FINAL REPORT

CAPSTONE PROJECT – THE BATTLE OF NEIGHBORHOODS

TITLE: ANALYSIS OF SCHOOLS IN BANGALORE

ANALYSIS OF SCHOOLS IN BANGALORE

INTRODUCTION

In this project I have tried to analyze the distribution of schools in Bangalore. It will help us to identify the areas where the infrastructure, in terms of school, needs to be developed. In order to obtain the distribution of schools in Bangalore I have used FourSquare API.

BACKGROUND

It is essential for every child to have quality education. Bangalore is one the IT-hub of India, which means more and more people are migrating to Bangalore to have a better job opportunity. An influx in migration will in turn lead people to search for a neighborhood which have good schooling access.

Thus, this project tries to answer to fold-problems. First, we would like to look at how schools are clustered in the neighborhood of Bangalore and if there are some regions, which lacks infrastructure in terms of schools. Then, it is important to identify these regions for the development of the city and to provide easy access to education for everyone. This is the long term aspect of the project.

While, in the short run we would like to know if a family is migrating to Bangalore, what will be the potential region to reallocate so that they can access good schooling infrastructure.

DATA SECTION

Data Link: "https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore"

I have used Bangalore's Neighborhood Wikipedia page to scrape data for different localities of Bangalore using beautifulsoup4 package. Next, using geocoder I have obtained the Latitudes and Longitudes of these neighborhoods.

Foursquare API Data:

In order to retrieve various school data in various neighborhoods of Bangalore I have used "Foursquare" location information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos.

As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

FourSquare API has been used to obtain the location of various schools in each neighborhood setting the limit to 100 for each neighborhood and a radius of 1km.

DATA PREPROCESSING

After obtaining the location of various schools in the city, there were some unwanted entries such as ‘driving schools’, ‘schools of music’, ’schools of engineering’ etc. These entries were dropped out. Also there was duplication of data because it was possible that a school was within 1000 meters distance from each neighborhood as a result of which it may appear twice with the same latitude and longitude in both these neighborhoods. These have been removed and I have retained only the first copy of such data. Also there were several schools whose category were mismatched or not clear. I have thus categorized them based on the data to whether they are high schools, primary or play schools, student centers or their category was not mentioned.

	Neighbourhood	Latitude	Longitude	School	School_Latitude	School_Longitude	School_category
0	Adugodi	12.94402	77.60800	Mary Immaculate School	12.945718	77.599380	High School
1	Adugodi	12.94402	77.60800	Indian Retail School	12.942211	77.608254	Not Mentioned
2	Adugodi	12.94402	77.60800	Dairy Colony Govt. School	12.938816	77.606491	Not Mentioned
3	Arekere	12.88568	77.59668	AECS Magnolia Maaruti Public School	12.885073	77.596503	Not Mentioned
4	Arekere	12.88568	77.59668	Aradhana School	12.884844	77.592087	Primary/Play
5	Arekere	12.88568	77.59668	Heritage Academy School	12.879765	77.594310	Primary/Play
6	Arekere	12.88568	77.59668	BGS National Public School	12.877473	77.599268	Not Mentioned
7	Arekere	12.88568	77.59668	nakshatra pre school	12.879890	77.603584	Primary/Play
8	Arekere	12.88568	77.59668	Cambridge School	12.886161	77.585502	Not Mentioned
9	Arekere	12.88568	77.59668	Brigade School @ JP Nagar	12.891879	77.585289	High School

Fig1: First 10 elements of the processed DataFrame

Each row corresponds to a particular school and its information, i.e. neighborhood it belongs to, its category, name of the school and its location. There are a total of 389 entries in the DataFrame. Using this final form of data I have analyzed the distribution of schools in Bangalore.

LIBRARIES USED

- ***Pandas***: For creating and manipulating dataframes.
- ***Folium***: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
- ***Scikit-Learn***: For importing k-means clustering.
- ***JSON***: Library to handle JSON files.
- ***XML***: To separate data from presentation and XML stores data in plain text format.
- ***Geocoder***: To retrieve Location Data.
- **Beautiful-Soup and Requests**: To scrap and library to handle http requests.
- ***Matplotlib***: Python Plotting Module.

METHODOLOGY

First, the data has been retrieved and cleaned as explained previously. The data consists of schools within a radius of 1km for each neighborhood 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 neighborhood in a more explanatory way using folium package for visualizing maps. This gives us a good idea of the distribution of schools. Then for those neighborhoods where there are no schools we find the distance to the closest school by incrementing the radius by 100 meters in each iteration until a school is found using Folium. On obtaining this refined data, consisting of the distance to nearest school for each neighborhood and the number of schools in each neighborhood we cluster this data using K-Means. This helps us to identify those regions where there is a need for improvement and those neighborhoods where the influx of migrant can be reallocated so that it offers good facilities in terms of schools.

ANALYSIS

1. EXPLORATORY DATA ANALYSIS

Using Folium, I have mapped the data onto a map to give a general idea of the distribution of schools in Bangalore.

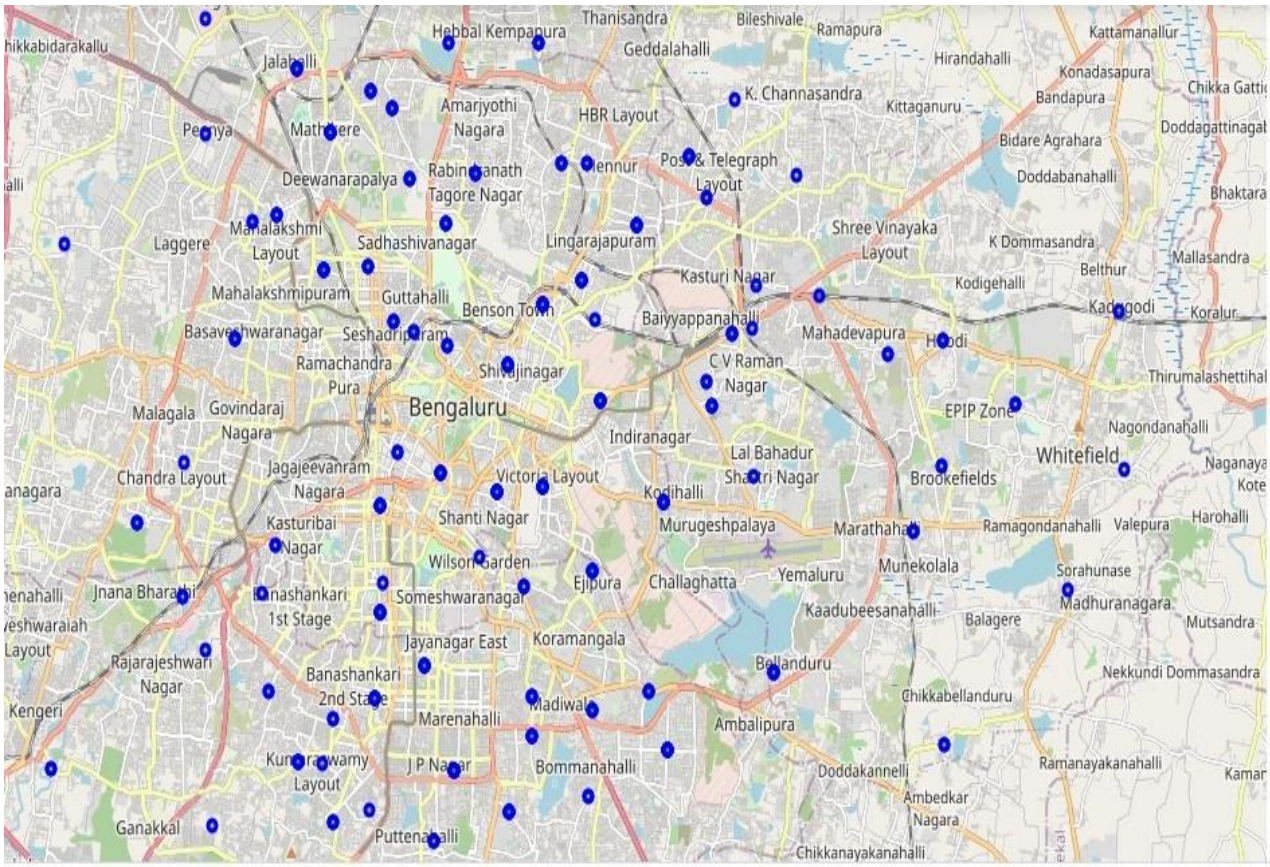

Fig2: General overview of distribution of schools in Bangalore

It can be observed that the schools are clustered more in the centre of the city especially near the Kormangala, Wilson Garden areas. We can also see that Whitefield which is a prominent destination for IT-hubs lacks infrastructure in terms of schools. Towards the outskirts of the city, the density of schools decreases.

For the ease of analysis I have broadly categorized schools into four groups. They are High School, Primary, Student Centre and the one which cannot be deciphered properly is taken as Not-Mentioned.

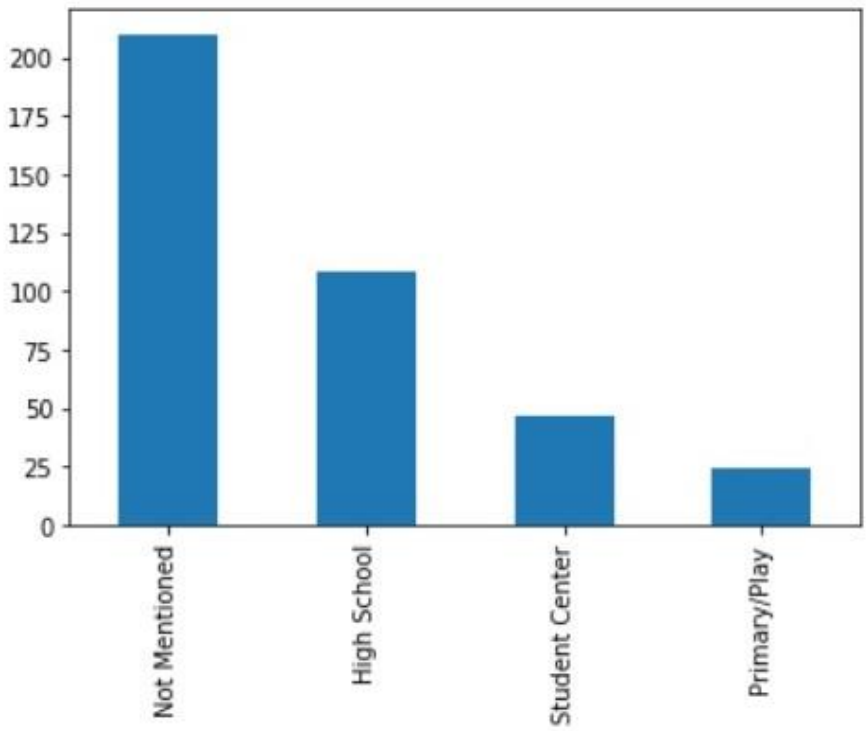

Fig3: Different categories of school in Bangalore

As we can see most of the schools have not mentioned their category because a lot of these schools provide education from kindergarten level to high school. Also, the number of High Schools are quite more in number than in case of Primary Schools.

2. CLUSTER ANALYSIS

In this section, I have clustered the neighborhoods using K-Means clustering algorithm into 3 clusters based on distance to nearest school and number of schools in each neighborhood. 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 meters until 2km. Also I have counted the number of schools in each neighborhood and created a data frame based on this information. In this method I have taken nearest distance to school to vary from 1km to 2km. For all those neighborhoods where the schools may be closer than 1km radius, I have assumed to start from 1km.

	Neighbourhood	Latitude	Longitude	Near_school
0	Agara, Bangalore	12.84283	77.48759	2000
1	Ananthnagar	12.95408	77.54135	1100
2	Anjanapura	12.85811	77.55909	1300
3	Bommasandra	12.81753	77.67879	1100
4	Devara Jeevanahalli	13.01444	77.59951	1100

Fig4: First five elements showing the nearest distance

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

label	Latitude	Longitude	Near_school	num_schools_in_1km
0	12.975056	77.610372	1034.862385	1.871560
1	12.936445	77.491425	1750.000000	0.000000
2	12.964732	77.597203	1000.000000	8.978723

Fig5: Number of schools within 1km radius under each cluster

We can see that the neighborhoods labeled 2 perform best with about an average of 9 schools in 1km radius. In comparison the neighborhood 1 has about an average of 2 schools within 1km radius. The neighborhoods labeled 1 are the ones that require improvements as the average distance to nearest schools is close to 2km with no schools within 1km of the neighborhood.

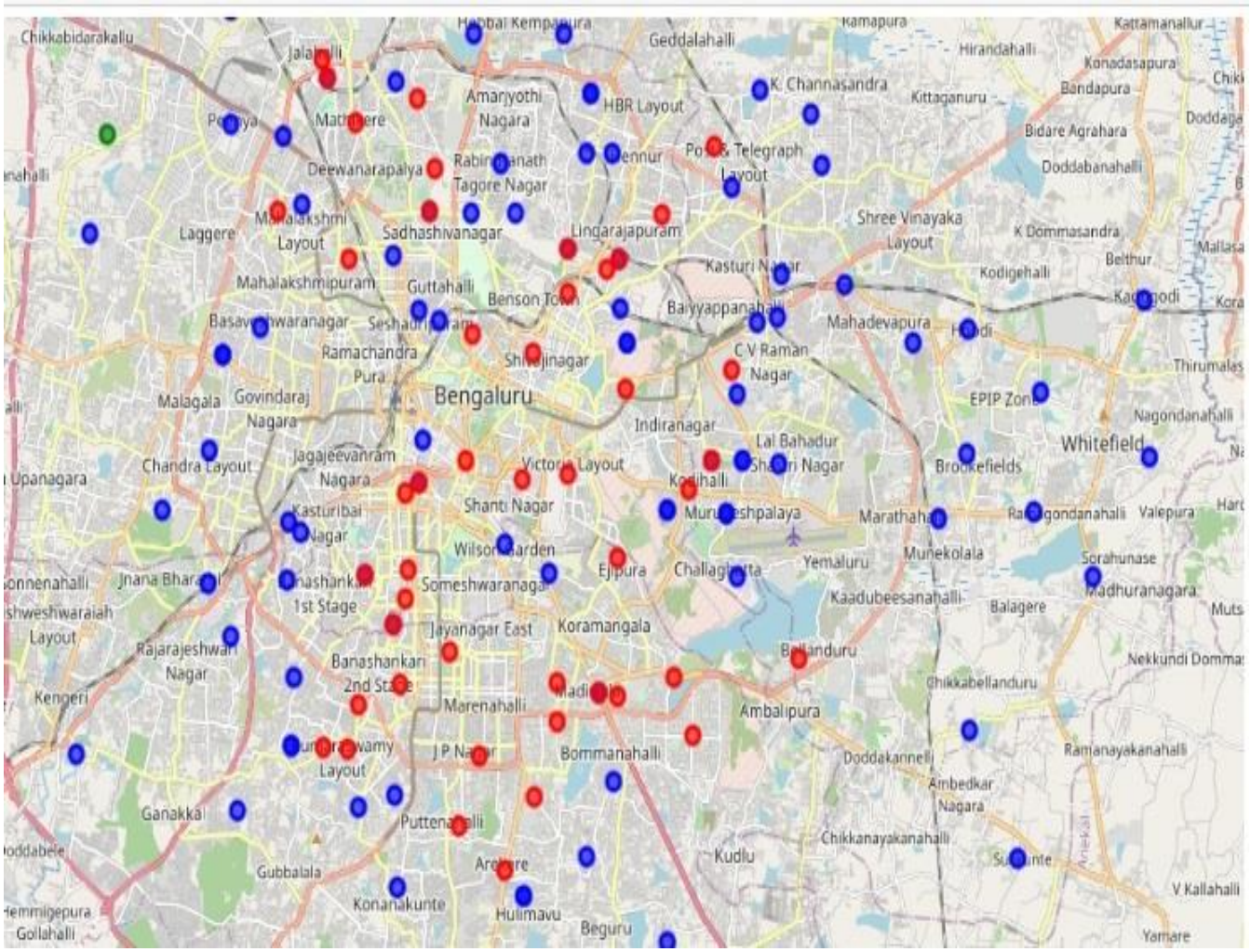

Fig6: Distribution of schools in different clusters

Here, the blue circles correspond to label 0, the green ones to 1 and the red ones to 2.

RESULT

Among the schools that have been observed in Bangalore, most of them are standard 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 compared to the High schools. 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.

Finally we have clustered the neighborhoods into 3 clusters and have observed that most of the neighborhoods fall under category with label 2 that is, the average distance to nearest school for these neighborhoods is 1000 meters and there are about 9 schools on average in these neighborhoods. The neighborhoods labeled 0 also have easy access to good education facilities at school level within almost 1km radius with an average of about 2 schools per neighborhood. The neighborhoods labeled 1 are the ones which requires improvement as there no schools within 2km of radius.

DISCUSSION

Neighborhood labeled 2 have the most number of schools in their vicinity and they are highly clustered around the central Bangalore. Some neighborhoods in this cluster are Kormangala, Jaynagar East, Victoria Layout etc. These are also the prominent IT-parks, thus people migrating to label-2 cluster will not find much difficulty in terms of good-schooling infrastructure.

We have also identified the neighborhoods requiring improvement which is the motive of this analysis. These neighborhoods have a school with an average radius to closest school close to 2km which is the upper limit set on the distance to nearest school feature. Some neighborhoods in this region are Bandapura, Balagere,Ramayanakanahalli etc. These are the regions which require attention in terms of improvement at school level. The government should concentrate on these areas. Also, Whitefield which is a prominent employment hub lacks school in its vicinity. So emphasis must be laid down to build new schools around its neighborhood.

CONCLUSION

In this project, using k-means cluster algorithm I separated the neighborhood into three clusters.This project has shown me a practical application to resolve a real situation that has a personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

The following detailed report concludes that firstly, Bangalore lacks Primary schools which is the first building block of education. Considering the mass of young work-force government along with private organization must build schools near to the neighborhood of major IT-hubs.
Secondly, people migrating to Bangalore should try to reallocate primarily in the label 2 cluster as shown in the analysis which provides good schooling facilities. The neighborhoods near the vicinity of label 1 clusters require massive help from government to lay the educational infrastructure.