ANALYSIS OF SCHOOLS IN HYDERABAD

This report is part of my capstone project for the final course in Applied Data Science Specialization on Coursera.

INTRODUCTION

In this report I have analyzed the school distribution in the city of Hyderabad. This is significant in identifying the areas where the school infrastructure needs to be developed.
Inorder to obtain the distribution of schools in Hyderabad I have used **FourSquare API**
(Note: I have used FourSquare APIs search feature for this)

BACKGROUND
It is essential for everyone to have access to quality education at the school level. This is the foundation of education. There are plenty of good schools in Hyderabad, however they are clustered in certain neighbourhoods.
There are several neighbourhoods in Hyderabad where there is lack of infrastructure in terms of schools etc. It is important to identify these regions for the development of the city and to provide easy access to education for everyone.

PROBLEM
The government needs to develop school infrastructure either by constructing schools in these regions or support in the development of private schools in these areas to provide easily accessible quality education available to everyone. However in order to do this, it is essential to identify these neighbourhoods. So in this report I have analyzed the distribution of schools in Hyderabad and have identified these neighbourhoods.
Also it is important information for the people also when they have moved in the city and are looking to buy a house or rent a place to live in because they would definitely want to choose a neighbourhood which has easy access in terms of transportation and number of options of good schools for their children.

WHO WILL BE INTERESTED?
As I have explained earlier, the target audience will be the government agencies looking to build school infrastructure, the private sector agencies looking to build schools in areas with good potential and the people looking for houses in different neighbourhoods to identify which ones will be closer to good schools.

DATA:Acquisition and Cleaning

DATA SOURCE
I have used beautiful soup to scrape the various neighbourhoods in Hyderabad from
https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India.
Then I have obtained the Latitudes and Longitudes of these neighbourhoods using geocoder.
I have then used FourSquare API to obtain the location of various schools in each neighbourhood setting the limit to 100 for each neighbourhood and a radius of 1km.

For plotting the choropleth map, I have acquired the geojson file of Hyderabad from https://raw.githubusercontent.com/datameet/Municipal_Spatial_Data/master/Hyderabad/ghmc-wards.geojson

PREPROCESSING THE DATA

After obtaining the location of various schools in the city, there were some unwanted entries such as 'driving schools', 'schools of sound' etc. I have dropped these entries. Also there was duplication of data because it was possible that a school was within 1000 metres distance from each neighbourhood as a result of which it may appear twice with the same latitude and longitude in both these neighbourhoods. I have removed such duplicates and have kept the first copy of such data. Also there were several schools whose category was mismatched or not clear. I have thus categorized them based on the data to whether they are high schools, primary/play schools, student centers or their category was not mentioned.

FINAL FORM OF DATA USED
This is an image of the first 9 entries of the dataframe used.

	Neighbourhood	Latitude	Longitude	School	School_Latitude	School_Longitude	School_category
0	A. S. Rao Nagar	17.411200	78.508240	Mennonite Brethem High School, Zamistanpur	17.410229	78.501984	High School
1	A. S. Rao Nagar	17.411200	78.508240	Sri Aurobindo International School	17.400952	78.505043	Not Mentioned
2	A.C. Guards	17.392977	78.456867	Genesis high school	17.395249	78.457319	High School
3	A.C. Guards	17.392977	78.456867	Govt High School, Vijaya Nagar Colony	17.394348	78.454649	High School
4	A.C. Guards	17.392977	78.456867	St. Anns Girls High School For Girls	17.394340	78.453383	High School
5	A.C. Guards	17.392977	78.456867	hyderabad international residential school	17.385760	78.460598	Not Mentioned
6	Abhyudaya Nagar	17.337650	78.564140	School	17.345389	78.559057	Not Mentioned
7	Abids	17.389800	78.476580	Diamond Jubilee High School	17.389743	78.474340	High School
8	Abids	17.389800	78.476580	John's High School	17.393322	78.476040	High School

Each row corresponds to a particular school and its information, i.e. neighbourhood it belongs to, its category, name and it location.
There are a total of 235 entries in the dataframe.
Using this data I have analyzed the distribution of schools in Hyderabad.
For the choropleth map I had access to the geospatial data of only those neighbourhoods in Hyderabad covered by GHMC. So some neighbourhoods in Secunderabad have not been covered in this map. The form of data is the same as above.

MAPPING THE DATA

Using Folium, I have mapped the data onto a map to give a general idea of the distribution of schools in Hyderabad.
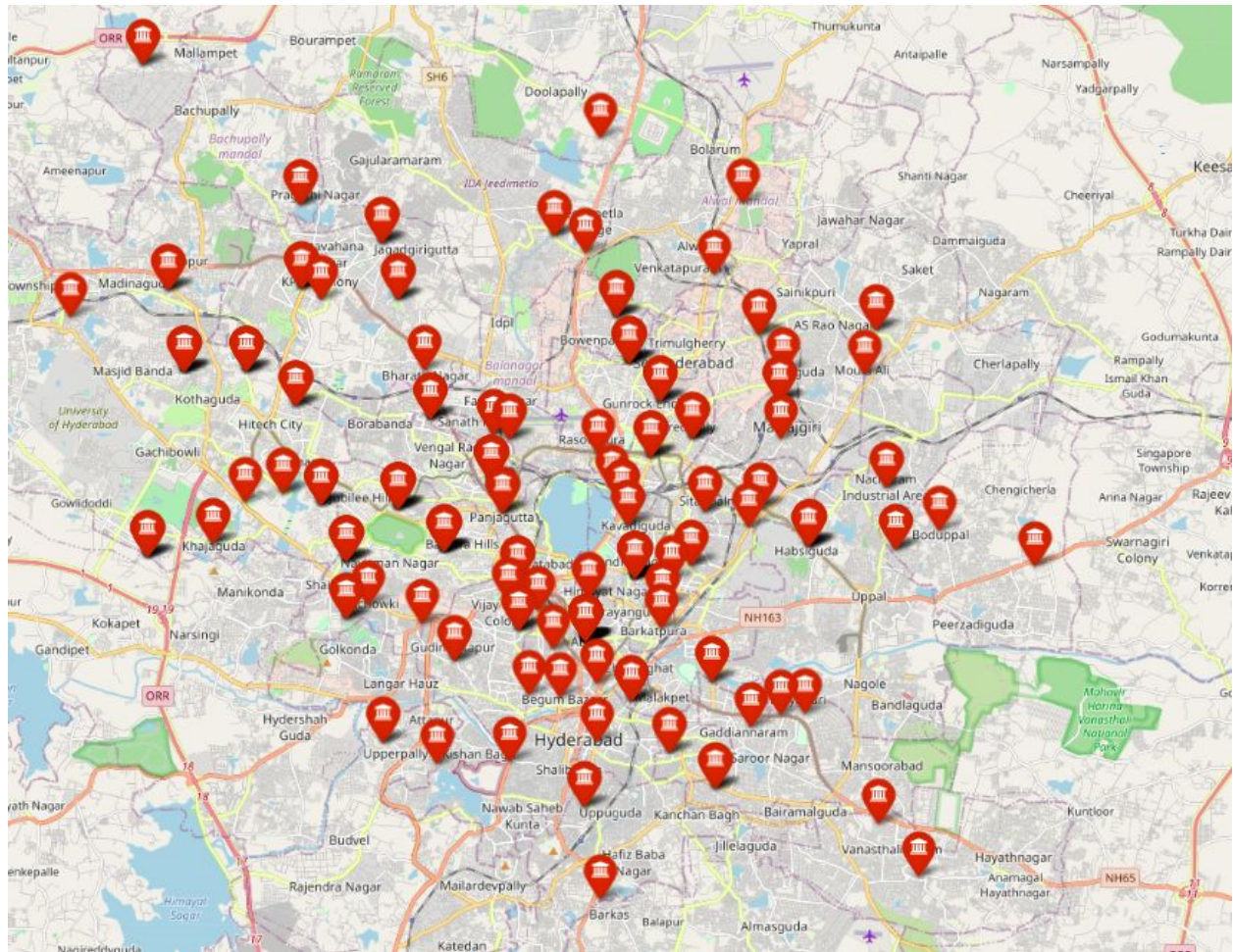
It can be observed that the schools are clustered more in the centre of the city which can be expected.

METHODOLOGY

First, the data has been retrieved and cleaned as explained previously. The data consists of schools within a radius of 1km for each neighbourhood with a limit set to 100.

Next, we analyze the data by exploring the various categories to which the schools belong.

After this we explore the distribution in each neighbourhood in a more explanatory way using choropleth maps for visualization. This gives us a good idea of the distribution of schools.

Then for those neighbourhoods where there are no schools we find the distance to the closest school by incrementing the radius by 100 metres in each iteration until a school is found using Folium.

On obtaining this refined data, consisting of the distance to nearest school for each neighbourhood and the number of schools in each neighbourhood we cluster this data using K-Means. This helps us identify those regions where there is a need for improvement and those neighbourhoods which have good facilities in terms of schools.
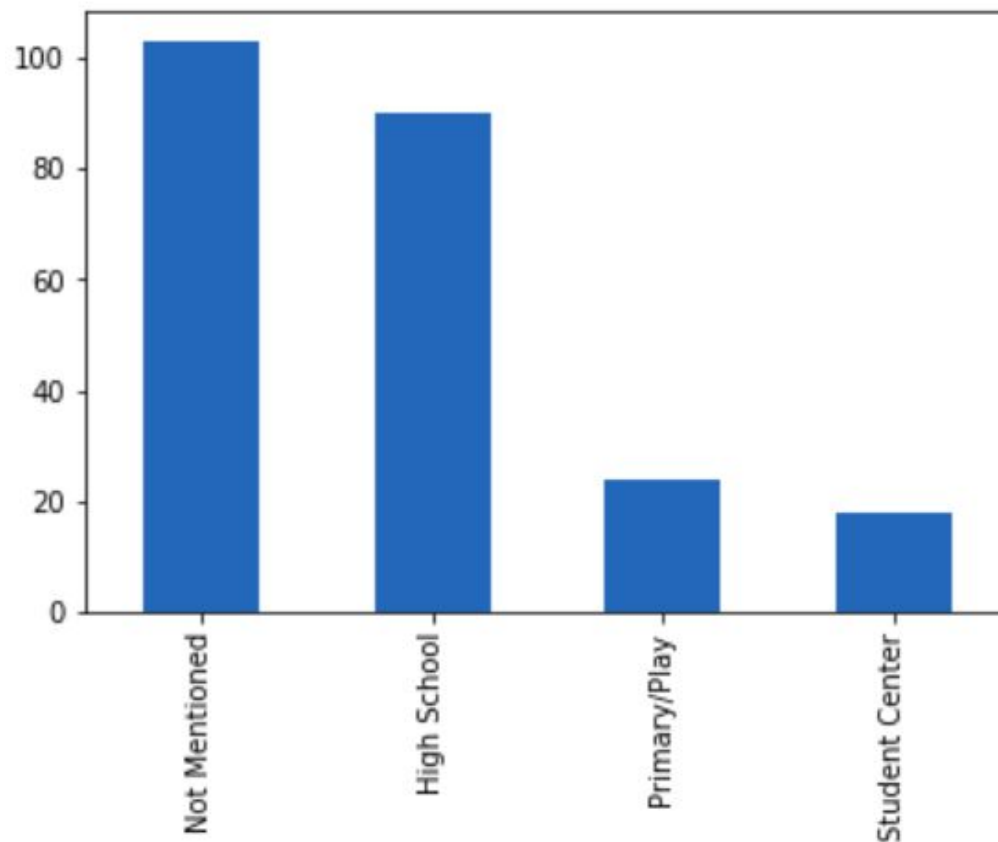
Finally we will discuss and conclude which neighbourhoods require an improvement of facilities in terms of schools and which regions have very good facilities.

ANALYSIS

Based on the data obtained from Folium about schools, I explored the categories to which schools belong.

CATEGORIES
The various categories of schools are as follows:

As we can see most of the schools have not mentioned their category. An interesting thing to note is that a lot of these schools provide education from kindergarten level to high school. Hence there is no surprise in finding only a few schools providing only primary/play level education as the big schools already cover this. Also "Student Center" categorized schools provide education from kindergarten level to high school (included).

DISTRIBUTION OF SCHOOLS IN NEIGHBOURHOODS
I have plotted the choropleth map to give a good idea of the distribution of schools in Hyderabad.
(Note: This map does not cover neighbourhoods in Secunderabad as I could obtain only the geospatial data of those neighbourhoods covered by GHMC.)
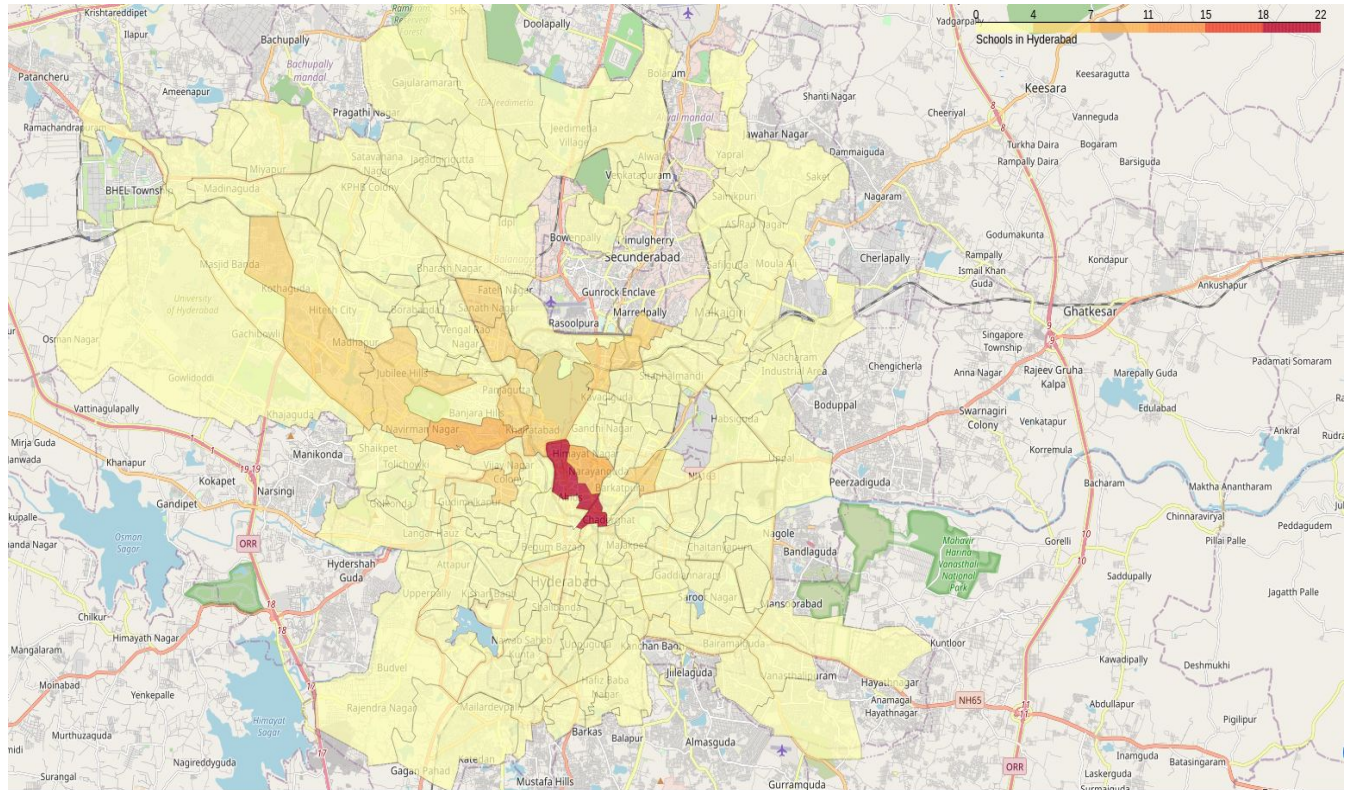
It is clear from the map that most of the schools are clustered in the centre of the city especially near the Abids, Himayath Nagar, Basheerbagh areas. In fact a lot of good schools are present in these neighbourhoods and they are close to each other.

Towards the outskirts of the city, the density of schools decreases which is clear in the map. This is expected as the neighbourhood near the center are more crowded than those in the outskirts.

Also these are older regions and the outskirts are being developed now.

CLUSTERING THE NEIGHBOURHOODS

In this section, I have clustered the neighbourhoods using K-Means clustering algorithm into 3 clusters based on distance to nearest school and number of schools in each neighbourhood.

For those regions that have no schools within the radius of 1km, I have identified the distance to nearest schools using Folium by incrementing the radius in steps by 100 metres until 2km.

Also I have counted the number of schools in each neighbourhood and created a dataframe based on this information.

In this method I have taken nearest distance to school to vary from 1km to 2km. For all those neighbourhoods where the schools may be closer than 1km radius, I have assumed to start from 1km. The first five elements of this dataframe are shown:

	Neighbourhood	Latitude	Longitude	Near_school	num_schools_in_1km
0	Adikmet	17.41061	78.51513	1100	0
1	Alwal	17.53543	78.54427	2000	0
2	Amberpet	17.38582	78.51836	1200	0
3	Asif Nagar	17.38514	78.44738	1200	0
4	Attapur	17.36917	78.43683	1300	0

The average of the distance to nearest school and the number of schools in each neighbourhood has been obtained when grouped by the label assigned in the clustering algorithm.

| label	Latitude	Longitude	Near_school	num_schools_in_1km
0	17.395322	78.478685	1000.000000	18.636364
1	17.412157	78.475229	1025.654450	2.937173
2	17.427935	78.463740	1905.555556	0.000000

Obviously, the average of Latitudes and Longitudes have no meaning in this context.

As it is clear, the neighbourhoods labeled 0 not only have schools within 1km (which let me remind you is the starting point for nearest distance and it can't get any better) but also have a number of options in terms of schools with an average of 18.63 . These neighbourhoods have hence access to excellent education facilities in terms of school level. Some neighbourhoods in this cluster are Abids, Himayath Nagar and Basheerbagh all of which are close to the city center.
The neighbourhoods labeled 1 also have easy access to good education facilities at school level within almost 1km radius with an average of about 3 schools per neighbourhood. Most of the neighbourhoods fall under this category.
The neighbourhoods labeled 2 are the ones requiring improvement as the average distance to nearest schools is close to 2km with no schools within 1km of the neighbourhood. These are present mostly in the outskirts of the city. Some neighbourhoods in this region are Alwal, Jeedimetla, Cherlapally, Patancheru, ECIL X Roads etc.
The plot of each neighbourhoods on map using Folium is as follows:
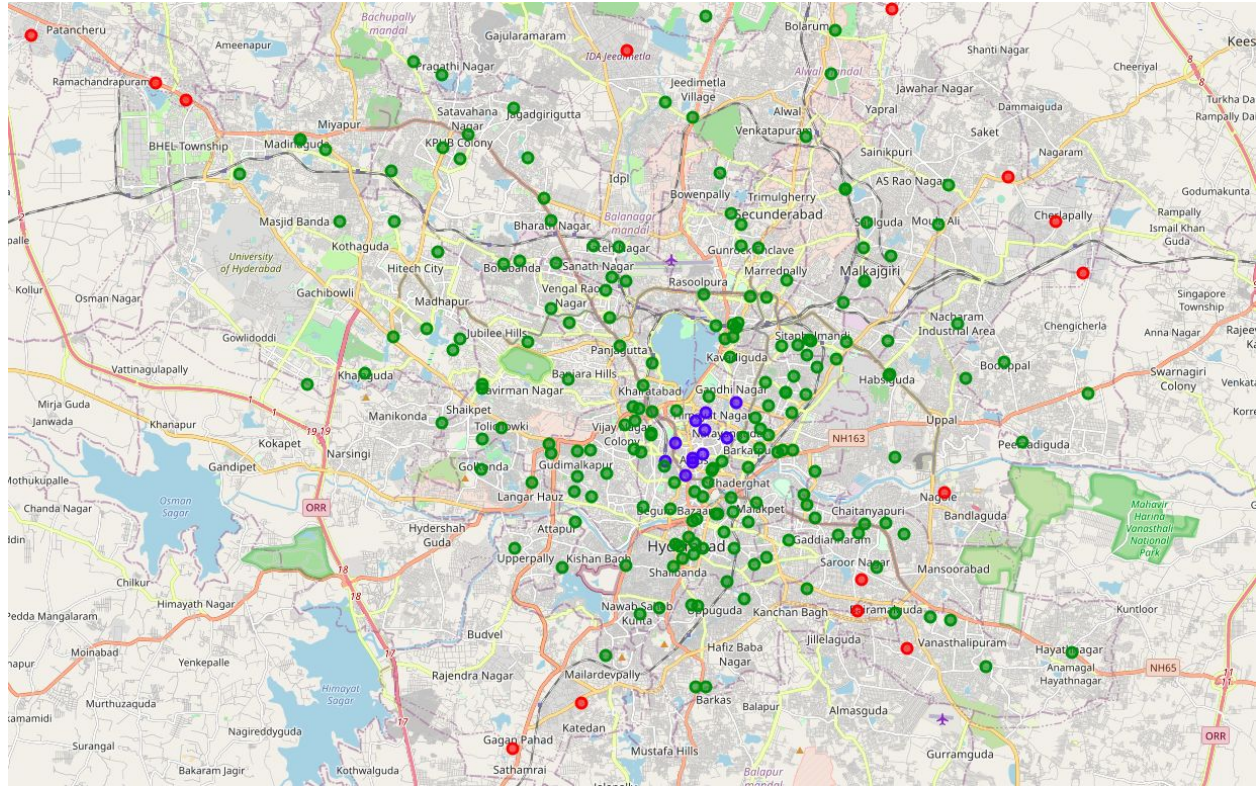
The blue circles correspond to label 0, the green ones to 1 and the red ones to 2.

RESULTS AND DISCUSSION

We have observed the various categories the schools belong to in Hyderabad, with most of them being more of general schools which provide education from kindergarten level itself till 12th grade. Also we observe that schools providing education only at primary level are comparatively less which is expected.

Next we have observed the distribution of schools and find that mostly they are concentrated towards the center of the city and are less dense in the outskirts. This is clear from the Choropleth map.

So when someone is looking to find a neighbourhood with easy access to good facilities, they should find such neighbourhoods closer to the center of the city.

Finally we have clustered the neighbourhoods into 3 clusters and have observed that most of the neighbourhoods fall under category with label 1, i.e, the average distance to nearest school for these neighbourhoods is 1025 metres and there are about 3 schools on average in these neighbourhoods. These regions consist of potential neighbourhoods for the private sector to build schools in, because they don't have many schools but at the same time there will be many people living in these neighbourhoods as they have access to good facilities.

We have also identified the neighbourhoods requiring improvement which is the motive of this analysis. These neighbourhoods have a school with an average radius to closest school close to 2km which was the upper limit set on the distance to nearest school feature.

Some neighbourhoods in this region are Alwal, Jeedimetla, Cherlapally, Patancheru, ECIL X Roads etc. These are the regions requiring attention in terms of improvement at school level. The government should concentrate in these areas.

The full list of neighbourhoods in each cluster can be found in the jupyter notebook I have attached.

CONCLUSION

In this analysis, the distribution of schools in Hyderabad has been analysed. Also the various neighbourhoods have been clustered based on their access to school facilities. The neighbourhoods requiring improvement have been identified along with those neighbourhoods which have potential for profitable schools. Also the various categories of schools have been observed.