CAPSTONE 2020

MAY 2

Location clustering based on hospital density

Authored by: Palagati Bhanu Prakash Reddy

Introduction:

Humans have many cravings and they will fight for all of them. However, we cannot create a quandary by bringing life on one hand and other things on the other hand. Hear most people will pick life over any other thing out there. That brings us no surprise, after all we are living organisms.

Even though we are willing to do anything in order to save our life, unreachability of hospitals is the paramount among all for loosing life in the first place. Consider, there is a accident almost 80% of the time the person who got involved in the accident can sustain if he gets the appropriate treatment at appropriate time.

This project is about analyzing various neighborhoods in Hyderabad, India. As a result, we get a detail idea about various neighborhoods opening doors for building a new hospital.

Business Problem:

We will never run out of investors on earth. Since the inception of the modern medical structure and the fear people having on their lives. Most people are considering hospitals as the primary sources to invest.

This project is about analyzing and clustering various locations in Hyderabad, India, which is a cosmopolitan city. Since it is a cosmopolitan city a lot of investors are willing to invest in this land to find out quests. Finally, after completing this project we will be classifying the locations with the density of hospitals. If somebody want to start a new hospital, they can use this analysis to start one at places where the density is very less. As a result, they will be making huge money.

Beneficiaries:

Broadly there are two groups of beneficiaries with this project.

1. Investors
2. General Public.

As we already discussed in the target audience investors are the direct people who are going to get benefit from the analysis.

However, if people are investing in building hospitals in the places where the density is less the people living in those places can access hospitals easily, as a result, their standards of living will get improve. That is a real win-win situation.

Data:

REQUIREMENT:

In order to perform the analysis, we need neighborhood data of Hyderabad, India. Precisely we need the following four columns:

1. Neighborhood
2. Latitude
3. Longitude
4. Hospitals count

STEPS:

1. Using the beautifulsoup library scrape the Wikipedia site where the neighborhoods data of Hyderabad, India, is available. Then cast it into a dataframe.

2. Using geolocator package get the location data then append those latitudes and longitudes to the dataframe.

3. Using foursquare get the venues belong to the medical category with a killometer radius.

4. Finally count the hospitals in each location and add this as a column to the dataframe.

As a result of the above steps, we get dataset containing neighborhoods, latitudes, longitudes, and number of hospitals.

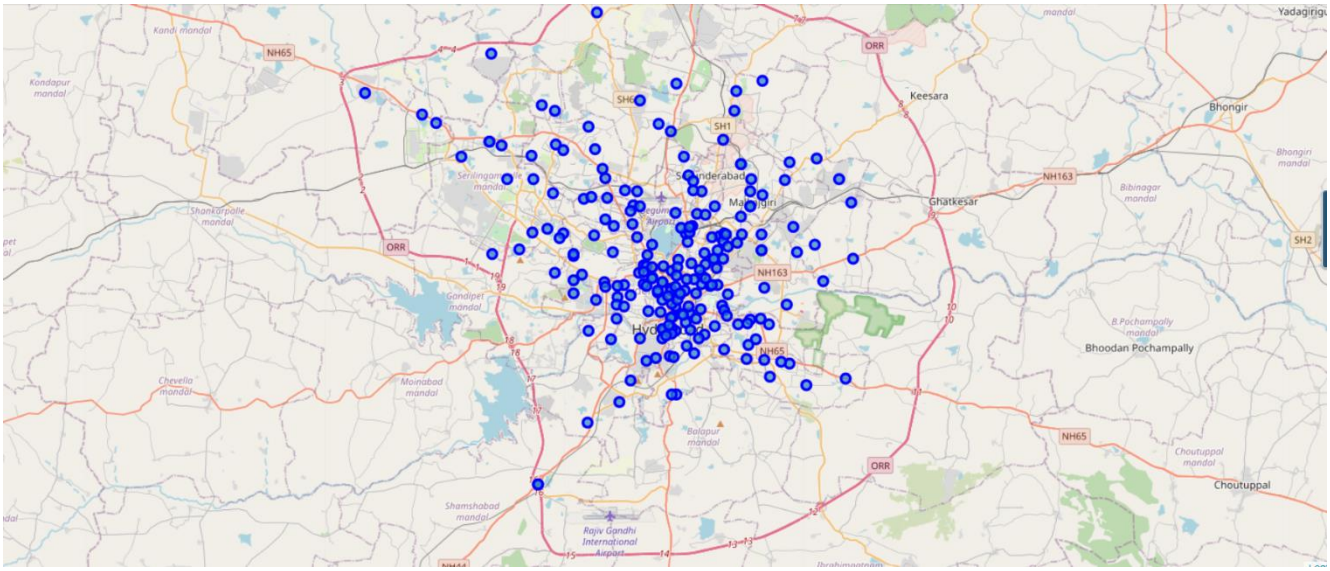
Methodology:

Firstly, we need data to do anything. As the project is based on the neighborhood in Hyderabad, India, Wikipedia is the best place to start. **https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India.**

The above links serves as a location of information which we can use. It consists of 220 names of neighborhoods in Hyderabad. However, that is not in a tabular format so, we cannot use pandas read_html function. As a result, I have used Beautifulsoup package to do web scraping. It has the result spread out in two pages so, there is a requirement of scraping two pages. After this step I am having a list of 220 neighborhoods in Hyderabad. After that it is converted into a dataframe to facilitate the work using the data scraped.

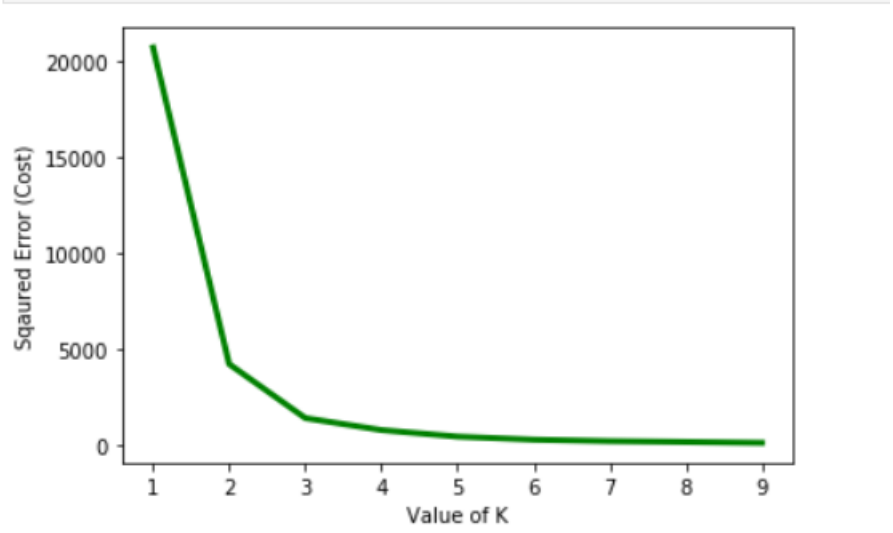
As a result, of the above scraping we are now having a pandas dataframe with neighborhoods. However, we need the exact geocoordinates in order to explore or map the data. So, I am using Geocoder package in python for which you need to pass the address in string format and it returns the coordinates of the address we passed. So, I have passed the address of all the neighborhoods to get the coordinates of all the neighborhoods. After that I have created Latitude and Longitude columns in the dataframe and appended these lists to those columns of the dataframe. As a result, now the dataframe is equipped with everything to work with the Foursquare API.

As this research is about the number of hospitals in a neighborhood. Let's communicate with the Foursquare API to get the venues of category hospitals. Then count how many hospitals were returned for each neighborhood within the radius of a kilometer. Now it is time to create a new column to store the results provided by the Foursquare API into the dataframe as a column "hospitalCount". In order to verify whether all the location results of geocoder and foursquare are accurate let's plot the points on map using the folium package.



Map view of the processed data before clustering.

Now it is time to do some analysis. Precisely let's make the location data into clusters based on the number of hospitals available. The next important question is finding the number of clusters for the Kmeans algorithm. So, let's train algorithm for different k values and find the best one.



Graph depicting the best K value i.e., 3

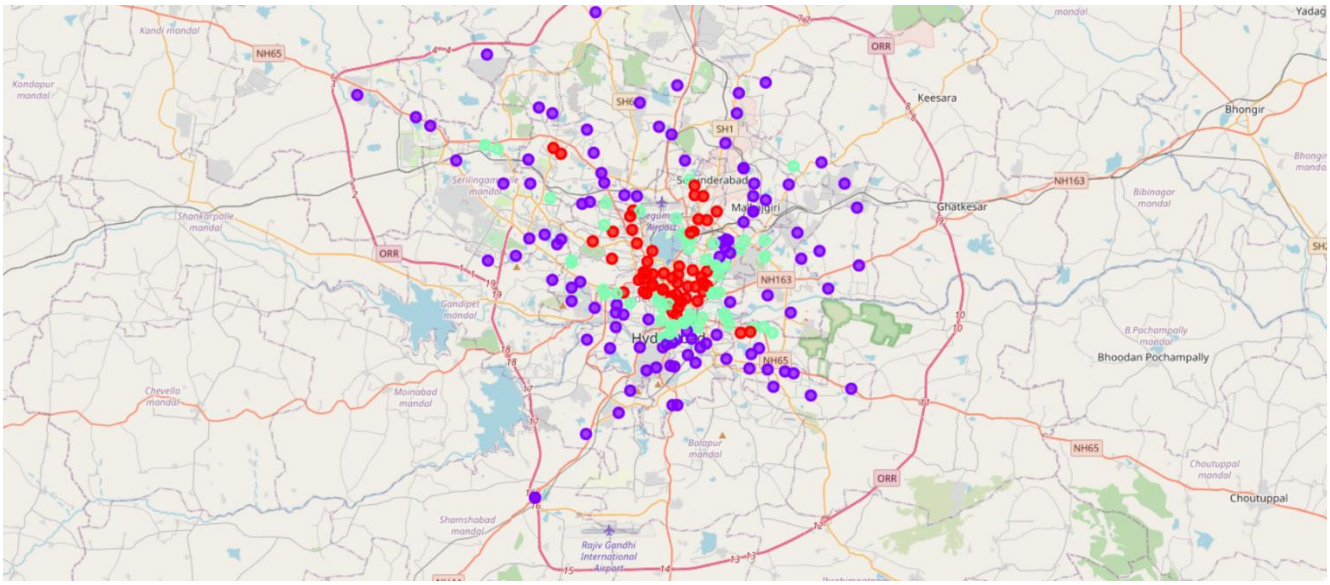
So, let's finalize the k value as 3 and train the algorithm. As a result we have labels of all the neighborhoods with 0, 1, and 2.

Results:

After training the algorithm it will result in the labels for each neighborhood. So, let's add the list of neighborhoods to the dataframe and call that column as "Labels". Now the dataframe is equipped with all the required columns.

Let's see the mean of each label

1. Cluster 0 – Red – 25 hospitals per kilometer.
2. Cluster 1 – blue – 2 hospitals per kilometer.
3. Cluster 2 – mint green – 13 hospitals per kilometer.



Map After clustering and labeling.

Discussion:

On observing the above depiction of various clusters on the map it is clear that high density of hospitals was in the very center of the city. The plausible reason is because that is the old city and it is densely populated. This cluster is colored with red. Then this cluster is surrounded by another cluster that is cluster 2 which is in mint green. This delineates the fact that the expansion of the city at this place started long back compared to the next layer. Here there is an average of 13 hospitals per km radius. Finally, there is another layer which is surrounding the previous two layers. Even though these are densely populated and the residential plot of high-income group. It was not colonized by many hospitals. This group is depicted with a blue color and it has an average of 2 hospitals per km radius.

Conclusion:

This project is about analyzing the frequency of the hospitals in different neighborhoods in Hyderabad, India. The main aim of this research is to give valuable information to the venture capitalists who are willing to invest in the hospital sector. If they are investing in the low frequency group, they are going to get high influx of the patients. Along with that it will also be helpful for the patients because the hospitals are evenly distributed so, it is easy to reach one. As a result, a patient can reach to a hospital on time. This is a kind of Win-Win situation where both the investors and the normal people are going to get benefitted.