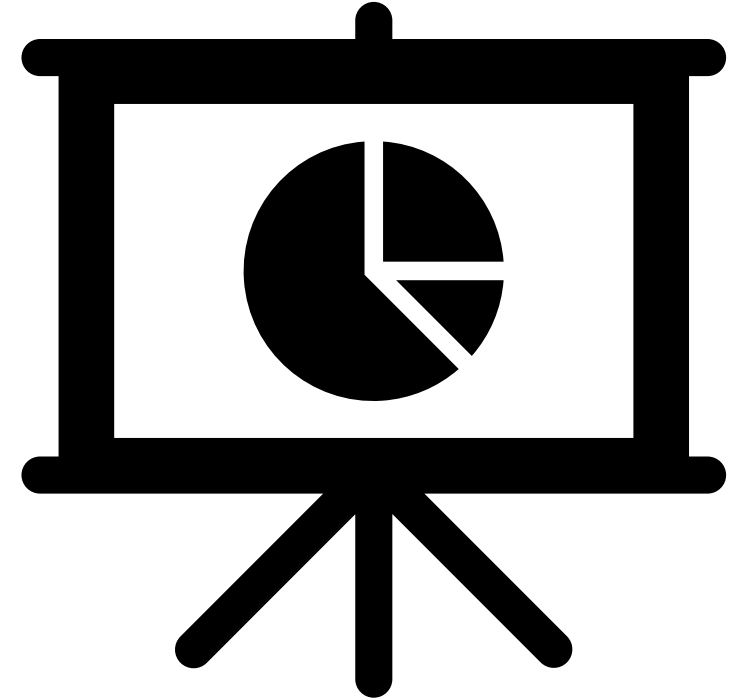
Find clusters with highest need for new Health Care infrastructure

- The clusters with the lowest number of mean hospitals were identified as the Highest priority clusters.
- The cluster centroids represented the mean number of hospitals value for its cluster.
- The centroids were sorted in ascending order and top 2 clusters were selected as the highest priority regions.

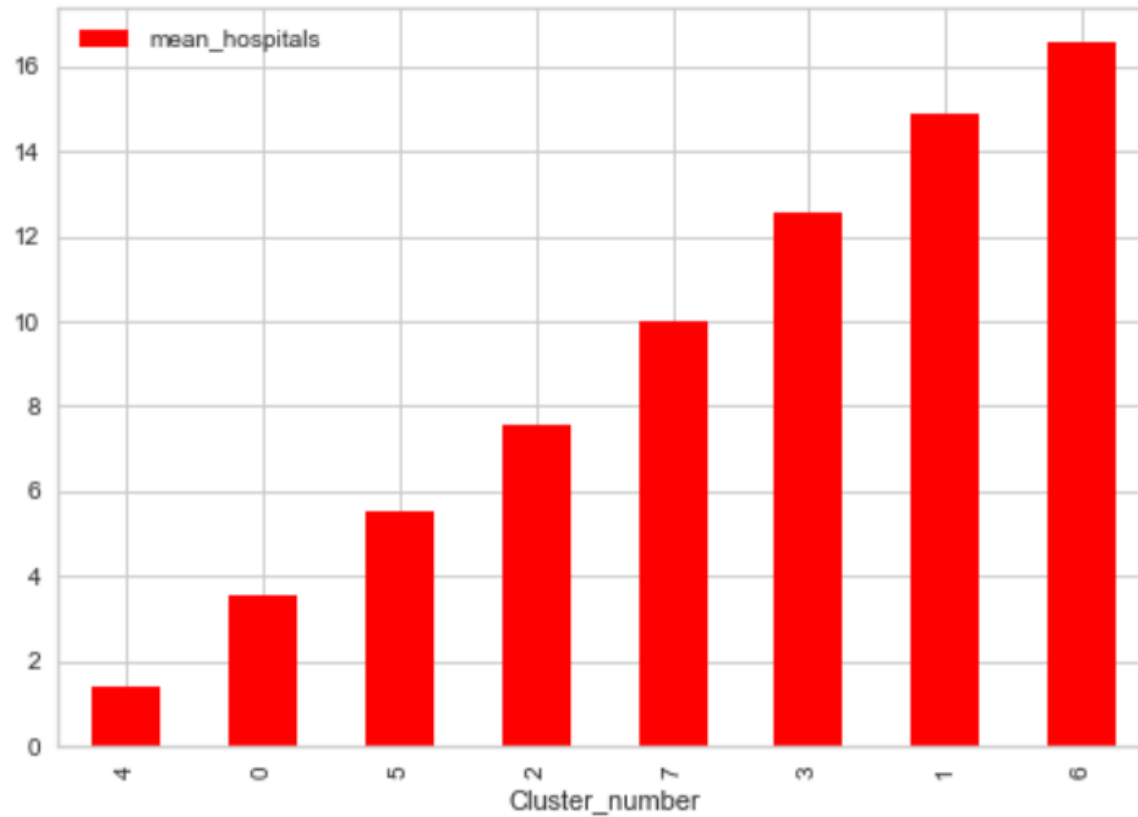


Results.

- 1) 8 Clusters of localities in the Pune city based on number of hospitals were formed.
- 2) Cluster number 4 had the lowest mean number of hospitals with a value of 1.42 hospitals.
- 3) Cluster number 0 had the second lowest mean number of hospitals with a value of 3.57 hospitals.
- 4) Clusters 7, 3, 1 and 6 had mean number of hospitals greater than 10.
- 5) Thus, localities belonging to Cluster 4 have the highest priority followed by those belonging to Cluster 0 for creation of new Health Care infrastructure



Clusters and mean hospital values



Cluster Number	Mean Hospitals
0	3.57
1	14.9
2	7.57
3	12.57
4	1.42
5	5.53
6	16.54
7	10.00


Table 1: Clusters and Mean Hospitals

Localities with highest priority.

Cluster4

Cluster Labels		Locality	Locality Latitude	Locality Longitude	Hospital Count
15	4	Dapodi	18.580846	73.832775	1
17	4	Dehu Road	18.680047	73.734331	1
37	4	Khadki	18.568175	73.850779	2
44	4	Laxmi Road	18.141836	74.561703	2
57	4	Pirangut	18.511282	73.679007	1
72	4	Vadgaon Budruk	18.467497	73.825365	1
78	4	Yerawada	18.565632	73.886576	2

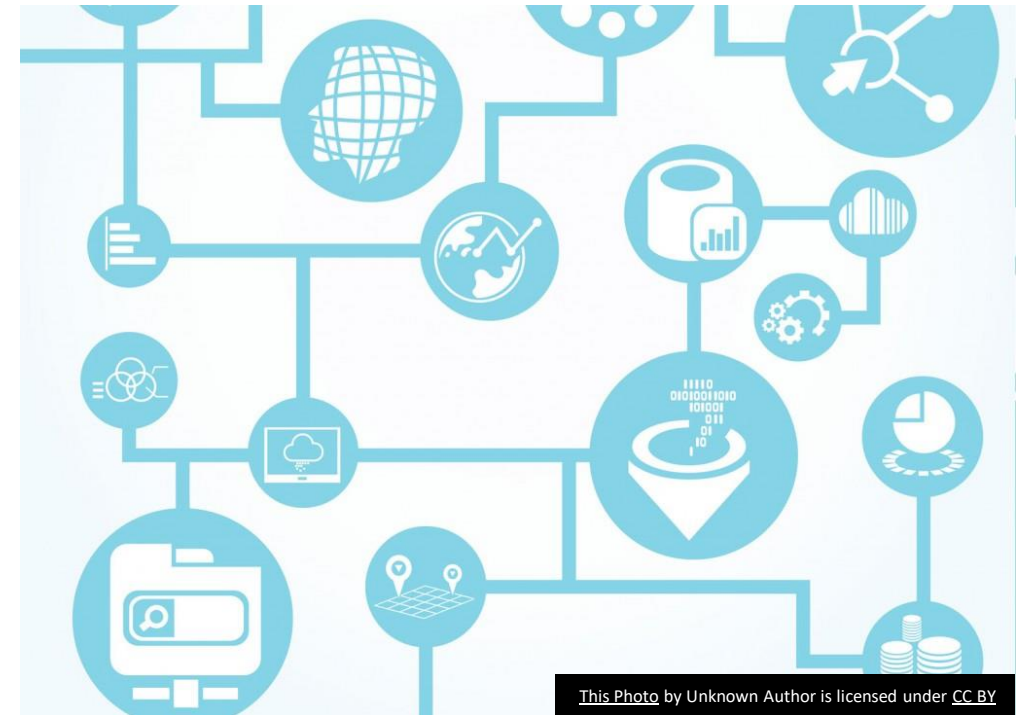
Cluster Labels		Locality	Locality Latitude	Locality Longitude	Hospital Count
1	0	Ambegaon Budruk	18.451118	73.837555	3
7	0	Baner road	18.564642	73.775079	4
10	0	Bibvewadi	18.478174	73.862105	3
14	0	Chandan Nagar	19.121709	72.923203	4
18	0	Dhayari Phata	18.460862	73.812755	3
29	0	Hadapsar	18.526967	73.927825	4
31	0	Hingne Khurd	18.479670	73.825099	3
32	0	Kalyani Nagar	18.548138	73.902551	4
36	0	Katraj	18.453679	73.856320	4
38	0	Kharadi	18.550518	73.942494	4
42	0	Kothrud	18.503889	73.807673	4
50	0	Nagar Road	18.561016	73.918279	4
54	0	Padmavati	18.477317	73.854770	3
56	0	Paud Road	18.502704	73.760385	4
63	0	Satara Road	18.475207	73.856165	3
68	0	Sinhagad Road	18.459608	73.809984	3
74	0	Wadgaon Sheri	18.550441	73.917167	4
75	0	Wagholi	18.580630	73.983310	3
77	0	Warje	18.482044	73.800170	4

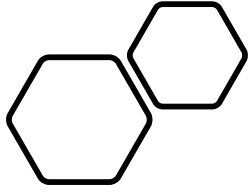


Localities with
second highest
priority.

Future Scope

- 1.The project analyzed the Pune city only,however the same analysis can be repeated for other cities which will help us improve the Health Care facilities throughout the World.
- 2.Once we have sufficient data present for all cities,a centralized web or stand alone application can be created wherein by entering the city name one could fetch such statistics in a single click and within seconds.





Conclusion



To conclude, this project was an attempt to find out the localities in Pune, India where new Health Care infrastructure needs to be set up urgently.



We found out that there are some localities where the number of hospitals are extremely less and efforts need to be taken to improve the situation.



We also understood that most of these localities lie in the outskirts of the city whereas the situation is comparatively better in the central part.



The End