DS 250: Data Analysis and Visualization

Business Recommender

Team

Utkarsh Gupta
Himani Madaan
Divyansh Khandelwal
Tanish Gupta
Shalini Kumari

Motivation

Business Setup

Aspects for building up a business:

1. Business Supply
2. Business Demand

Business Problem

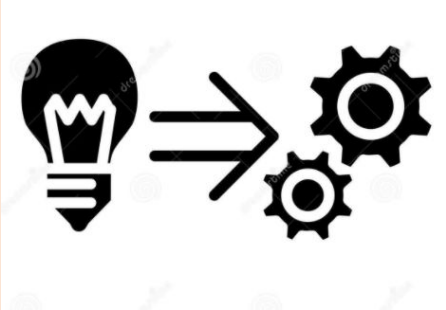
In the city of Ahmedabad, if a property developer is looking to open a new hospital or school, where would you recommend that they open it?

We measure the quality of recommendation in terms of average service rating and it’s quantity.

Implementation

The project is divided in 3 stages :

1. Data Collection
2. Data Analysis and Visualization
3. Modelling[Unsupervised Learning]



End Goal-

On giving the coordinate of a point in Ahmedabad, our model can tell the quality of service of that business and if the business supply is sufficient enough.

Data Collection

- Finding neighbourhoods in Ahmedabad using Web Scraping

  Wikipedia link

- GeoCoder

Data Collection

- Finding neighbourhoods in Ahmedabad using Web Scraping

  Wikipedia link

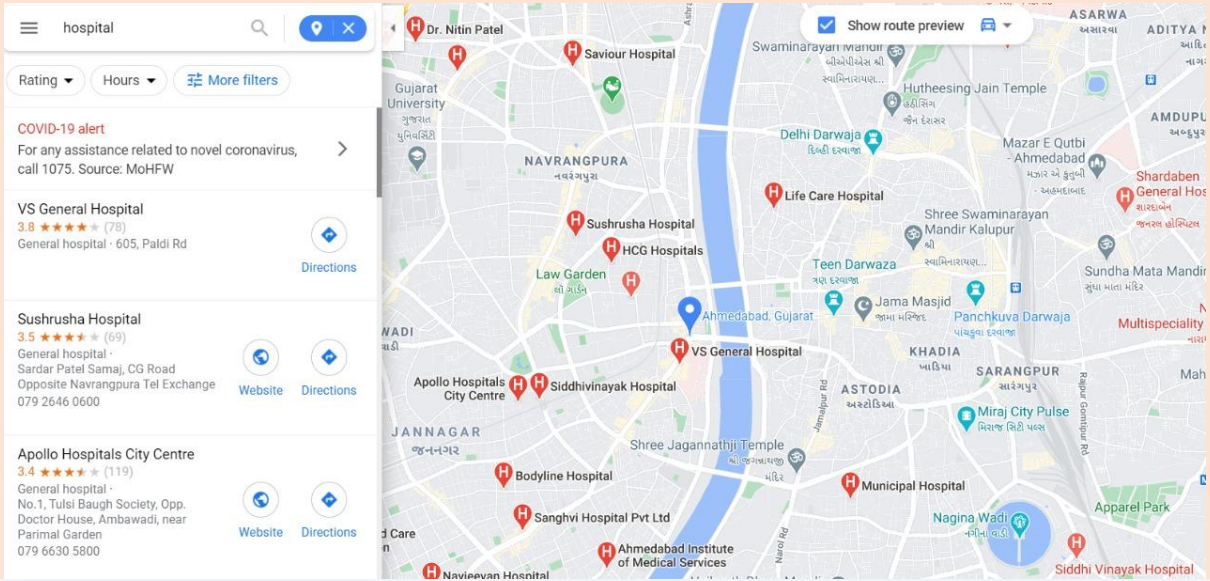
- GeoCoder

(81, 3)

	Neighborhood	Latitude	Longitude
0	Agol	23.027760	72.600270
1	Ahmedabad Cantonment	23.027760	72.600270
2	Alam Roza	23.002120	72.549790
3	Ambawadi	23.018850	72.554410
4	Amraiwadi	23.007350	72.622680
5	Anand Nagar (Ahmedabad)	23.011390	72.517120
6	Asarwa	23.047080	72.604810
7	Asarwa Chakla	23.042257	72.604566
8	Badarkha	22.841280	72.454530
9	Bahiyal	23.027760	72.600270
10	Bapunagar	23.034760	72.630240

Data Collection

Places API (Documentation link)



Data Collection

Places API (Documentation link)

Data is collected over 3 major categories- Healthcare, Education and Food

Features for Health care sector 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

Data Collection

	Neighbourhood	Latitude	Longitude	Venue_type	Venue_Name	Venue_Rating
0	Agol	23.027760000000058	72.60027000000008	hospital	Victoria Jubilee Hospital	4.0
1	Agol	23.027760000000058	72.60027000000008	hospital	Al Ameen Hospital|	NaN
2	Agol	23.027760000000058	72.60027000000008	hospital	Dr.Tanumati Shah Hospital	NaN
3	Agol	23.027760000000058	72.60027000000008	hospital	Lokhandwala General Hospital	4.2
4	Agol	23.027760000000058	72.60027000000008	hospital	shreeShreeji Pathology Laboratory	NaN
...	...	...	...	...	...	...
11088	Vejalpur	23.007820000000038	72.51818000000003	physiotherapist	Dr Binal Shah Desai/Happy Healing Physio Clinic	5.0
11089	Vejalpur	23.007820000000038	72.51818000000003	physiotherapist	Dr.Nirali`s PhysioRehab	NaN
11090	Virochannagar	23.093770000000063	72.22700000000003	hospital	SC Virochannagar	NaN
11091	Virochannagar	23.093770000000063	72.22700000000003	hospital	phc virochannagar	NaN
11092	Virochannagar	23.093770000000063	72.22700000000003	hospital	PRIMARY HEALTH CENTRE VIROCHANNAHAR	NaN

Data Processing

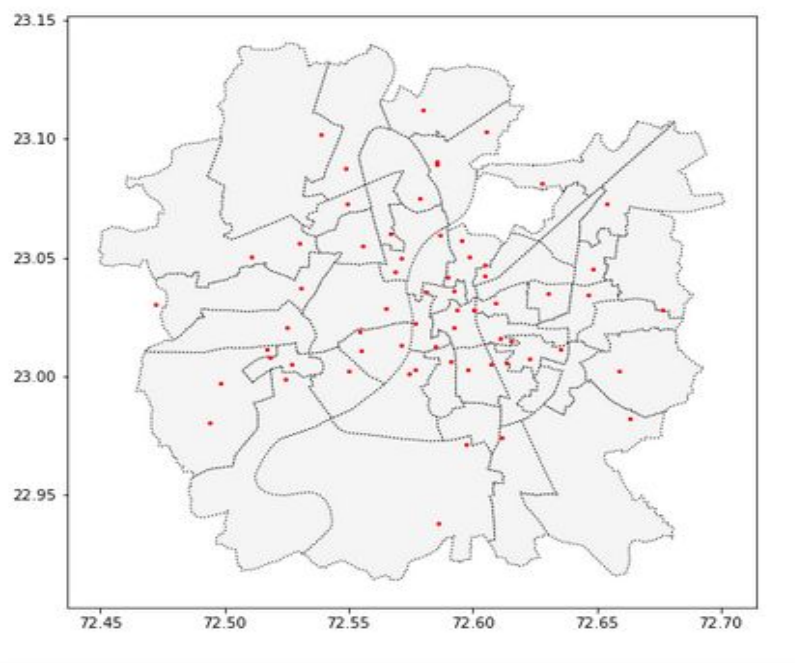
We groupby the dataframe according to neighbourhood name and venue-type to calculate the number of hospitals, doctors etc and their mean rating.

	Neighbourhood	Latitude	Longitude	Hospital_Count	Mean_hospital_rating	Doctor_Count	Mean_doctor_rating	Dentist_Count	Mean_dentist_rating	Pharmacy_Count	M
0	Agol	23.02776	72.60027	33	4.475000	45	4.712500	12	4.812500	39	
1	Ahmedabad Cantonment	23.02776	72.60027	33	4.475000	45	4.712500	12	4.812500	39	
2	Alam Roza	23.00212	72.54979	60	4.489189	60	4.169444	32	4.942857	60	
3	Ambawadi	23.01885	72.55441	60	4.578571	60	4.369767	56	4.493548	60	
4	Amraiwadi	23.00735	72.62268	47	4.144444	27	4.607692	16	4.655556	40	
...	...	...	...	...	...	...	...	...	...	...	
76	Usmanpura	23.04981	72.57120	60	4.407500	60	4.532353	27	4.905000	33	
77	Vastral	23.00238	72.65865	60	4.153488	17	4.516667	26	4.633333	35	
78	Vastrapur	23.03717	72.53085	60	4.426316	60	4.639394	45	4.744118	54	
79	Vejalpur	23.00782	72.51818	60	4.274419	60	4.502703	54	4.723529	55	
80	Virochannagar	23.09377	72.22700	3	0.000000	0	0.000000	0	0.000000	0	

- Calculating z-score for every column

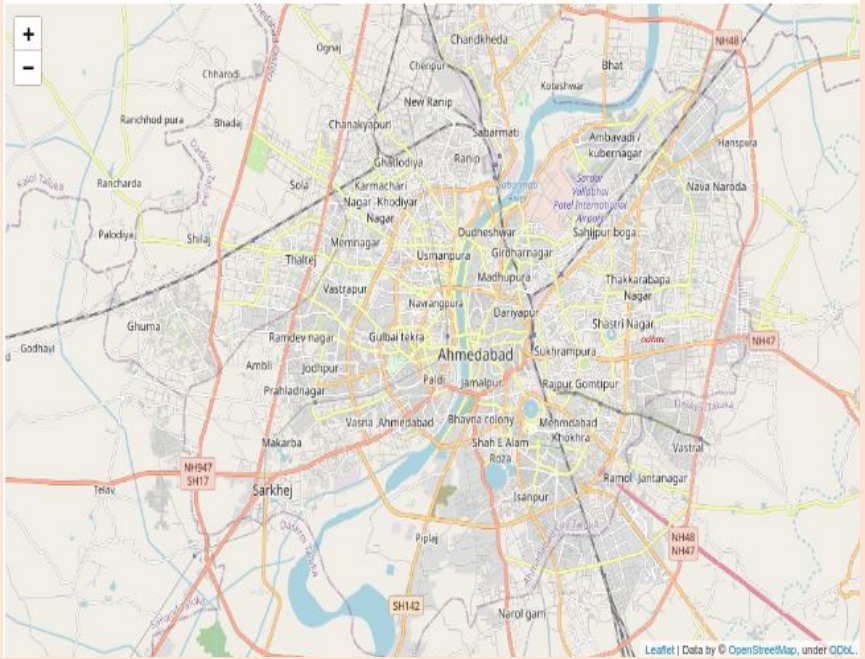
Data Visualization

We have used **Geopandas** for showing the location of neighbourhood which is merged within the Ahmedabad region plot.



Data Visualization

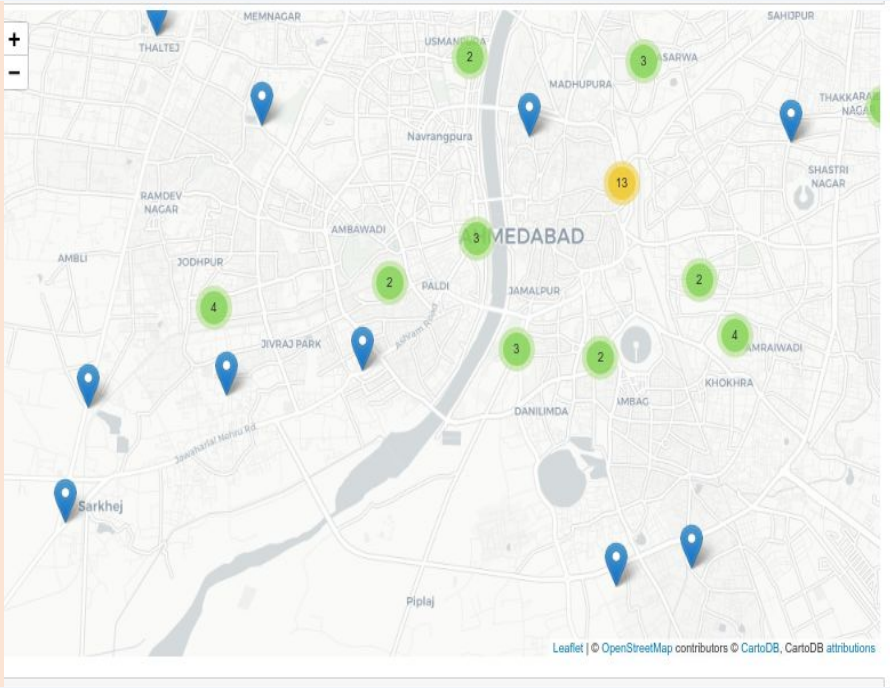
Then, we have done some visualization using **Folium.**



Data Visualization

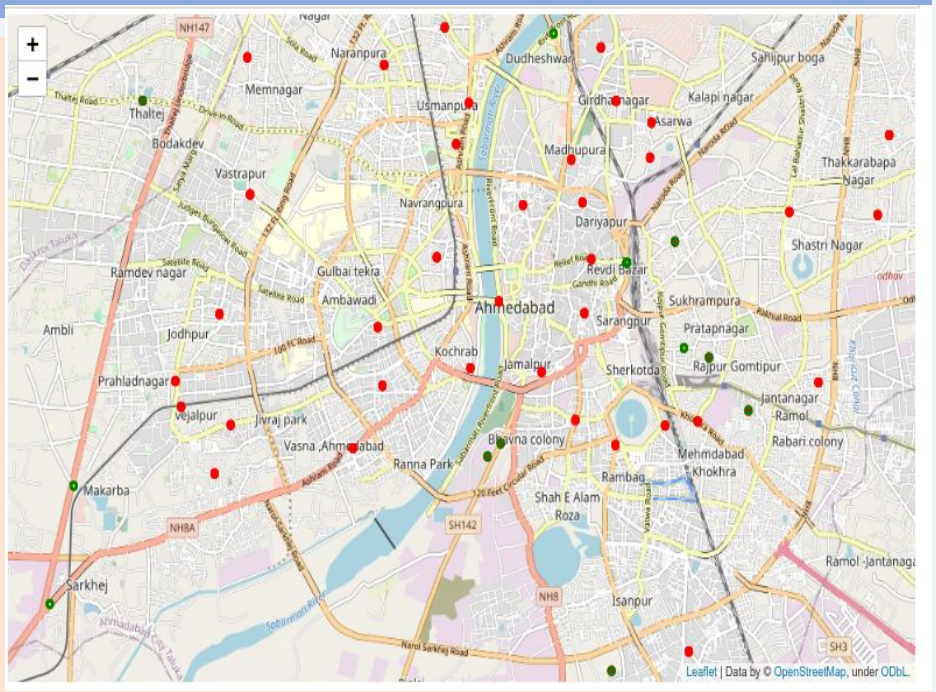
This folium plot is showing some numbers,which is basically the number of neighbourhoods at that particular position.

So if we zoom in/out that number will vary according to the number of neighbourhood present.



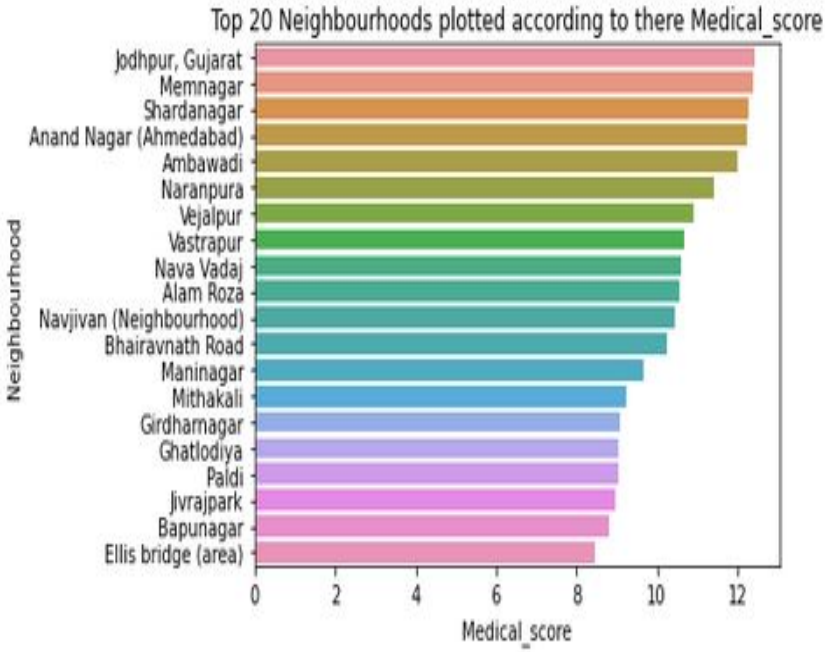
Data Visualization

Further, we have shown the location of neighbourhood having hospital count greater than 50 with red colour and less than 50 with green. (Same is done for pharmacy and doctor count part)



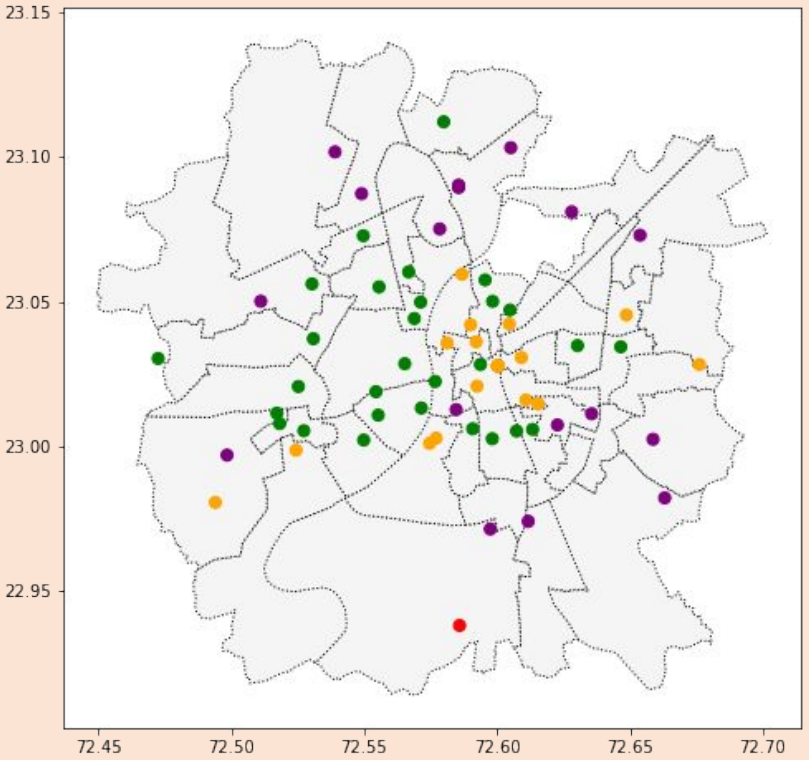
Data Visualization

Here is the **Bar Plot** showing the Top 20 neighbourhood according To our medical score.



Data Visualization

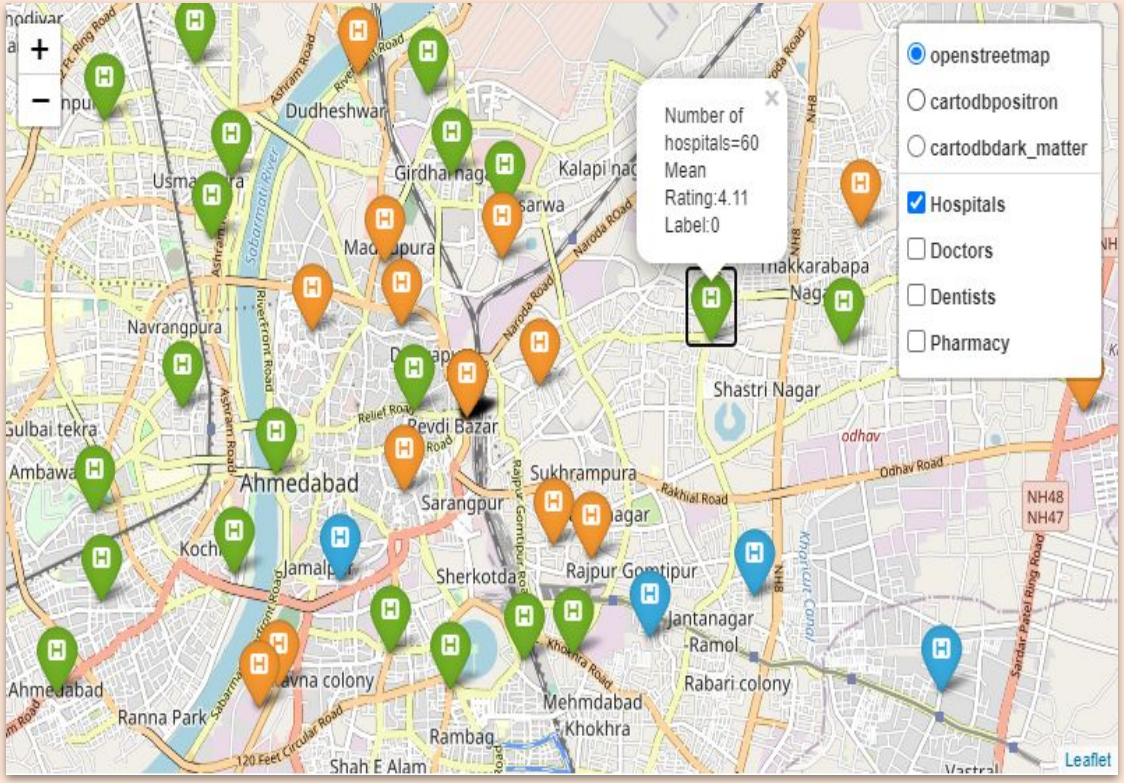
- Merged graph of regions in our dataset with wards of Ahmedabad to find out which of the regions lie outside Ahmedabad.
- Longitudes on x-axis and Latitudes on y-axis.
- Labelling is done using k-means.



Data Visualization

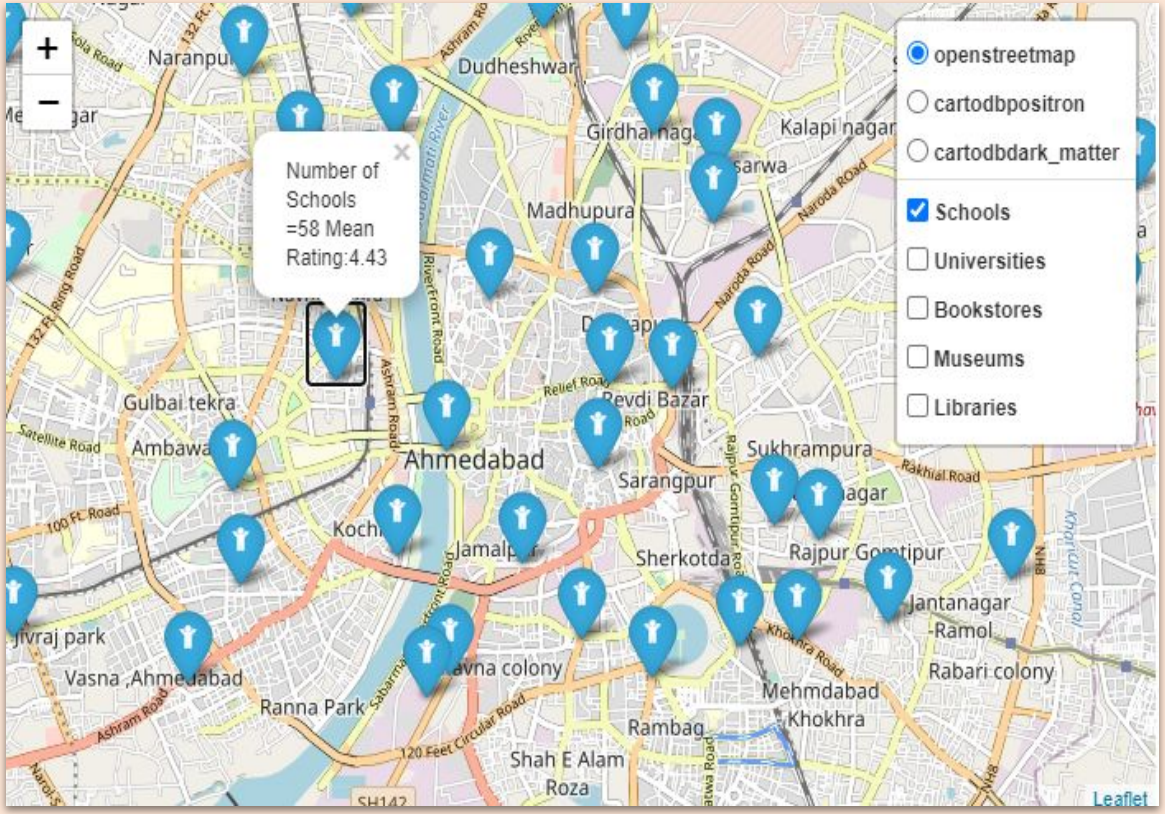
Using geopandas and folium,plotted this all-in-one graph showing the summary of the medical dataset like number of hospitals and their mean rating in a region.

Coloring of icons done using k-means.



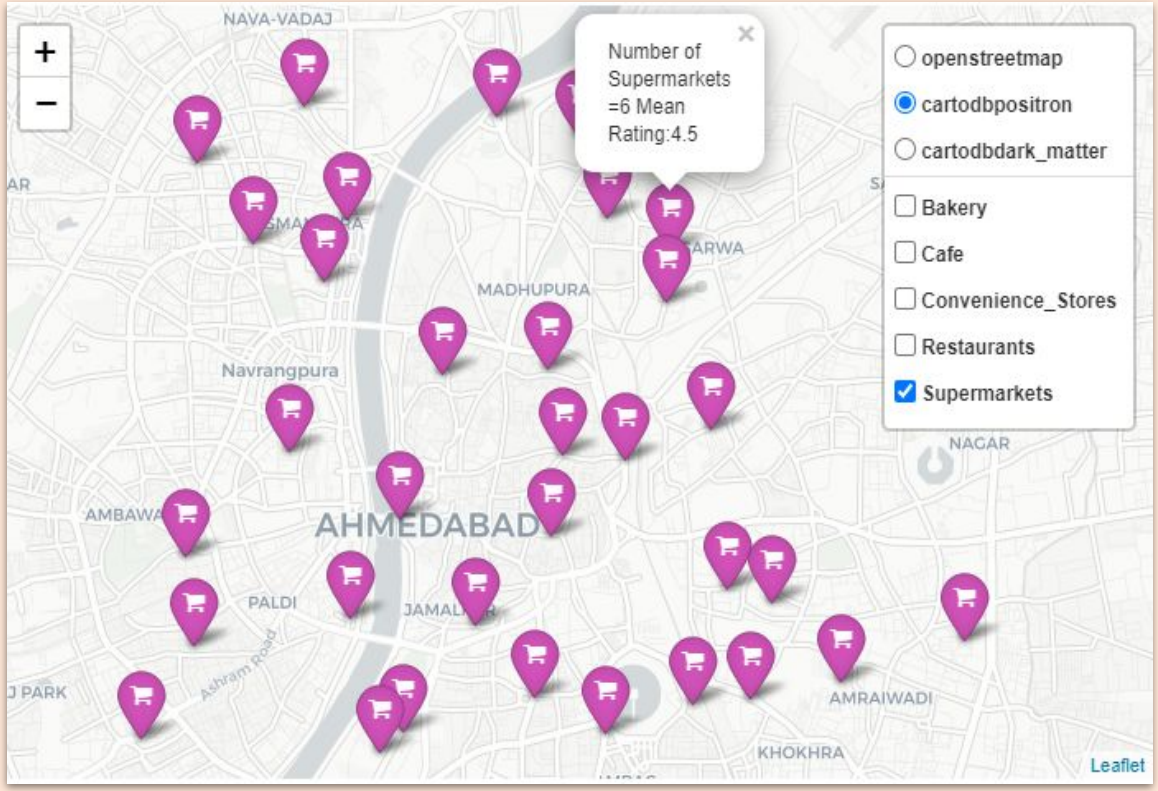
Data Visualization

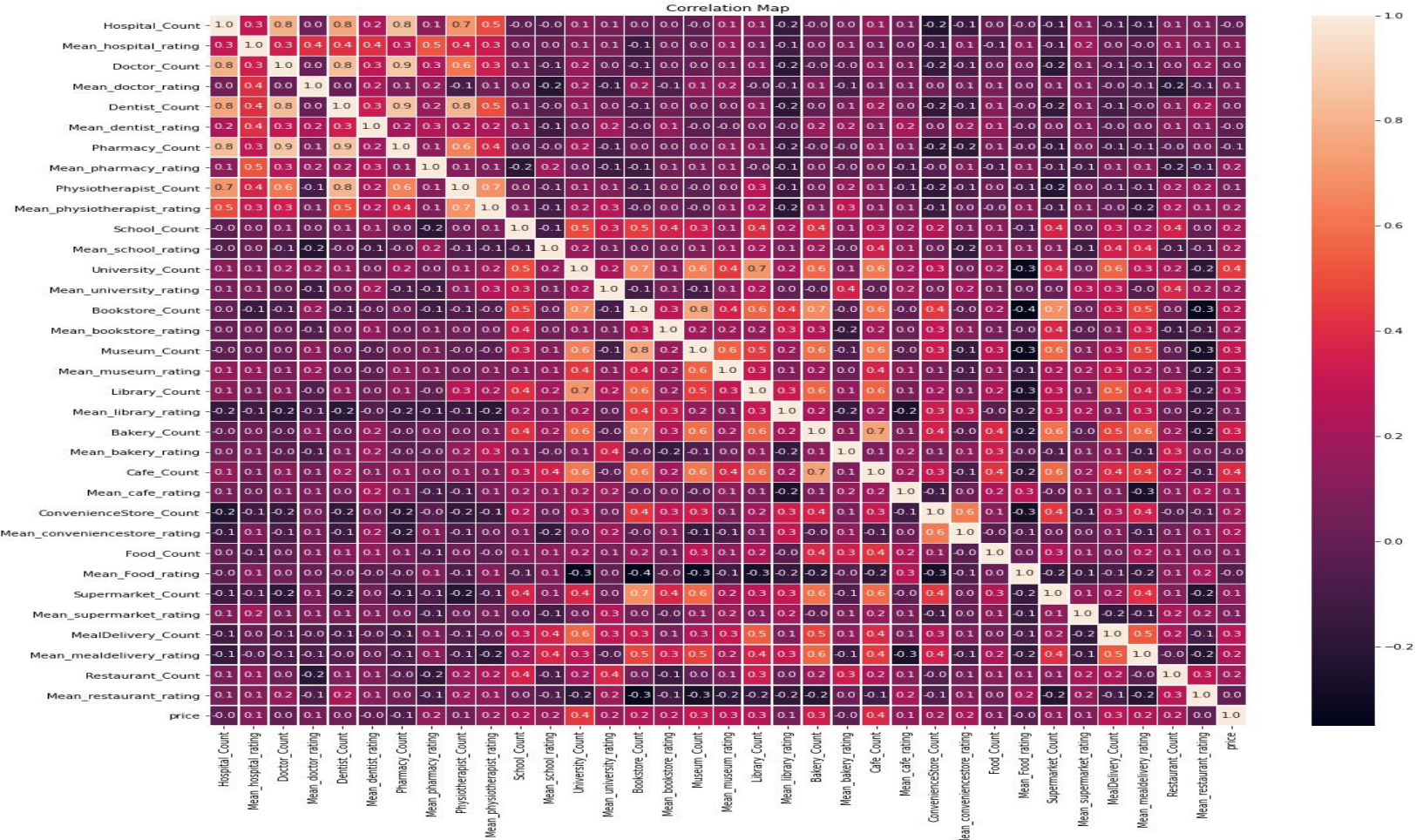
Similarly for school dataset also,we have plotted same kind of graphs where one can the see the entire details of the dataset in just one graph.



Data Visualization

One can also toggle between different themes of graphs like openstreetmap, cartodb positron,etc to get a better view.




Correlation Map
Hospital_Count
Mean_hospital_rating
Doctor_Count
Mean_doctor_rating
Dentist_Count
Mean_dentist_rating
Pharmacy_Count
Mean_pharmacy_rating
Physiotherapist_Count
Mean_physiotherapist_rating
School_Count
Mean_school_rating
University_Count
Mean_university_rating
Bookstore_Count
Mean_bookstore_rating
Museum_Count
Mean_museum_rating
Library_Count
Mean_library_rating
Bakery_Count
Mean_bakery_rating
Cafe_Count
Mean_cafe_rating
ConvenienceStore_Count
Mean_conveniencestore_rating
Food_Count
Mean_Food_rating
Supermarket_Count
Mean_supermarket_rating
MealDelivery_Count
Mean_mealdelivery_rating
Restaurant_Count
Mean_restaurant_rating
price
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

EDA : Key Points Notable among data

1. Some Business 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 highest correlation with price which means that people of posh area have demand of 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.

Model Explanation :

- We chose K-means clustering to give a visual representation of locality around a particular place in which user wants to setup a particular business, bcz neighbourhood is always necessary while choosing a business.

- 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 spearman correlation between different column 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.

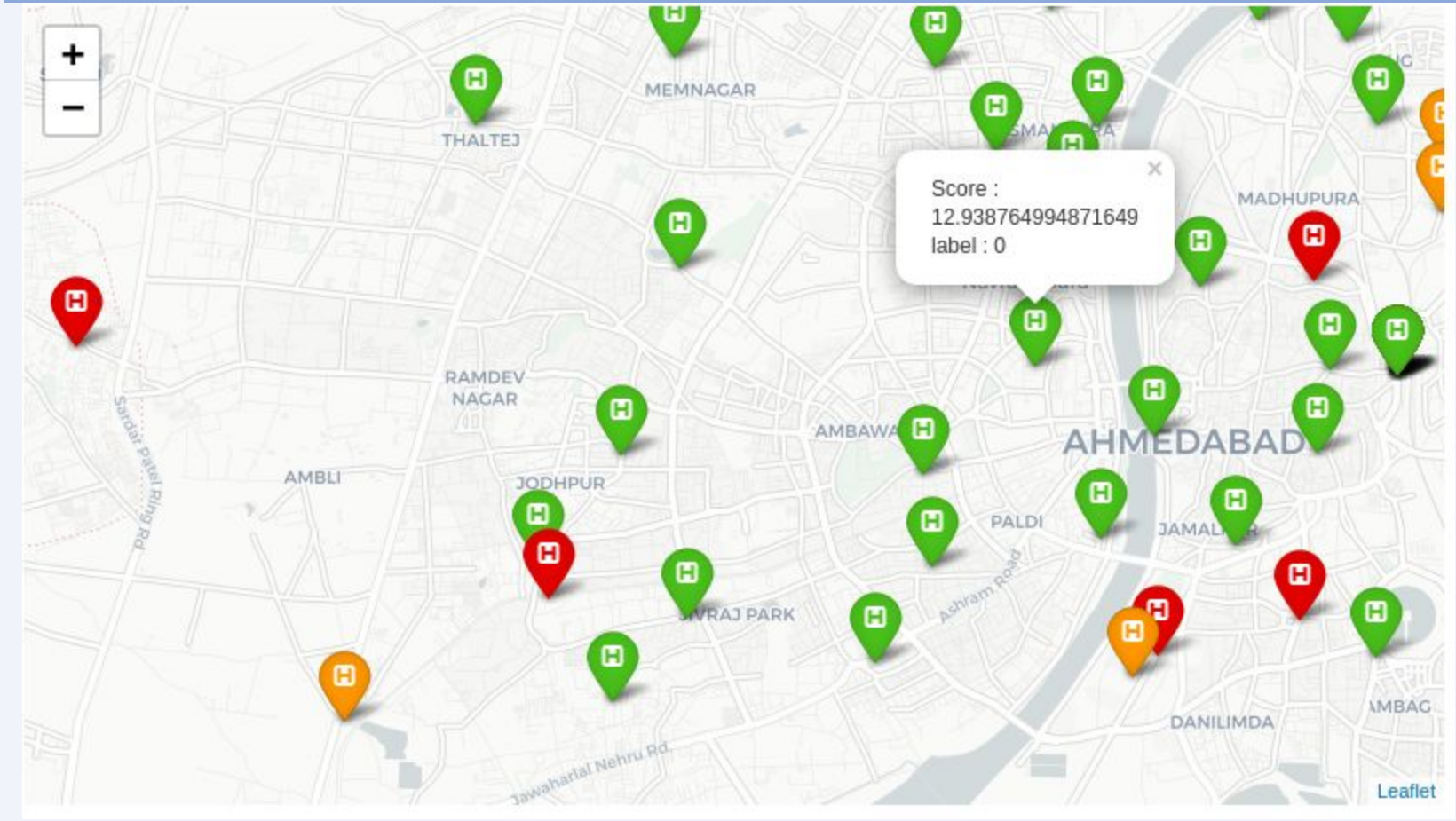
Difficulties Faced and comparison between K_means and our scoring Function .

1. It was difficult to come to a scoring function which generalize well with cluster labels.

2. The way we chose scoring function was not monotonically accordance with cluster labels, but it did pretty well.

3. Both the models are equally important while choosing the best location for the business. One was for qualitative analysis while another was for quantitative.

Final Result



Future Scope:

1. Adding population in scoring matrix calculation.

2. Making this code modular so that 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 type of business.

3. Integrating this analysis in businesses like Justdial as a new feature as they have huge reservoir of business demand data.

Conclusion

- Currently the project deals with the Business Supply part of the problem. An improvement or genuine extension can be inclusion of **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 medical facility and thus we can inform this to the property dealer.
- Through this project we contributed a **new dataset to the Data Science** realm that is the neighbourhood analysis of the city of Ahmedabad.
- Getting a clear understanding of the Places API, Folium for visualization and clustering & segmentation.