Project Report

Author- Utkarsh Gupta, Himani Madaan, Shalini Kumari, Divyansh Khandelwal, Tanish Gupta

Date- 18th November,2020.

Definition

Project Overview

Google Maps API is a geospatial API and the Places API specifically can be used to find about the neighbourhood of a place. Places API can be used to find out about nearby places like restaurants, hospitals, schools, cafes, bakeries and so on. We used clustering and segmentation to find about a specific quality of service (for example healthcare including hospitals, doctors, pharmacies etc.) in the neighbourhood. We used data visualisation techniques to get an idea of different neighbourhoods in the city. At the end we achieved a model in which if we input a specific latitude and longitude of the city of Ahmedabad, it will tell the user about the quality of service in that area in the field of medical care, education and food. It will also label the area taking into account the overall development.

Business Problem

The objective of this project would be to analyse the neighbourhoods of the city of Ahmedabad in India, to tell us about the quality of hospitals, food and schools in any particular area. Moreover, using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question:

In the city of Ahmedabad, if a property developer is looking to open a new hospital, restaurant or school, where would you recommend that they open it?

In this project, we implemented an algorithm to find the best place to start a business where there is high demand and no (or very few) supply. We measure the quality of recommendation in terms of average service rating and customer-business ratio.

Metrics

After collecting a much of features for medical care which includes:

1. Number of hospitals & their mean rating
2. Number of dentists & their mean rating
3. Number of doctors & their mean rating

4. Number of pharmacies & their mean rating
5. Number of physiotherapist & their mean rating

Similarly features for food include:

1. Number of bakeries & their mean rating
2. Number of cafes & their mean rating
3. Number of convenience stores & their mean rating
4. Number of tiffin services & their mean rating
5. Number of groceries/supermarkets & their mean rating
6. Number of meal delivery services & their mean rating
7. Number of restaurants & their mean rating

And features for education include:

1. Number of schools & their mean rating
2. Number of university & their mean rating
3. Number of book stores & their mean rating
4. Number of museums & their mean rating
5. Number of libraries & their mean rating

We used these features and formulated a scoring function which will imply a final medical score (and other two as well) for the neighbourhood and then we will use K-means clustering to analyse how our data is clustered and labelled. This will thereby divide the neighbourhoods based on the quality of medical service into different tiers.

Analysis

Data Collection and Exploration

Web Scraping

We firstly used Web Scraping to extract all the neighbourhoods of the city of Ahmedabad for analysis. After that we used GeoCoder to find out all the latitudes and longitudes of the neighbourhoods. After this point we will have the names of all the neighbourhoods of the city and their respective latitude and longitude.

Places API

We now used Google Places API to find more about the neighbourhoods. Use this link to read the documentation of the API. Let's look at the type of data we can get from the API.

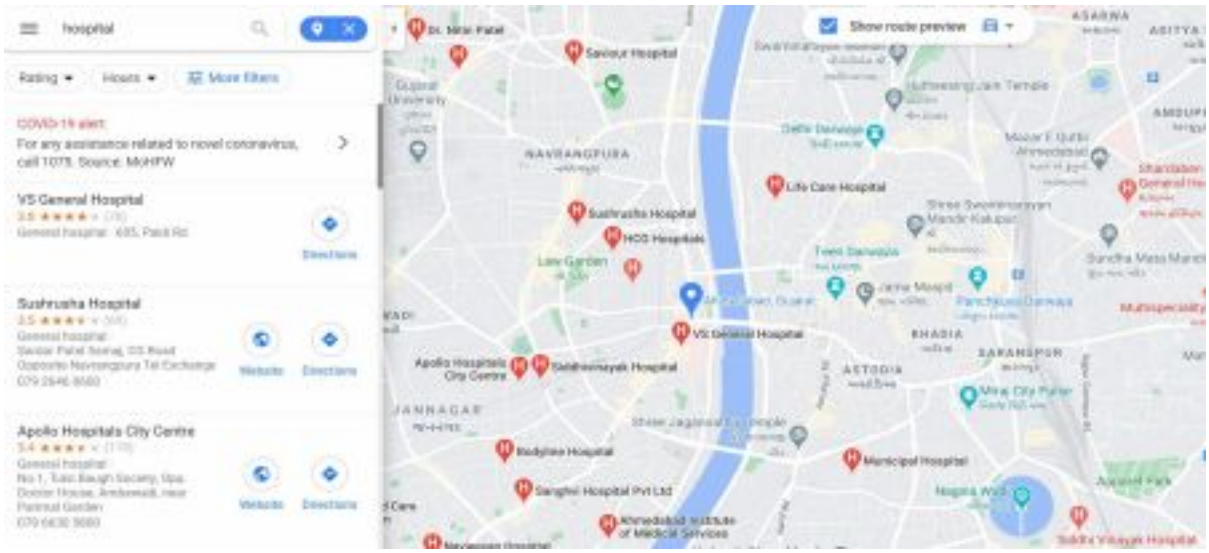


Figure 1 Google Maps

We can see in the image above that we can surf through nearby hospitals around a particular latitude and longitude. We can thus collect the names of the hospitals and their ratings as well. You will find it in the documentation of the Places API that we can adjust the radius of search around a particular point. In this project I've chosen the radius to be 1km.

	A	B	C	D	E	F
1	Neighbourhood	Latitude	Longitude	Venue_typ	Venue_Name	Venue_Rating
2	Agol	23.02776	72.60027	hospital	Victoria Jubilee Hospital	4
3	Agol	23.02776	72.60027	hospital	Al Ameen Hospital|	
4	Agol	23.02776	72.60027	hospital	Dr.Tanumati Shah Hospital	
5	Agol	23.02776	72.60027	hospital	Lokhandwala General Hospital	4.2
6	Agol	23.02776	72.60027	hospital	shreeShreeji Pathology Laboratory	
7	Agol	23.02776	72.60027	hospital	Government dental College Ahmedabad	5
8	Agol	23.02776	72.60027	hospital	Victoria Jubilee Hospital Trust	5
9	Agol	23.02776	72.60027	hospital	Chinubhai Baronet Hospital	5
10	Agol	23.02776	72.60027	hospital	Shree Jee Eye & Orthopedic Hospital	
11	Agol	23.02776	72.60027	hospital	Mahavir Medical & Surgical	5
12	Agol	23.02776	72.60027	hospital	General Medical Store	5
13	Agol	23.02776	72.60027	hospital	Kalupur Railway Health Unit	4.7
14	Agol	23.02776	72.60027	hospital	Rahi sarvjanik clinic	
15	Agol	23.02776	72.60027	hospital	Divya Children`s Hospital	5
16	Agol	23.02776	72.60027	hospital	Shaheen dispensary	
17	Agol	23.02776	72.60027	hospital	ghatdi	4.8
18	Agol	23.02776	72.60027	hospital	Mahak Hospital	
19	Agol	23.02776	72.60027	hospital	Shreeji Childrenâ€™s Hospital Dr Kushal M	5
20	Agol	23.02776	72.60027	hospital	Samast Bangali Samaj Accosion Trust	
21	Agol	23.02776	72.60027	hospital	Gulab Bhai Hospital	3

	A	B	C	D	E	F
166	Ahmedabad Cantonment	23.02776	72.60027	doctor	Atul Plastic	
167	Ahmedabad Cantonment	23.02776	72.60027	doctor	Dinesh S Agrawal	
168	Ahmedabad Cantonment	23.02776	72.60027	doctor	Shaan Pathology Laboratory	5
169	Ahmedabad Cantonment	23.02776	72.60027	doctor	Bhatt Dr Yashesh Rameshchandra	
170	Ahmedabad Cantonment	23.02776	72.60027	doctor	Dr Nayan Kantilal Bhatt	
171	Ahmedabad Cantonment	23.02776	72.60027	doctor	Dani Dr Dineshchandra Ambalal	
172	Ahmedabad Cantonment	23.02776	72.60027	doctor	Prajapati Dr Gaurang R	5
173	Ahmedabad Cantonment	23.02776	72.60027	doctor	Nayak Pravinchandra C	
174	Ahmedabad Cantonment	23.02776	72.60027	doctor	Limbachiya Dr Vinubhai R	
175	Ahmedabad Cantonment	23.02776	72.60027	doctor	Deep Orthopaedics	
176	Ahmedabad Cantonment	23.02776	72.60027	doctor	Shah Dr Rajendra H	
177	Ahmedabad Cantonment	23.02776	72.60027	doctor	Dr Jamil Ahmed Shaikh	
178	Ahmedabad Cantonment	23.02776	72.60027	doctor	Rajesh M Shah	
179	Ahmedabad Cantonment	23.02776	72.60027	doctor	Atul R Parikh	
180	Ahmedabad Cantonment	23.02776	72.60027	doctor	Dr. Imran Mansuri (Hanifa Clinic)	4.7
181	Ahmedabad Cantonment	23.02776	72.60027	doctor	HIJAMA CENTRE	5
182	Ahmedabad Cantonment	23.02776	72.60027	doctor	Pandya Dr Nimish C	
183	Ahmedabad Cantonment	23.02776	72.60027	doctor	Omkar Jyotish	
184	Ahmedabad Cantonment	23.02776	72.60027	doctor	Dr Shailesh D Ravat	5
185	Ahmedabad Cantonment	23.02776	72.60027	doctor	Shah Dr Raju Haribhai	4.3
186	Ahmedabad Cantonment	23.02776	72.60027	doctor	Gandhi Dr Govindlal M	
187	Ahmedabad Cantonment	23.02776	72.60027	doctor	Dr.Asiya .A. Kazi	
188	Ahmedabad Cantonment	23.02776	72.60027	doctor	HARDVED AND MALAM PATTA	5

We can observe in the above images the type of data we collect. It contains the neighbourhood names, latitude and longitude of the neighbourhoods, Venue Type, name of the venue and it's rating.

Algorithms & Scoring Techniques

Algorithms include the following things:

- Formulate Scoring Function.
- K-Means and PCA for **Unsupervised Learning.**
- **Haversine Formula** to calculate distance between two coordinates.

After collecting all the data points, we have to formulate a scoring function so as to create an overall medical score for that particular neighbourhood. Before that we first have to find out the mean rating for each service in the neighbourhood and then we can assign different weights to different features to calculate overall score.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Neighbour	Latitude	Longitude	Hospital_Count	Mean_hospital_rating	Doctor_Count	Mean_doctor_rating	Dentist_Count	Mean_dentist_	Pharmacy	Mean_pha	Physiother	Mean_physiotherapist_rating		
2	Agol	23.02776	72.60027	33	4.475	45	4.7125	12	4.8125	39	4.420833	0	0		
3	Ahmedaba	23.02776	72.60027	33	4.475	45	4.7125	12	4.8125	39	4.420833	0	0		
4	Alam Roza	23.00212	72.54979	60	4.489189189	60	4.168444444	32	4.942857143	60	4.351613	10	4.76		
5	Ambawadi	23.01885	72.55441	60	4.578571429	60	4.366767442	56	4.491548387	60	4.405714	10	4.542857		
6	Amraiwadi	23.00735	72.62268	47	4.144444444	27	4.607692308	16	4.655555556	40	3.968182	3	4.95		
7	Anand Nag	23.01139	72.51712	60	4.568181818	60	4.548717949	60	4.780487805	60	4.253488	8	4.94		
8	Asarwa	23.04708	72.60481	60	4.34047619	44	4.666666667	9	4.275	55	4.310714	2	5		
9	Asarwa Ch	23.04226	72.60457	60	4.294444444	36	4.807142857	5	4.4	37	4.227778	1			
10	Badarkha	22.84128	72.45453	0	0	0	0	0	0	1	5	0	0		
11	Bahiyal	23.02776	72.60027	33	4.475	45	4.7125	12	4.8125	39	4.420833	0	0		
12	Bapunagar	23.03476	72.63024	60	4.110810811	60	4.405714286	21	4.521428571	60	4.475758	5	4.7		
13	Bareja (Jan	22.8557	72.5949	17	4.672727273	1	1	2	3	3	4.75	0	0		
14	Behrampu	23.00278	72.57706	43	4.118181818	14	4.942857143	2	4.4	20	4.257143	0	0		
15	Bhadlad	22.3159	72.10697	0	0	0	0	0	0	0	0	0	0		
16	Bhairavnat	23.00258	72.59816	60	4.229545455	60	4.275	37	4.589655172	60	4.463889	8	4.75		
17	Bhojva	23.15832	72.01855	1		0	0	0	0	0	0	0	0		
18	Bopal	23.03032	72.47247	60	4.564814815	60	4.558333333	34	4.607407407	39	4.277273	3	3.633333		
19	Calico Mill	23.00098	72.57459	37	3.97	5	4.95	1	4	18	3.938462	0	0		
20	Choloda	22.80689	72.42511	0	0	0	0	0	0	0	0	0	0		
21	Chandkhe	23.11214	72.57989	60	4.263043478	33	4.419230769	27	4.7	50	3.933333	10	5		
22	Chandlodi	23.08729	72.54899	45	4.581081081	43	4.643333333	25	4.83	27	4.368421	7	4.9		
23	Dabhoda	23.02776	72.60027	33	4.475	45	4.7125	12	4.8125	39	4.420833	0	0		
24	Daniapur (	23.03807	72.59213	60	4.216129032	57	4.610526316	13	4.377777778	60	3.825926	0	0		

After this, we have to normalise the data by calculating z score for each column and then finally creating the medical score.

O	P	Q	R	S	T	U	V	W	X	Y
Hospital_Count_zscore	Mean_hospital_rating_	Doctor_Count_zsc	Mean_doctor_rating_	Dentist_Count_zscc	Mean_dentist_	Pharmacy_	Mean_pha	Physiother	Mean_phy	Medical_score
0.793972658	0.449003978	1.055490647	0.418442109	2.413634164	0.511188052	1.169094	0.187826	2.182731	0.88051	12.42033703
0.793972658	0.429042217	1.055490647	0.276254724	2.413634164	0.4974212	1.169094	0.215584	2.736967	0.831893	12.40565733
0.793972658	0.367520722	1.055490647	0.227936351	2.413634164	0.413956604	1.169094	0.209559	3.014085	0.858481	12.27613447
0.793972658	0.530063308	1.055490647	0.369417969	2.413634164	0.495721653	1.169094	0.295241	1.351377	0.93399	12.26900749
0.793972658	0.537254132	1.055490647	0.248678889	2.180335397	0.324341464	1.169094	0.419058	1.905613	0.765043	12.0137351
0.793972658	0.388816624	1.055490647	0.295656343	2.413634164	0.377967282	1.169094	0.126031	1.351377	0.868356	11.43507694
0.793972658	0.326744847	1.055490647	0.338371181	2.063686013	0.461702137	0.926531	0.296684	1.351377	0.93399	10.92887193
0.793972658	0.431875443	1.055490647	0.430597655	1.538763786	0.473998863	0.878018	0.204644	1.628495	0.945334	10.6768291
0.793972658	0.339385892	1.055490647	0.386026131	1.597088478	0.558995848	1.169094	0.270326	0.520024	0.848909	10.60243369
0.793972658	0.475391181	1.055490647	0.113519649	0.780542791	0.59269992	1.169094	0.375053	1.905613	0.857417	10.55910922
0.793972658	0.256205266	1.055490647	0.328609926	1.947036629	0.473246003	1.169094	0.204809	0.242906	0.916974	10.4483605
0.793972658	0.295687385	1.055490647	0.184738675	1.07216625	0.381743125	1.169094	0.466176	1.351377	0.853163	10.24387247
0.793972658	0.292880672	1.055490647	0.329965147	0.547244023	0.237182308	1.169094	0.407438	1.351377	0.821258	9.674280328
0.793972658	0.114488766	1.055490647	0.342971427	1.305465018	0.311603995	1.169094	-0.01472	1.351377	0.427758	9.24129471
0.793972658	0.340124431	1.055490647	0.371246844	0.430594639	0.384778251	1.169094	0.292403	-0.03421	0.959514	9.099982585
0.793972658	0.621296889	1.055490647	0.354218536	0.955516866	0.435205618	0.150329	0.348135	1.628495	0.902794	9.047015079
0.793972658	0.524764806	1.055490647	0.163978473	0.197295871	0.387921776	1.169094	0.310889	0.797142	0.508585	9.033414053
0.793972658	0.357907888	1.055490647	0.210040171	0.663893407	0.422051466	0.926531	0.425707	0.242906	0.885068	8.96186525
0.793972658	0.213509141	1.055490647	0.27293246	0.13897118	0.340993451	1.169094	0.476029	0.520024	0.831893	8.820712964
0.793972658	0.319282277	1.055490647	0.169646554	-0.094327588	0.04747811	1.169094	0.376	1.351377	0.59792	8.440670061
0.793972658	0.365929648	1.055490647	0.452211631	-0.210976972	0.420500117	1.169094	-0.03164	-0.31133	0.959514	8.042573103
0.793972658	0.292945793	1.055490647	0.281727742	0.372269947	0.364030992	0.635455	0.36265	-0.03421	0.938244	7.814409801
0.793972658	0.527732948	1.055490647	0.375905532	0.897192174	0.392345995	0.150329	0.314586	-0.03421	0.378127	7.516057382
0.793972658	0.29830904	-0.059464262	0.301562866	0.430594639	0.488998168	1.169094	0.42603	-0.03421	0.895704	7.496194664
0.793972658	0.318871931	-0.148660654	0.28205212	0.488919331	0.447648734	0.683968	0.034834	1.905613	0.959514	6.851959937
0.793972658	0.418852719	1.055490647	0.358376391	0.488919331	0.570089	-0.14075	0.266985	-0.31133	0.895704	6.637543674
0.793972658	0.209669517	0.698705076	0.420304326	0.022321796	0.268467858	0.295867	0.455079	0.520024	0.810622	6.51590544
0.793972658	0.372464413	0.341919505	0.448998743	-0.560925124	0.193809159	0.926531	0.341787	-0.31133	0.959514	6.478835154

Exploratory Visualisation

After collecting and playing with the data we have to visualise the data as well so as to understand important features. Also, visualisations provide interesting plots which can result in interesting results about the city. Python library Folium is used for the given visualisations.

- We looked into the different wards of the city and then imposed our data points on the regions of Ahmedabad.
- We plotted the data points on the map to show the location of each place for better understanding.
- We plotted all the places with different colours according to a certain criterion, let's say the number of pharmacies in the neighbourhood greater than 40 is encircled in green whereas less than 40 encircled in red.
- We plotted visualisation related to clusters on the map
- Visualisation depicting heatmap according to the number of services present will also help us analyse the neighbourhoods.
- Exciting bar plots will also tell us the best service providers' information.

Figure 2 Location of neighbourhoods on the map.

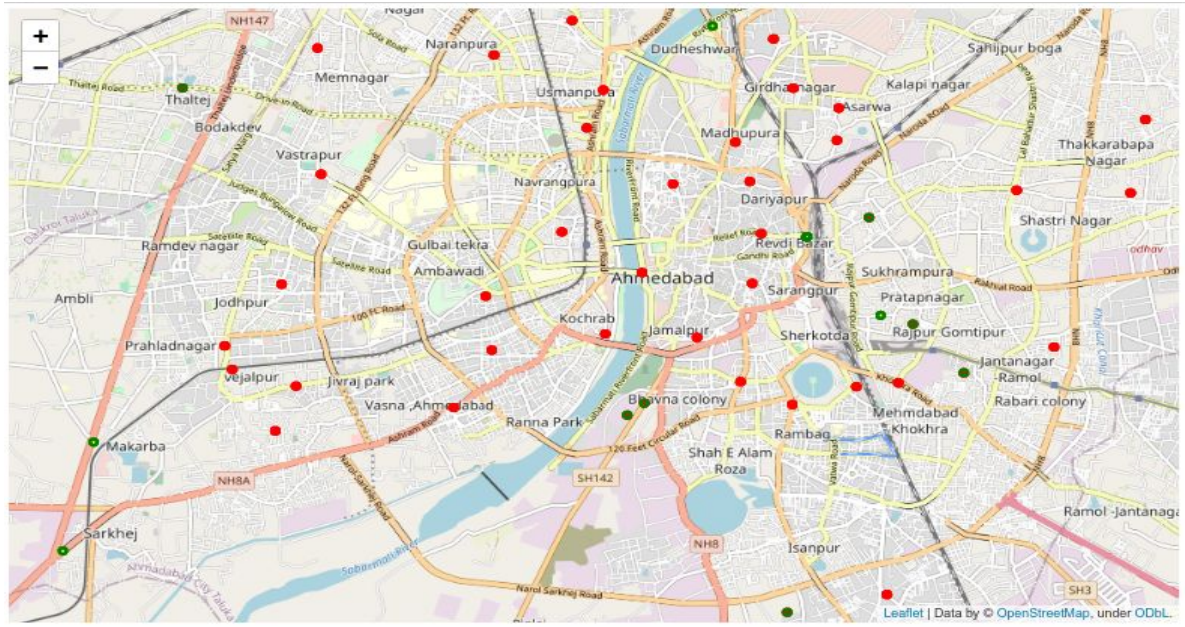


Figure 3 a – Clustering of different locations on the map

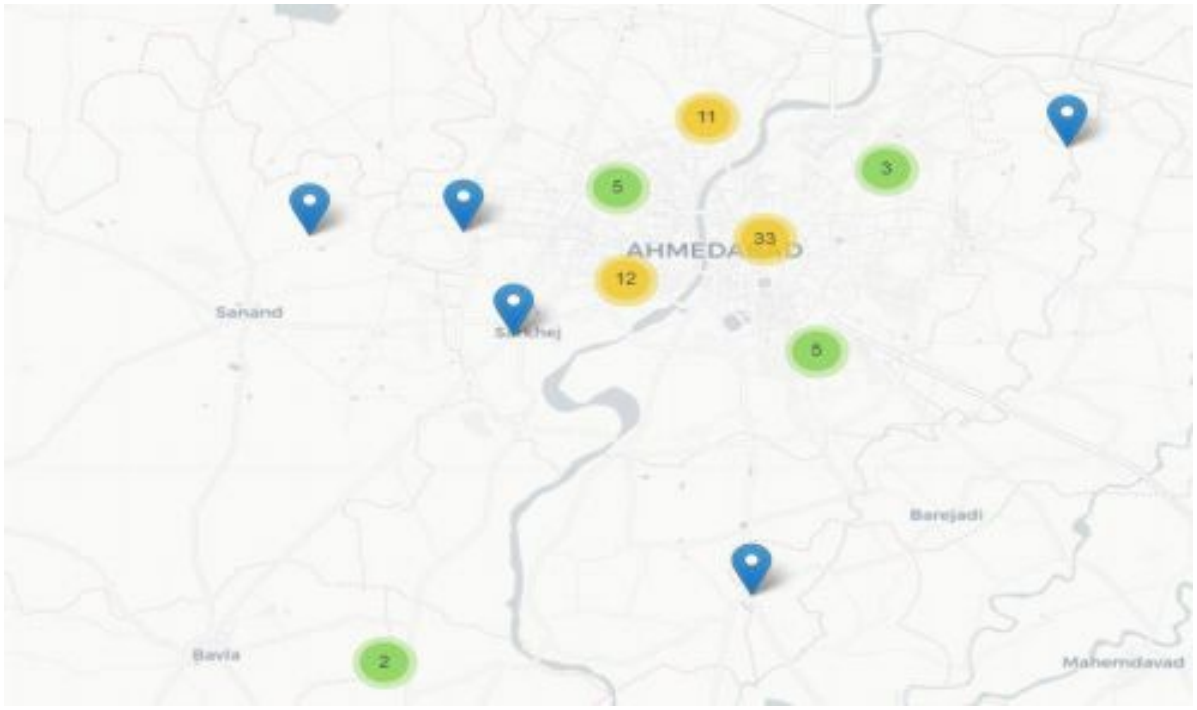


Figure 3 b – Clustering similar to Figure 3a but zoomed out

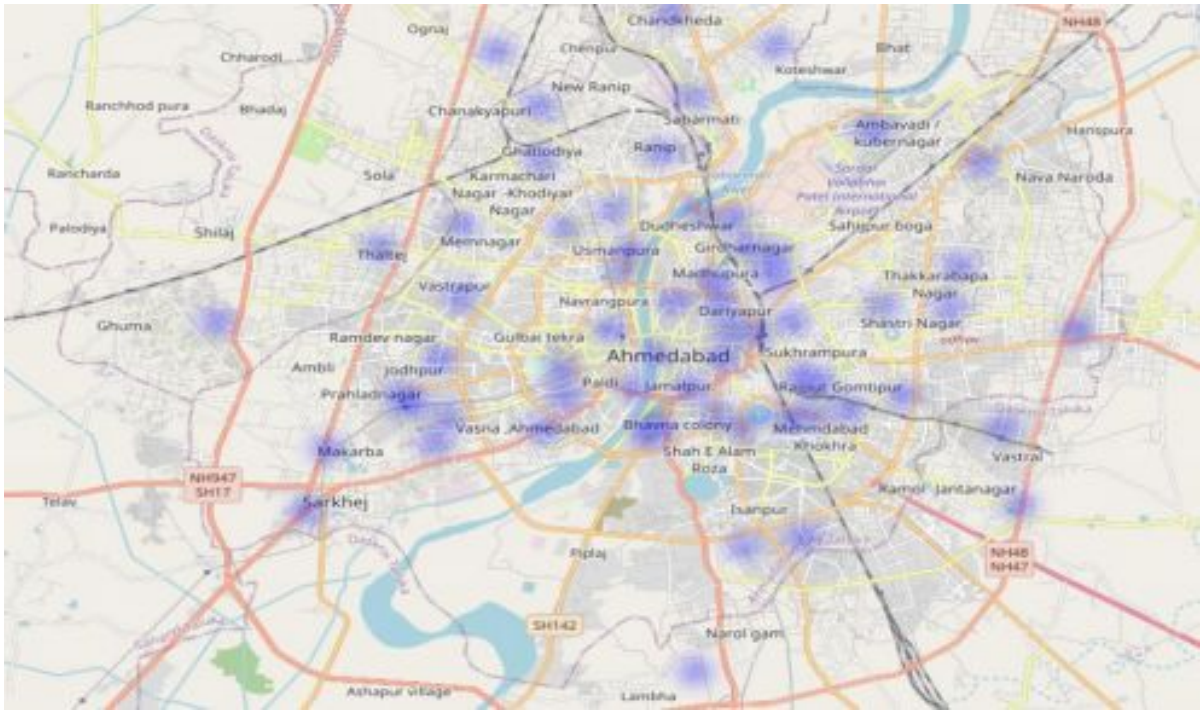


Figure 4 – Heat Map based on density

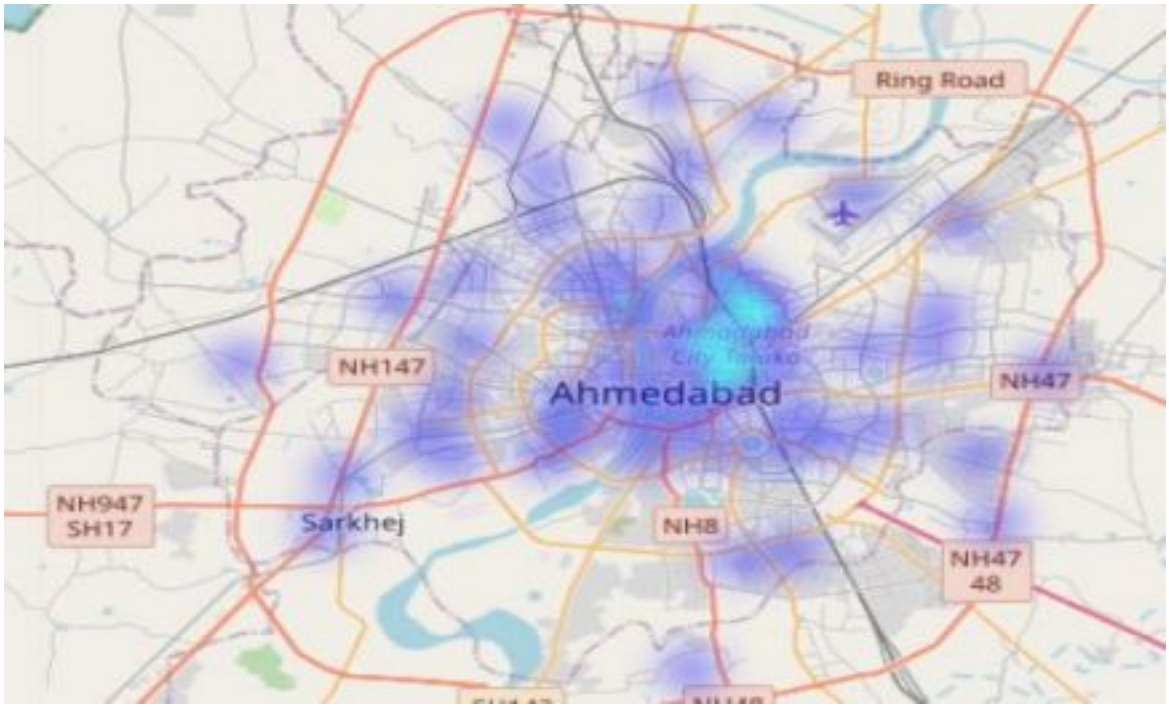


Figure 5 – Heat map of the same Figure 4 zoomed out. This shows the part that turned blue provides more service on medical care terms.

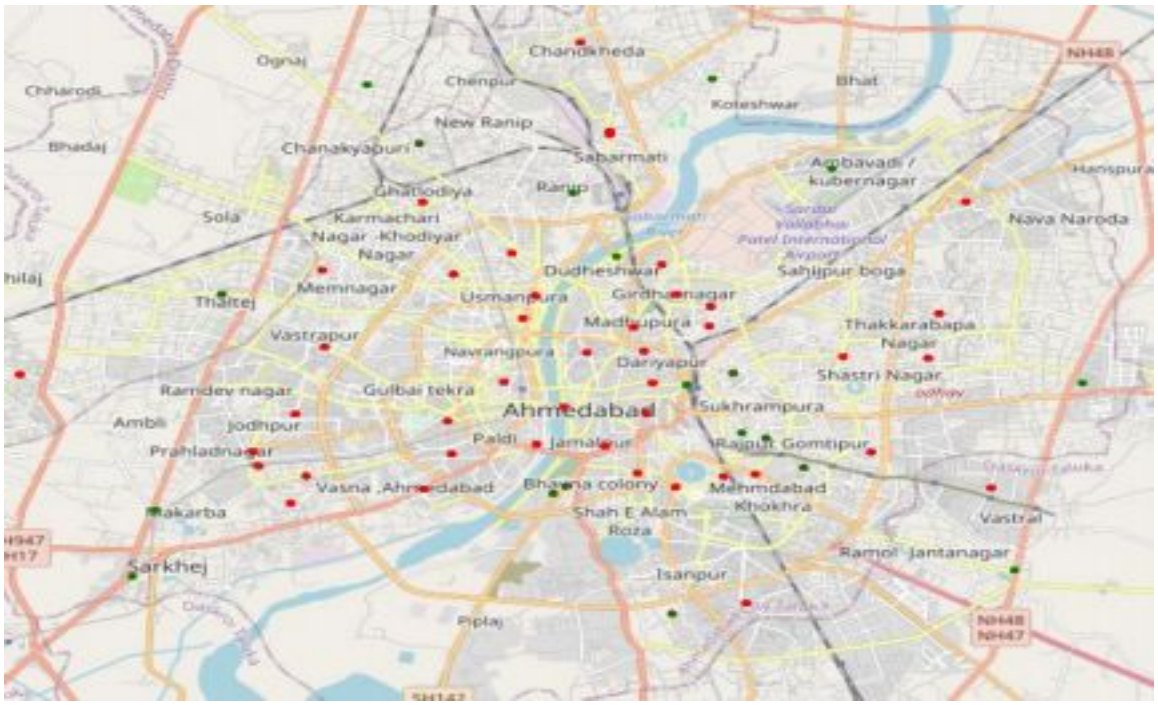


Figure 6 – Red and green dots depicting whether the number of pharmacies are less or greater than 40 respectively.

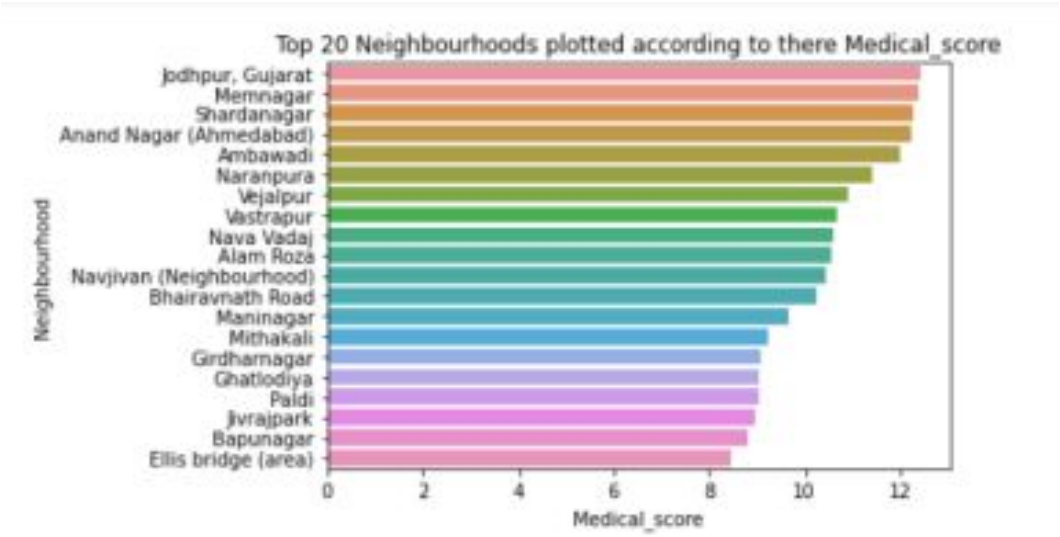


Figure 7 – Bar Plot showing the Top 20 neighbourhood accordingTo our medical score.

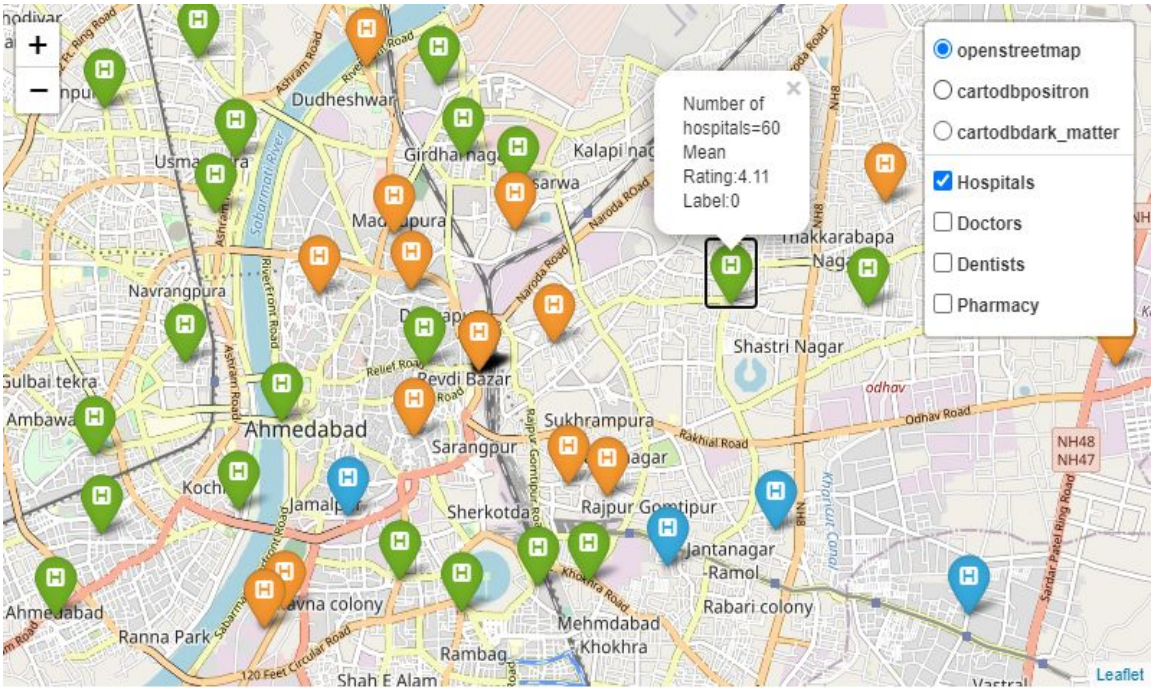


Figure 8 – An all-in-one map showing details of hospitals,doctors,pharmacy,dentists with icons coloured according to labels assigned in the dataset.One can also toggle between different themes in the map.

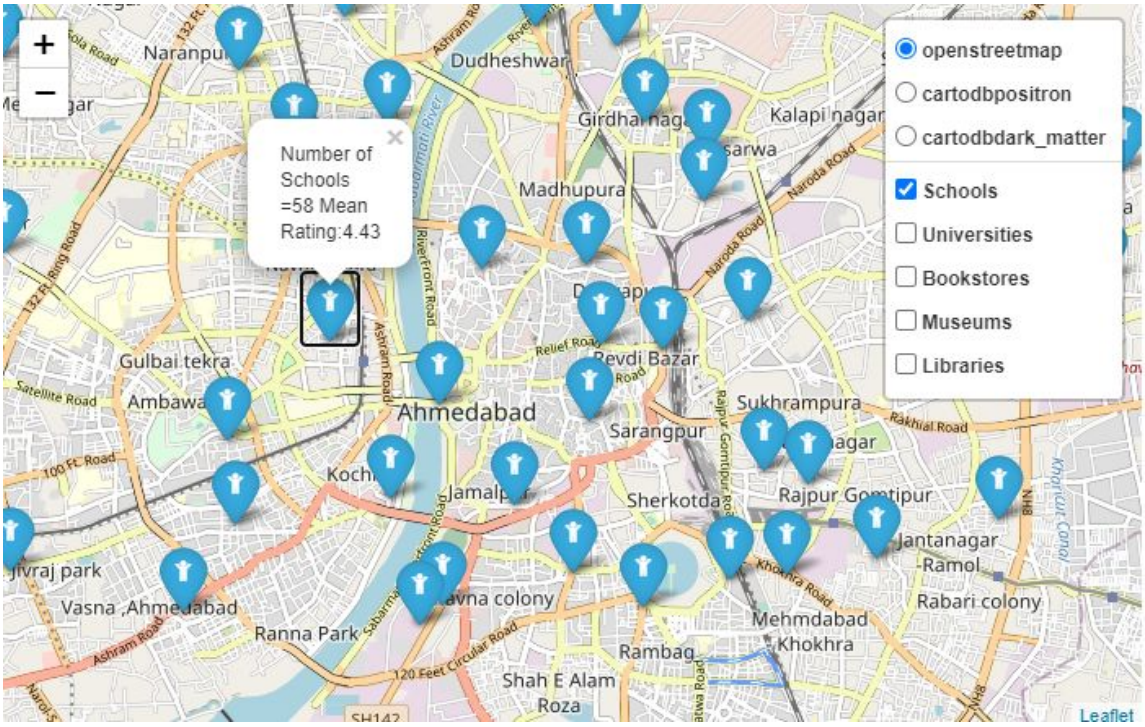


Figure 9 – An all-in-one map showing details of schools,universities,museums,bookstores and libraries.

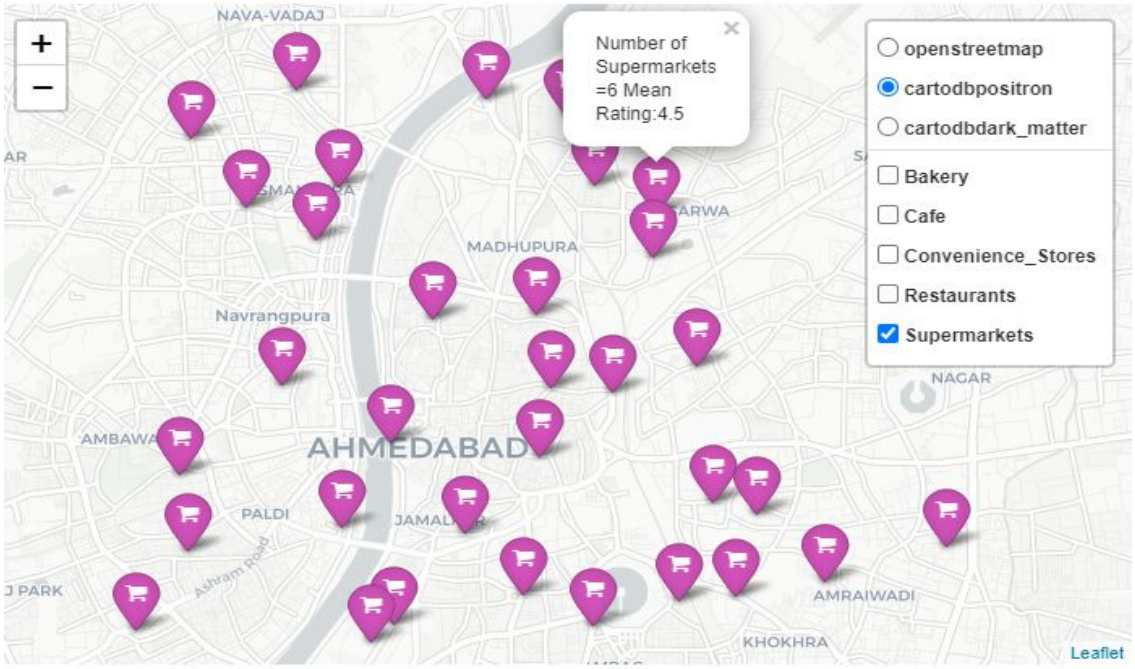


Figure 10 – An all-in-one map showing details of Bakery, Cafes, Convenience Stores, Supermarkets and Restaurants in regions of Ahmedabad.

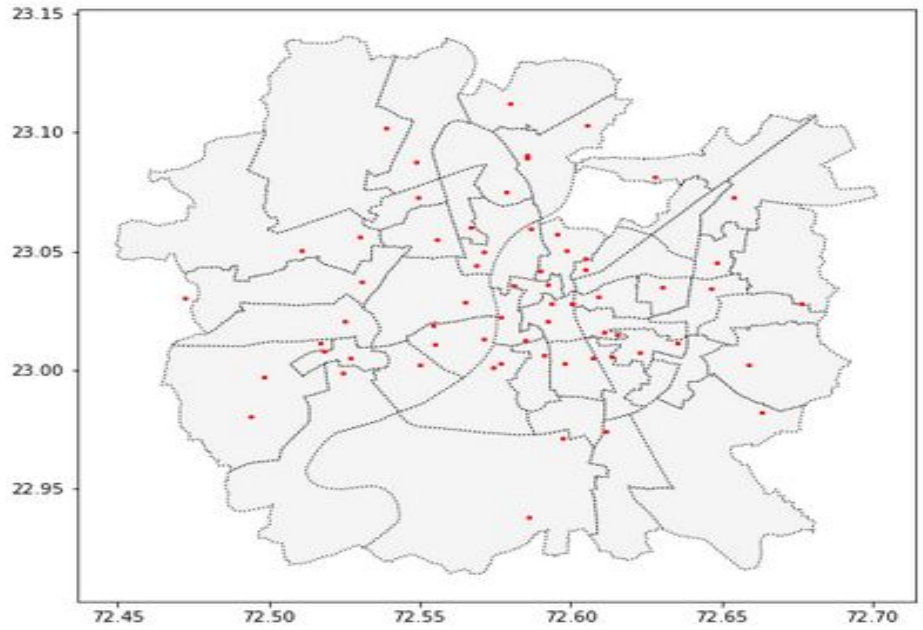


Figure 11 - Shows the location of neighbourhood inside the Ahmedabad region plot (plotted using Geopandas)

Benchmark

We use K-means clustering to use the features gathered and formulated from the data so that we can now label our data and form clusters. We then sort our score column to find out how our labels fir into the data. We observe that label 3 shows that data points depict places with the best quality of medical service, the label 1,2 depict a medium quality of service and the label 0 depict the worst quality of service in the neighbourhood.

Methodology

Data Pre-processing

The raw dataset contains names of the hospital, dentists, doctors, pharmacies, physiotherapist names and their individual rating.

This data is processed to fill the NaN values with the mean rating so that we can then find the total number of hospitals, dentists, doctors etc in a radius of 1 km around each neighbourhood point. After that, the columns were normalised so that a scoring function can be made to formulate a medical score which depicts the overall medical facility score.

Implementation

The implementation process can be divided into 3 stages:

1. Data Collection and Exploration
2. Data Visualisation
3. Unsupervised Learning through K-means

The Places API used to collect medical information around a point was quite troublesome. The API had a daily quota limit to the number of queries which can be made for data collection.

After Data Collection the scoring function was formulated with a particular formula

Medical Score= w1*nhos + w2*mhos +w3*ndoc+ w4*mdoc +w5*ndent+ w6*mdent+ w7*nph+w8*mph

Where,

nhos= number of hospitals

mhos= mean rating of hospitals

ndoc= number of doctors

mdoc=mean rating of doctors

ndent= number of dentists

mdent= mean rating of dentists

nph= number of physiotherapists

mph= mean rating of physiotherapist

Data Visualisation part has already been explained in detailed before.

Unsupervised Learning through K-means is also a fairly simple process and has been discussed before.

Refinement

Since this project is based on **Unsupervised Learning,** K-means is only used for clustering the neighbourhoods based on the features and that seems to be perfectly alright.

Result

Let us see the result of the clustering algorithm in the image below

Hospital_Count_zscore	Mean_hospital_rating_	Doctor_Count_zsc	Mean_doctor_rating_	Dentist_Count_zsc	Mean_dentist_	Pharmacy_	Mean_pha	Physiother	Mean_phy	Medical_score	labels
0.793972658	0.449003978	1.055490647	0.438442109	2.413634164	0.511188052	1.169094	0.187826	2.182731	0.88051	12.42033703	0
0.793972658	0.429042217	1.055490647	0.276254724	2.413634164	0.4974212	1.169094	0.215584	2.736067	0.831893	12.40565733	0
0.793972658	0.367520722	1.055490647	0.227936351	2.413634164	0.413956604	1.169094	0.209559	3.014085	0.858481	12.27813447	0
0.793972658	0.530063308	1.055490647	0.369417969	2.413634164	0.495721653	1.169094	0.295241	1.351377	0.93399	12.24900749	0
0.793972658	0.537254132	1.055490647	0.248678889	2.180335397	0.324341464	1.169094	0.419058	1.905613	0.765043	12.0137351	0
0.793972658	0.388816624	1.055490647	0.295656343	2.413634164	0.377967282	1.169094	0.126031	1.351377	0.868356	11.43507694	0
0.793972658	0.326744847	1.055490647	0.338371181	2.063686013	0.461702137	0.926531	0.296684	1.351377	0.93399	10.92887193	0
0.793972658	0.431875443	1.055490647	0.430597655	1.538763786	0.473998863	0.878018	0.204644	1.628495	0.945334	10.6768291	0
0.793972658	0.339385892	1.055490647	0.386026131	1.597088478	0.538995848	1.169094	0.270326	0.520024	0.848909	10.60243369	0
0.793972658	0.475391181	1.055490647	0.113519649	0.780542791	0.59269992	1.169094	0.375053	1.905613	0.857417	10.55910922	0
0.793972658	0.256205266	1.055490647	0.328609926	1.947038629	0.473246003	1.169094	0.204809	0.242906	0.916974	10.4483605	0
0.793972658	0.295687385	1.055490647	0.184738675	1.07216625	0.381743125	1.169094	0.466376	1.351377	0.853163	10.24387247	0
0.793972658	0.292880672	1.055490647	0.329965147	0.547244023	0.237182308	1.169094	0.407438	1.351377	0.821258	9.674280328	0
0.793972658	0.114488766	1.055490647	0.342973427	1.305465018	0.311603995	1.169094	-0.01472	1.351377	0.427758	9.24129471	0
0.793972658	0.340124431	1.055490647	0.371246844	0.430594639	0.384778251	1.169094	0.292403	-0.03421	0.959514	9.099982585	0
0.793972658	0.621296889	1.055490647	0.354218536	0.955516866	0.435205618	0.150329	0.348135	1.628495	0.902794	9.047015079	0
0.793972658	0.524764806	1.055490647	0.163978473	0.197295871	0.387921776	1.169094	0.310889	0.797142	0.508585	9.033414053	0
0.793972658	0.357907988	1.055490647	0.210040171	0.663893407	0.422051466	0.926531	0.425707	0.242906	0.885068	8.96186525	0
0.793972658	0.213509141	1.055490647	0.27293246	0.13897118	0.340993451	1.169094	0.476029	0.520024	0.831893	8.820712964	0
0.793972658	0.319282277	1.055490647	0.169646554	-0.094327588	0.04747811	1.169094	0.376	1.351377	0.59792	8.440670061	0
0.793972658	0.365929648	1.055490647	0.452211631	-0.210976972	0.420500117	1.169094	-0.03164	-0.31133	0.959514	8.042573103	0
0.793972658	0.292943793	1.055490647	0.281727742	0.372269947	0.364030992	0.635455	0.36265	-0.03421	0.938244	7.814409801	0
0.793972658	0.527732948	1.055490647	0.375905532	0.897192174	0.392345995	0.150329	0.314586	-0.03421	0.378127	7.516057382	0
0.793972658	0.29830904	-0.059464262	0.301562866	0.430594639	0.488998168	1.169094	0.42603	-0.03421	0.895704	7.496194664	0
0.793972658	0.318871931	-0.148660654	0.28205212	0.488919331	0.447648734	0.683968	0.034834	1.905613	0.959514	6.851959937	0
0.793972658	0.418852719	1.055490647	0.358376391	0.488919331	0.570089	-0.14075	0.266985	-0.31133	0.895704	6.637543674	0
0.793972658	0.209669517	0.698705076	0.420304326	0.022321796	0.268467858	0.295867	0.455079	0.520024	0.810622	6.51590544	0
0.793972658	0.372464413	0.341919505	0.448998743	-0.560925124	0.193809159	0.926531	0.341787	-0.31133	0.959514	6.478835154	0

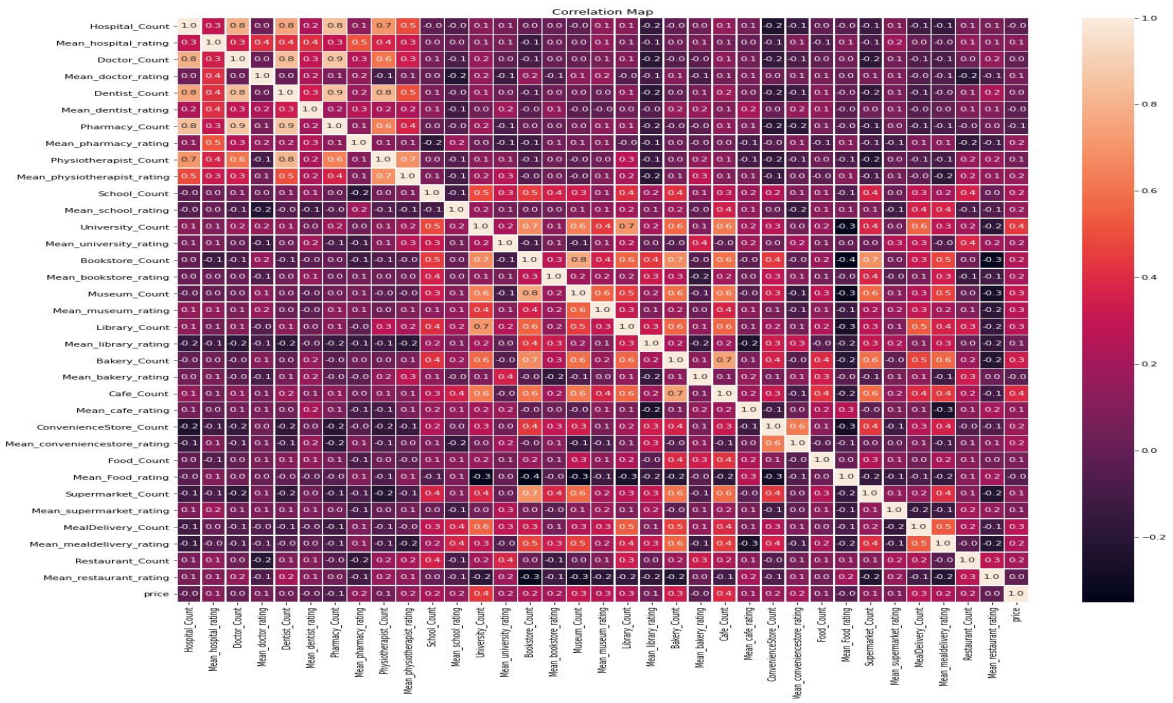
Figure 8 – Label 0 is given to the ones whose medical score is quite good (arranged in descending order).

-1.864309983	0.828931934	-1.219017366	0.511971355	-0.969197967	0.328194817	-1.54761	0.902436	-0.31133	0.895704	-5.929940908	3
-1.071488844	0.009927032	-0.951428188	0.091199811	-0.852548583	0.552170911	-0.81992	0.123918	-0.86557	-1.16751	-7.011973683	2
-1.72440037	0.828931934	-1.575802937	0.673900928	-1.085847351	-2.359518335	-1.25654	0.563529	-0.58845	0.959514	-8.693723584	3
-1.211398457	0.602420975	-1.575802937	-2.024925297	-0.969197967	-0.567709567	-1.59612	0.699092	-0.86557	-1.16751	-12.21368244	2
-2.004219595	-2.631652161	-1.620401133	-2.699631854	-1.085847351	-2.359518335	-1.69315	0.902436	-0.86557	-1.16751	-23.54033089	1
-1.957583058	-2.631652161	-1.486606544	0.673900928	-1.085847351	-2.359518335	-1.69315	-3.16445	-0.86557	-1.16751	-24.26307665	1
-2.004219595	-2.631652161	-1.575802937	-2.699631854	-1.085847351	-2.359518335	-1.64464	-0.31763	-0.86557	-1.16751	-25.20650681	1
-1.864309983	-2.631652161	-1.620401133	-2.699631854	-1.085847351	-2.359518335	-1.64464	-0.52097	-0.86557	-1.16751	-25.22864648	1
-1.864309983	-2.631652161	-1.620401133	-2.699631854	-1.085847351	-2.359518335	-1.74166	-3.16445	-0.86557	-1.16751	-29.38791025	1
-1.957583058	-2.631652161	-1.620401133	-2.699631854	-1.085847351	-2.359518335	-1.74166	-3.16445	-0.86557	-1.16751	-29.62109294	1
-2.004219595	-2.631652161	-1.620401133	-2.699631854	-1.085847351	-2.359518335	-1.74166	-3.16445	-0.86557	-1.16751	-29.73768428	1
-2.004219595	-2.631652161	-1.620401133	-2.699631854	-1.085847351	-2.359518335	-1.74166	-3.16445	-0.86557	-1.16751	-29.73768428	1
-2.004219595	-2.631652161	-1.620401133	-2.699631854	-1.085847351	-2.359518335	-1.74166	-3.16445	-0.86557	-1.16751	-29.73768428	1
-2.004219595	-2.631652161	-1.620401133	-2.699631854	-1.085847351	-2.359518335	-1.74166	-3.16445	-0.86557	-1.16751	-29.73768428	1

Figure 9 – Label 1 for those whose medical score is the least. As you can see it goes into negative.

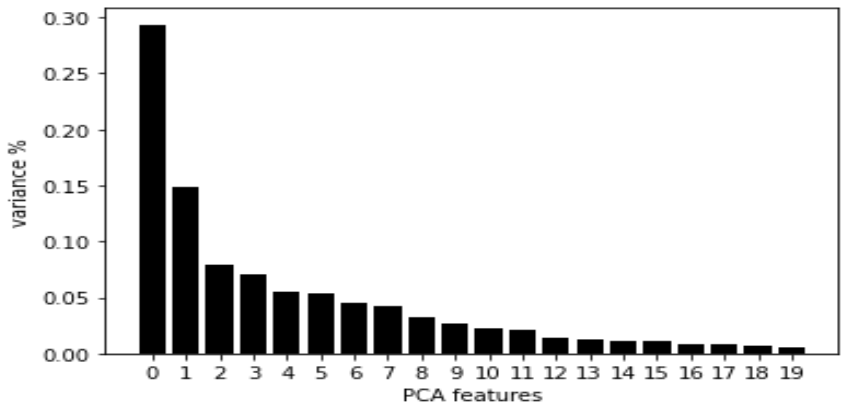
EDA : Key points notable among data:

1. Some businesses have shown high correlation among themselve. Which means that they are interdependent. For example Doctors count and Pharmacy count have a correlation of 0.9.
2. University Count and Cafe count shows the highest correlation with price which means that people of posh area have demand for these services more that other services.
3. However one point to be noted here is that while doctors count and pharmacy count is highly correlated , doctors rating and pharmacy rating have very little correlation.

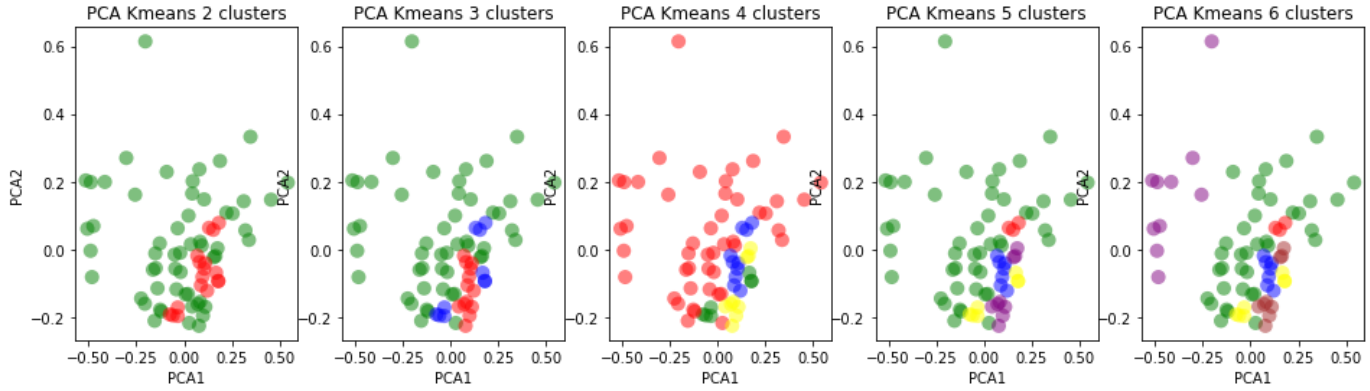


Wrangling through different model :

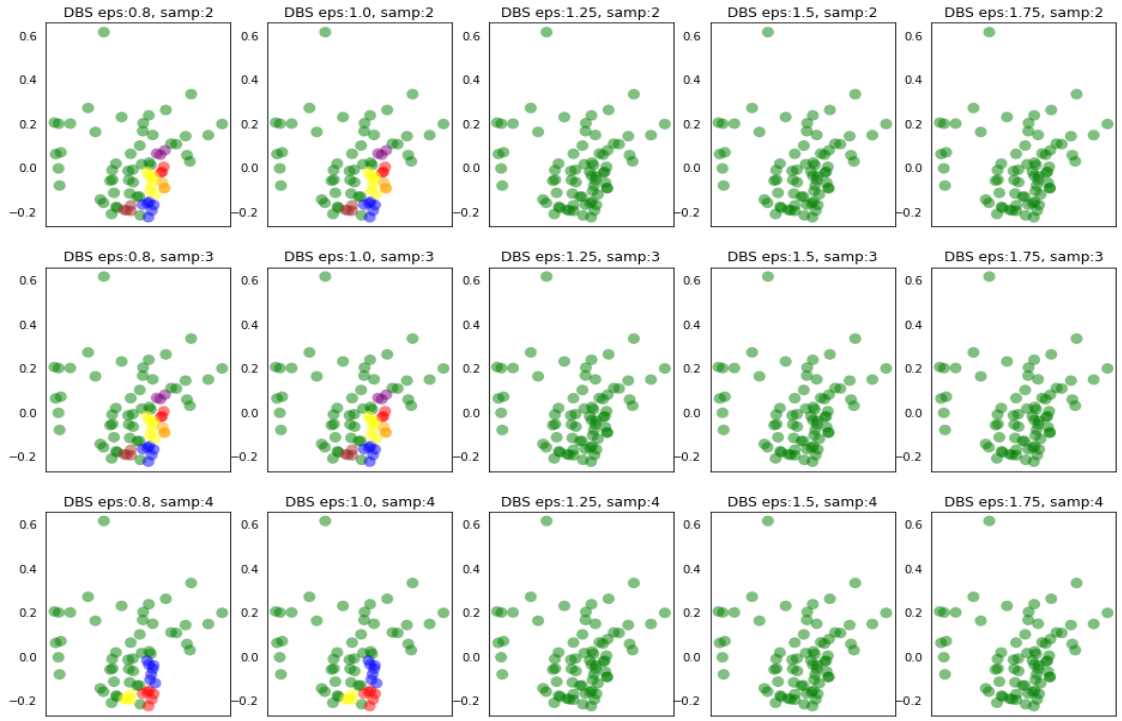
First we did PCA for getting an overview of data.



In the above figure we can see that 2 components explain majority of variance. Then we fit first two components of principal components in K-means for different no clusters for visualization.



Also we fit the first two components of principal components in DB-scan algorithm for different epsilons and no of samples in the cluster for better Understanding.



From the above figure we can say that eps = 0.8 and 4 samples in the cluster would be a choice for optimal result.

Then we have computed our scoring function based on the spearman correlation which is described in the next section. But we find out that our scoring function is more consistent with K-means rather than DB-scan. So we choose K-means finally.

Final Model Description :

- We chose K-means clustering to give a visual representation of locality around a particular place in which the user wants to set up a particular business, bcz neighbourhood is always necessary while choosing a business. We have also done hypothesis testing for finding which correlations are significant.

- We chose a scoring matrix, whose pseudo code can be given by :

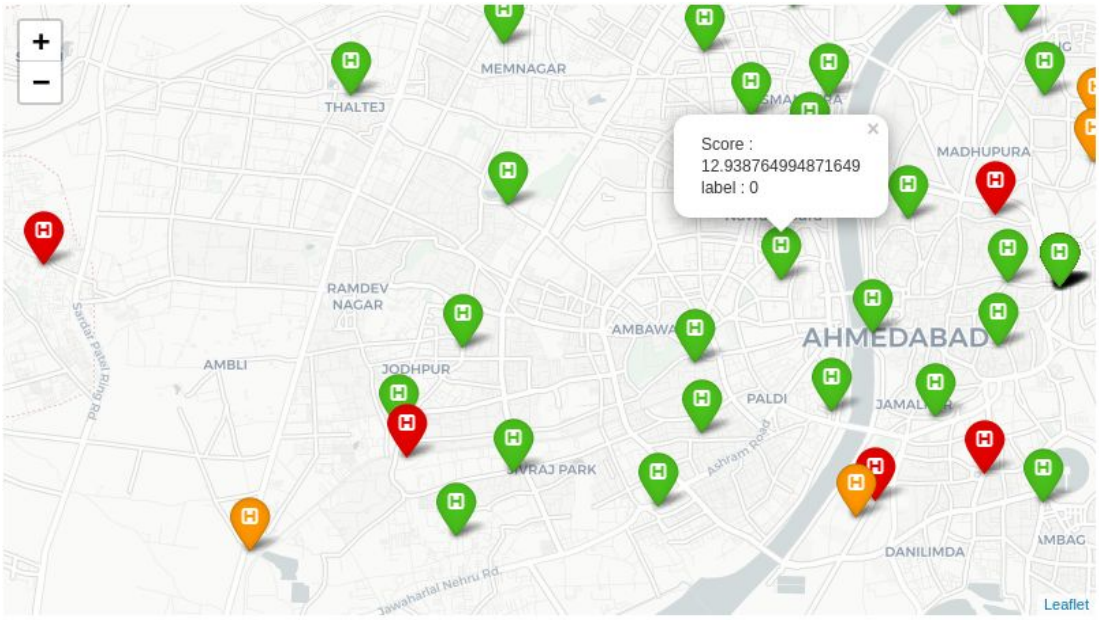
- df_final2['Score'] = 0
- for column in df_final2.columns:
- df_final2['Score'] = +df_final.columns x correlation[column]['price']
- df_final2['Score'] += normalizedPrice

Model Explanation :

We chose this scoring matrix after seeing a spearman correlation between different columns and price , as we can think property price as a good indicator of development index in that area, and added normalized price to the score bcz logistic is a key attribute while setting up a business.

We chose k-means for visual analysis of business location and for quantitative analysis we chose our scoring function with k = 4. But then we merged k = 1 and k = 2 for a more consistent result.

Final Result : As it is a business problem so qualitative and quantitative analysis both are important. So we have given the result accordingly.



Model Evaluation and Validation

Now the task will be to give us the coordinates of a point in Ahmedabad, and we will see which neighbourhoods are the closest to this point. The properties of the closest neighbourhoods will match with the properties of the nearby areas around the given coordinate. Using the labels of our clustering process we can comment on the quality of service existing around this particular point. This will in turn help a property developer check if he should invest in building up a hospital nearby.

Justification

We can conclude that if it results in the label 3 then the area already has the best quality of service in a Medical Facility. In other words, business supply is saturated.

If the result is a label 0 then the area has the worst quality of Medical service. In other words, business supply is in a bad condition.

If the result is a label 1 or 2 then the area has a moderate quality of Medical service. In other words, business supply is just at a threshold condition.

This is justifiable as we can cross check that labels are fit perfectly based on the medical score of sorted.

Conclusion

There are two aspects to building up a business in a particular place.

1. Business Supply
2. Business Demand

In this project we have gone through the process of exploring, visualising and modelling on the basis of business supply.

By looking how much supply of medical facilities are available in the neighbourhood can get us an idea of business supply.

Future Scope:

1. Adding population in scoring matrix calculation.
2. Making this code modular so that the open source community can use it for their visual and quantitative analysis of business. They may also use their own set of latitude and longitude for comparison and they may also extend it up for different types of business.
3. Integrating this analysis in businesses like Justdial as a new feature as they have a huge reservoir of business demand data.

Improvement

An improvement to this project or in other words an extension would be to get enough data for business demand which can inform us how many users are trying to look for a hospital in this place. If the number of searches exceeds a particular threshold, we can say that this neighbourhood is in need of a medical facility and thus we can inform this to the property dealer.

Teamwork

Meetings were held on every alternate day for smooth workflow. Although each team member played role in each stage of the project, yet major contributions can be delivered as follows:

Data Collection, Exploration and Cleaning- majorly done by Himani Madaan and Utkarsh Gupta.

Data Visualization- majorly done by Tanish Gupta and Divyansh Khandelwal.

Modelling and Exploratory Data Analysis - Shalini Kumari and Utkarsh Gupta.