# Most Similar Neighborhood In Bangalore

Namrata Choudhary

April 29th, 2020

# Table of Contents

# 1. Introduction

### 1.1 Background

Bengaluru is sometimes referred to as the "Silicon Valley of India" (or "IT capital of India") because of its role as the nation's leading information technology (IT) exporter. Indian technological organisations ISRO, Infosys, Wipro and HAL are headquartered in the city.The establishment and success of high technology firms in Bangalore has led to the growth of Information Technology (IT) in India. IT firms in Bengaluru employ about 75% of India's pool of 2.5 million IT professionals and account for the highest IT-related exports in the country.Every month so many IT engineers are relocating to bangalore from all other parts of India alongwith their family or alone.India is the country of diversity and the cultural difference between South and North India is so obvious, that you feel they belong to two different continents.So the people travelling from all other parts of India are willing to relocate to the similar neighborhood where they used to stay. Therefore, it is advantageous for people to identify whether and how much a
Neighborhood they are shifting to is similar to the one they are living in now.

### 1.2 Problem Statement

Exploring the neighborhood of the person and comparing it with the neighborhood of bangalore.This project aims to recommend the person which locations in bangalore are having similar neighborhoods to the one they are living in now.We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that the best possible final location can be chosen by the person.

# 2. Data acquisition and cleaning

### 2.1 Data sources

We decided to use locations(All SO & BO) centered around PinCodes of Bangalore, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

1. For exploring neighborhoods of Bangalore and the user location the pincode data of bangalore has been extracted from the internet.
2. Centers of candidate areas will be based on pincodes and approximate addresses of centers of those areas will be obtained using geopy.

3. Number of venues in the neighborhood and their type and location in every neighborhood will be obtained using the Foursquare API.
4. Coordinate of the Bangalore center will be obtained using geopy.

### 2.2 Data cleaning

The pin code dataset has been downloaded from https://finkode.com/ka/bangalore.html.The dataset contains the Pincode,PostalOfficeName and the District column. The dataset has been modified by grouping the data on the basis of Pincode and adding different PostalOfficeName corresponding to Pincode in a string column named Neighborhood. The latitude and longitude of the pincodes has been fetched using a geopy library.Some pincodes whose coordinates were not available/incorrect we have dropped those pin codes as the data was incorrect.By using FourSquareAPI , nearby venues within 500m radius has been fetched and explored.FourSquareAPI is an API which returns the nearby venue categories based on the coordinates/addresses provided. We have used pincode to fetch nearby venues.

## 3. Methodology

In this project we will direct our efforts on detecting areas of Bangalore that have neighborhoods similar to the users neighborhood.

1. In first step we have collected the required **data: pincode, areas in the pincode as neighborhood,venues nearby the pincode(according to Foursquare categorization).
2. Second step in our analysis will be calculation and exploration of 'neighborhood' across different areas of bangalore - ,what are the most frequent venues per pincode,doing proper encoding of data, and then applying k means clustering on the top of it with k=5. That means we are going to divide our entire bangalore area in 5 clusters.
3. In the third and final step we will focus on the most similar area as per the user input and we will take the current place and current pincode as input from the user and then we will identify the areas in the bangalore which is most similar to the one specified by the user..

## 4. Exploratory Data Analysis

### 4.1 Rawdata :

We explored our pincode data along with coordinates using a folium map library. An empty map centred at bangalore has been created .All the pin codes have been plotted as a data point circle in the map.

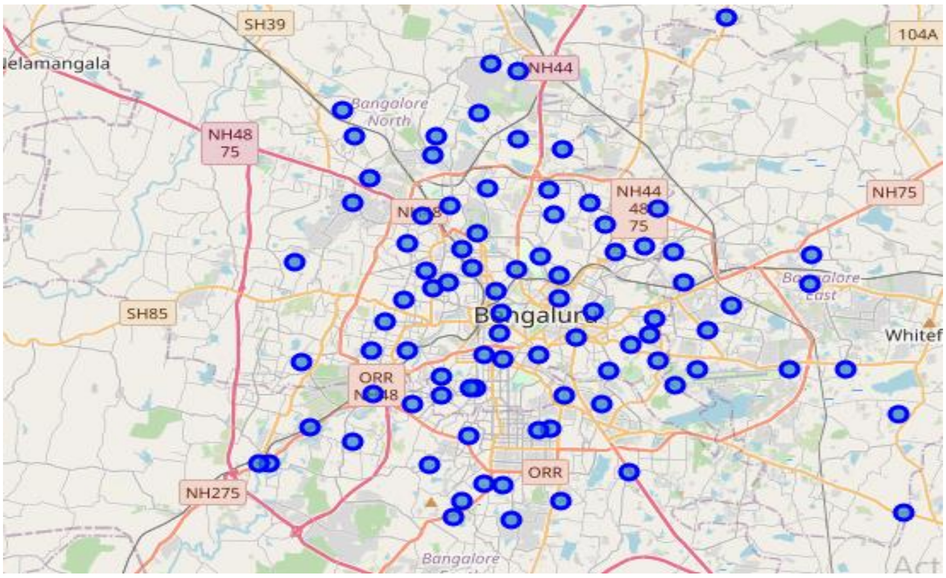
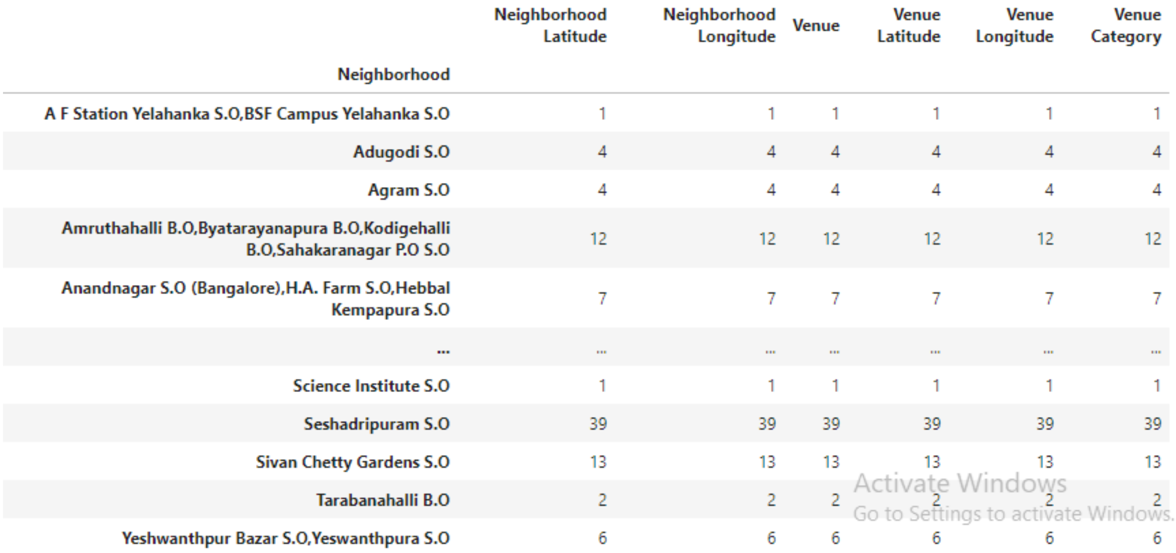


Fig 1 Bangalore Pincodes

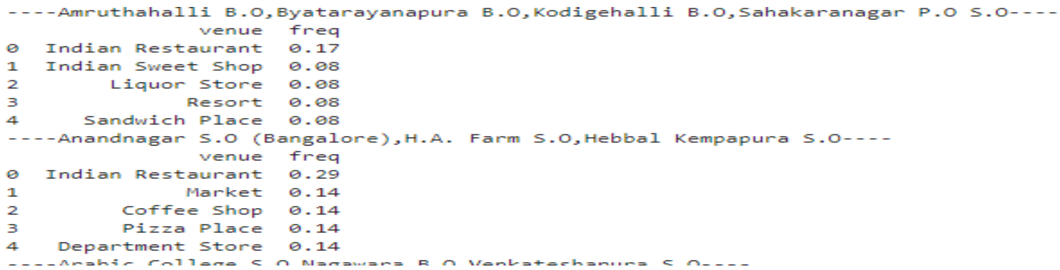
We explored the count of venues per neighborhood.



| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| A F Station Yelahanka S.O,BSF Campus Yelahanka S.O | 1 | 1 | 1 | 1 | 1 | 1 |
| Adugodi S.O | 4 | 4 | 4 | 4 | 4 | 4 |
| Agram S.O | 4 | 4 | 4 | 4 | 4 | 4 |
| Amruthahalli B.O,Byatarayanapura B.O,Kodigehalli B.O,Sahakaranagar P.O S.O | 12 | 12 | 12 | 12 | 12 | 12 |
| Anandnagar S.O (Bangalore),H.A. Farm S.O,Hebbal Kempapura S.O | 7 | 7 | 7 | 7 | 7 | 7 |
| ... | ... | ... | ... | ... | ... | ... |
| Science Institute S.O | 1 | 1 | 1 | 1 | 1 | 1 |
| Seshadripuram S.O | 39 | 39 | 39 | 39 | 39 | 39 |
| Sivan Chetty Gardens S.O | 13 | 13 | 13 | 13 | 13 | 13 |
| Tarabanahalli B.O | 2 | 2 | 2 | 2 | 2 | 2 |
| Yeshwanthpur Bazar S.O,Yeswanthpura S.O | 6 | 6 | 6 | 6 | 6 | 6 |

Fig2: EDA2

After adding the nearby venues to each pincode we also explored the top 5 most frequent venues per neighborhood. Some results are:



```
----Amruthahalli B.O,Byatarayanapura B.O,Kodigehalli B.O,Sahakaranagar P.O S.O----
               venue  freq
0  Indian Restaurant  0.17
1  Indian Sweet Shop  0.08
2       Liquor Store  0.08
3             Resort  0.08
4     Sandwich Place  0.08
----Anandnagar S.O (Bangalore),H.A. Farm S.O,Hebbal Kempapura S.O----
               venue  freq
0  Indian Restaurant  0.29
1             Market  0.14
2        Coffee Shop  0.14
3        Pizza Place  0.14
4   Department Store  0.14
----Arabic College S.O,Nagawara B.O,Venkateshapura S.O----
```

Fig3: EDA3

We have transformed the data using one hot coding and then used a clustering algorithm to group the similar neighborhoods in bangalore

### 4.2 Clustering  Algorithm :

K-Means clustering is an unsupervised learning algorithm that finds a fixed number (k) of clusters in a set of data. A cluster is a group of data points that are grouped together due to similarities in their features.In our case we have used K-Means clustering algorithm to partition our entire bangalore areas into 5 (k) neighborhoods.
As already mentioned we have considered the nearby venues  within 500m of a particular pincode as features input. On the basis of the similarity/dissimilarity of these features K-Means  has clustered the data into 5 groups.

Areas after clustering can be visualize using folium maps as below:

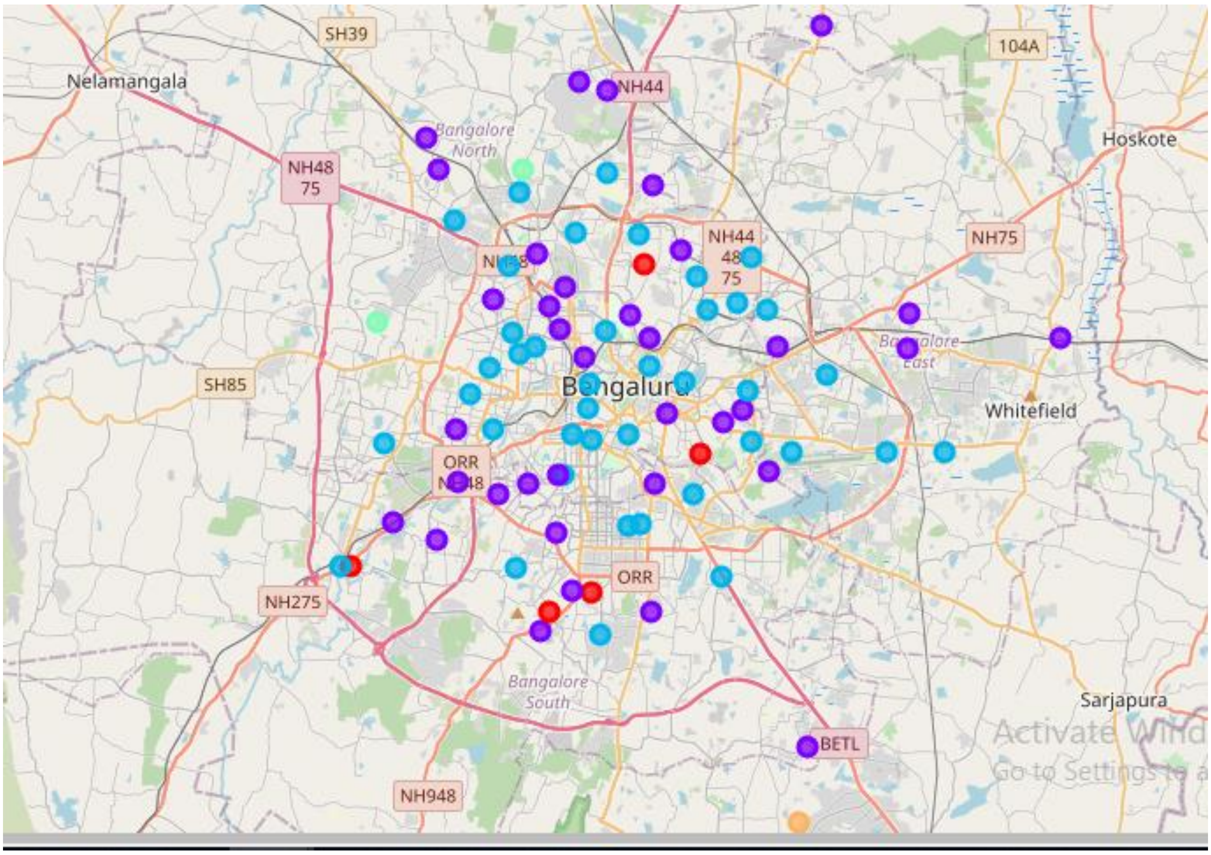


Fig4 Clustered areas with different colors as labels

### 4.3 Recommending Neighborhood:

We can take three parameters as input from users: current city, current pincode and current area. By using these parameters we can predict the cluster label it belongs too.

Just for demonstration purpose I have taken an input as:

current_place='Sarojini Nagar'
current_pincode=110023
current_city='Delhi'

After fetching the latitude and longitude, nearby venues of the input data and performing proper transformation we have predicted using the model we built on bangalore data.We found that the predicted label is 2. That means the neighborhood similar to Sarojini Nagar is in the area enclosed by cluster 2. Cluster 2 contains 40 pincode and the user can relocate to any of the areas shown below:

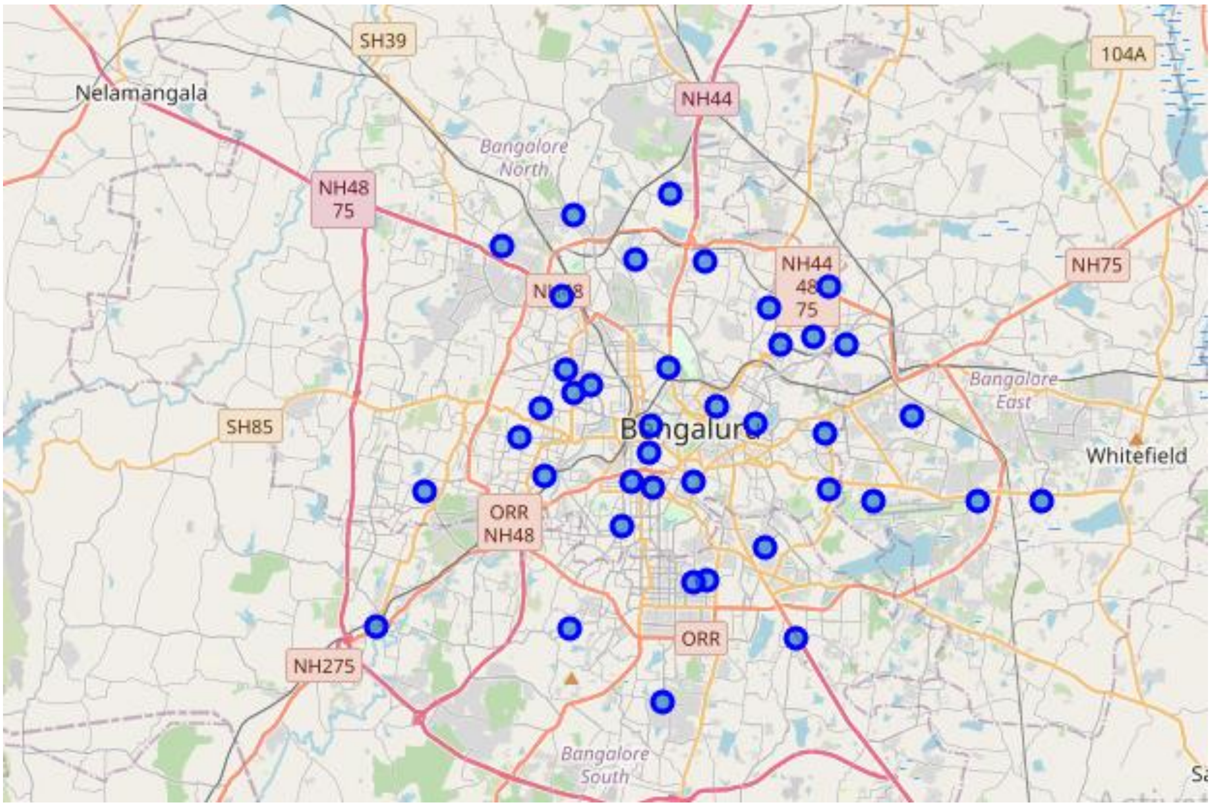


Fig5 : Cluster 2

### 4.4 Measuring Accuracy(Similarity Score):

In clustering algorithms it is very difficult to quantify your results, how well your model is performing how similar they are from the predicted clusters.Informally, accuracy is the fraction of predictions our model got right.Keeping this in mind, I tried to quantify the accuracy , better to say similarity_score as follows:

Among the list of all venues near by user location(test data) within 500m range how many venues are actually located near a particular bangalore area(pincode) within 500m range, where bangalore area(pincode) is from the predicted cluster.

So let’s say we have 10 venues in user_data and out of that the Jalahalli H.O ( a point in the cluster 2) is having 8 venues similar to user_data. So, accuracy here is 0.8.

So we added one more parameter to the existing results as similarity_score.
After calculating similarity scores for all the data points available in the predicted clusters , we selected one  with the max accuracy as most recommended.

Although the user can choose any area from Fig 5 or cluster 2 but among them we tried to recommend the most similar on the basis of similarity_score.In our case we found 560024 Anandnagar S.O (Bangalore),H.A. Farm S.O,Hebbal has the most similar neighborhood as Sarojini Nagar , having similarity_score=0.91. The visual is shown below:

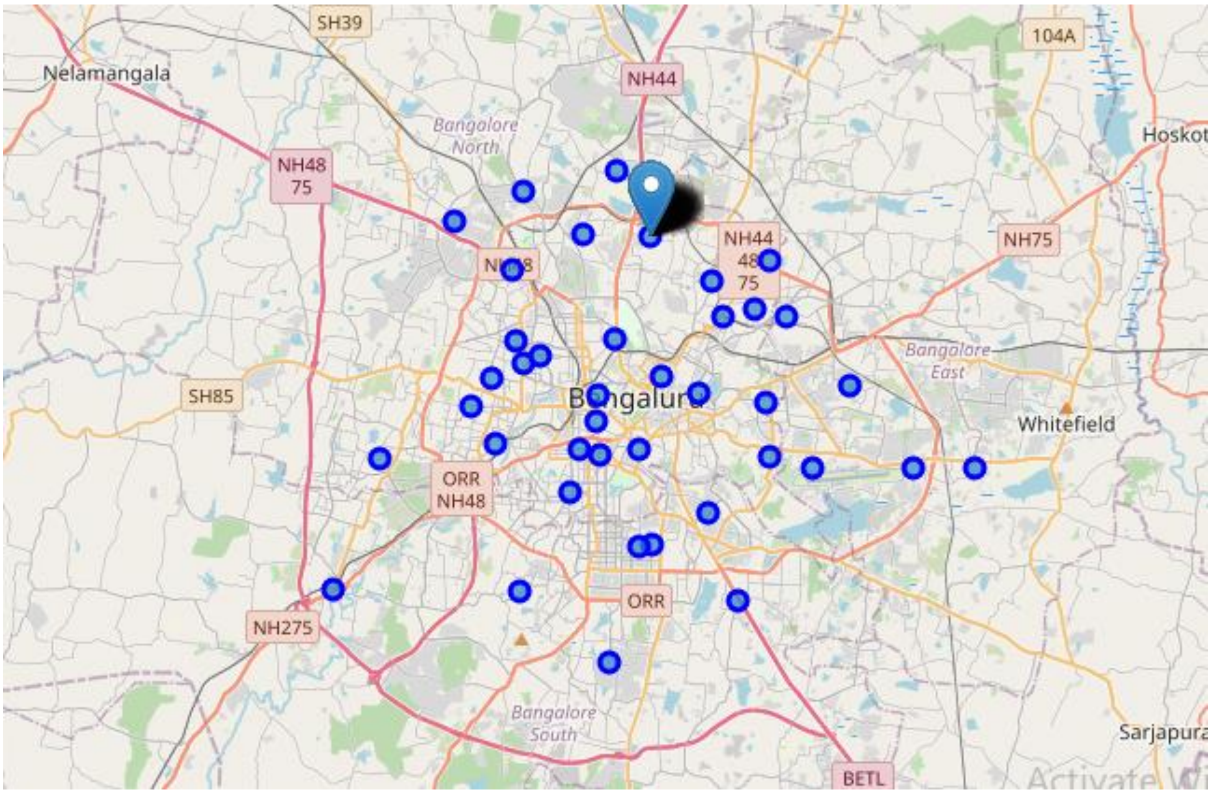


Fig 6 : 560024  Anandnagar S.O (Bangalore),H.A. Farm S.O,Hebba highlighted as marker.

## 5.Results and Discussions

In our analysis  we have taken a demo input let's say the person living nearby Sarojini Nagar ,Delhi(PinCode-110023) and willing to relocate to bangalore we got all the nearby venues of Sarojini Nagar in 500m radius using FoursquareAPI. It is going to work with any user provided input.

We have taken all the pincodes of Bangalore  as a data point and found the nearby 50 venues within 500m of those pin codes.Those location candidates were then clustered in 5 groups to create

zones of interest which contain similar kinds of neighborhoods. All the pincodes are labelled on the basis of their cluster group. Considering our demo case, we found the zone areas(cluster group 2) which is having neighborhood similar to Sarojini Nagar,Delhi.After identifying the clusters we tried to find the similarity score(accuracy)for each area(pincode) in the given cluster on the basis of percent of venues matched in the user location and predicted areas.
On the basis of Similarity scores we have highlighted the area having maximum similarity.

Result of all this is 40 areas similar to the user location has been predicted alongwith their similarity score.We have found Anandnagar S.O (Bangalore),H.A. Farm S.O,Hebbal(560024) is very similar to Sarojini Nagar alongwith similarity of 91%. This, of course, does not imply that those areas are actually optimal locations for a relocation! Purpose of this analysis was to only provide info on bangalore areas similar to the neighborhood of the user area, but other factors can also be included like commuting facilities, distance from office,levels of noise etc.

## 6.Conclusions:

Purpose of this project was to identify the Bangalore city areas similar to the neighborhood area of the user in another city,the information can be utilised for relocation purposes. By using the FoursquareAPI we clustered the bangalore areas on the basis of nearby venues(similar neighborhood) and then recommended the similar areas to the user along with similarity score.

Final decision on relocation will be made by users based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, commuting facilities etc**.**