IBM Applied Data Science Capstone

Clustering Bus Stations in the City of Bengaluru, India

By: Himanshu Baswal

June 2021

Introduction

A bus station is a structure where city or intercity buses stop to pick up and drop off passengers. While the term bus depot can also be used to refer to a bus station, it generally refers to a bus garage. A bus station is larger than a bus stop, which is usually simply a place on the roadside, where buses can stop. It also often provides a convenient point where services can be controlled from. The size and nature of a terminal may vary, from a roadside bus stop with no facilities for passengers or bus crews, to a purpose-built off-road bus station offering a wide range of facilities. Expanding the Network of buses over the entire city can help in connecting the small localities of the cities to the mainland area.

Business Problem

The objective of the capstone project is to analyse and select the best locations in the city of Bengaluru, India to open up bus stations. Using the Machine Learning methods like clustering, this project aims to provide solution to answer business question: In Bengaluru, after the expansion of the city to 741km^2, the newer area of the city needs to connect to the mainland, where would you recommend that they set up bus stations?

Data Description

To solve the problem, we will require the following data:

- List of Neighbourhoods in Bengaluru. This defines the scope of the project that is limited to Bengaluru in South Asia.
- Latitude and Longitude of the neighbourhoods. This is used for plotting the map and also to acquire the venue data.
- The Venue here is the bus stations, so we will need them and we shall perform clustering on these neighbourhoods.

Sources of data and methods to extract them:

This following Wikipedia page:

(https://commons.wikimedia.org/wiki/Category:Suburbs_of_Bangalore) contains a list of neighbourhoods in Bengaluru, with a total of 58 neighbourhoods. We make use of web scraping techniques to extract the data from Wikipedia page, with help of python requests and beautiful soup packages. Then we will get the geographical coordinates of neighbourhoods using Python geocoder package which gives us the latitude and longitude coordinates of the neighbourhoods.

After this, we make use of foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide us many categories of venue data, we are interested particularly in Bus Stations in order to help us solve the problem put forward. This project will make use of *Pandas, NumPy, Json, Matplotlib and Sci-kit learn* packages. *Folium* for map visualization will also be used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Bengaluru. Fortunately, that list is available in Wikipedia page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Bangalore). We will web scrap using python requests and beautiful soup packages to extract the list of neighbourhood data. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API.

To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a panda Dataframe and then visualize the neighbourhoods in a map using Folium package. This allows us to make sure that geographical coordinates data returned by geocoder are correctly plotted in the city of Bengaluru.

Next, we use Foursquare API to get the top 100 venues that are within a radius of 5000 meters. We register a Foursquare Developer account to obtain the Foursquare ID and Foursquare Secret key. We then make API calls to Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and venue longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Bus Stations” data, we will filter the “Bus Station” as venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering algorithm. It identifies k number of centroids and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence of “Bus Station”. The results will allow us to identify which neighbourhoods have higher concentration of Bus Stations while which neighbourhoods have lower

concentration of Bus Stations. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Bus Stations.

Results

The results from k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Bus Station”:

- Cluster 0: Neighbourhoods with highest concentration of Bus Stations
- Cluster 1: Neighbourhoods with lowest concentration of Bus Stations
- Cluster 2: Neighbourhoods with moderate concentration of Bus Stations

The results of the clustering are visualized in the map below with cluster 0 as red colour, cluster 1 as green colour and cluster 2 as purple colour.

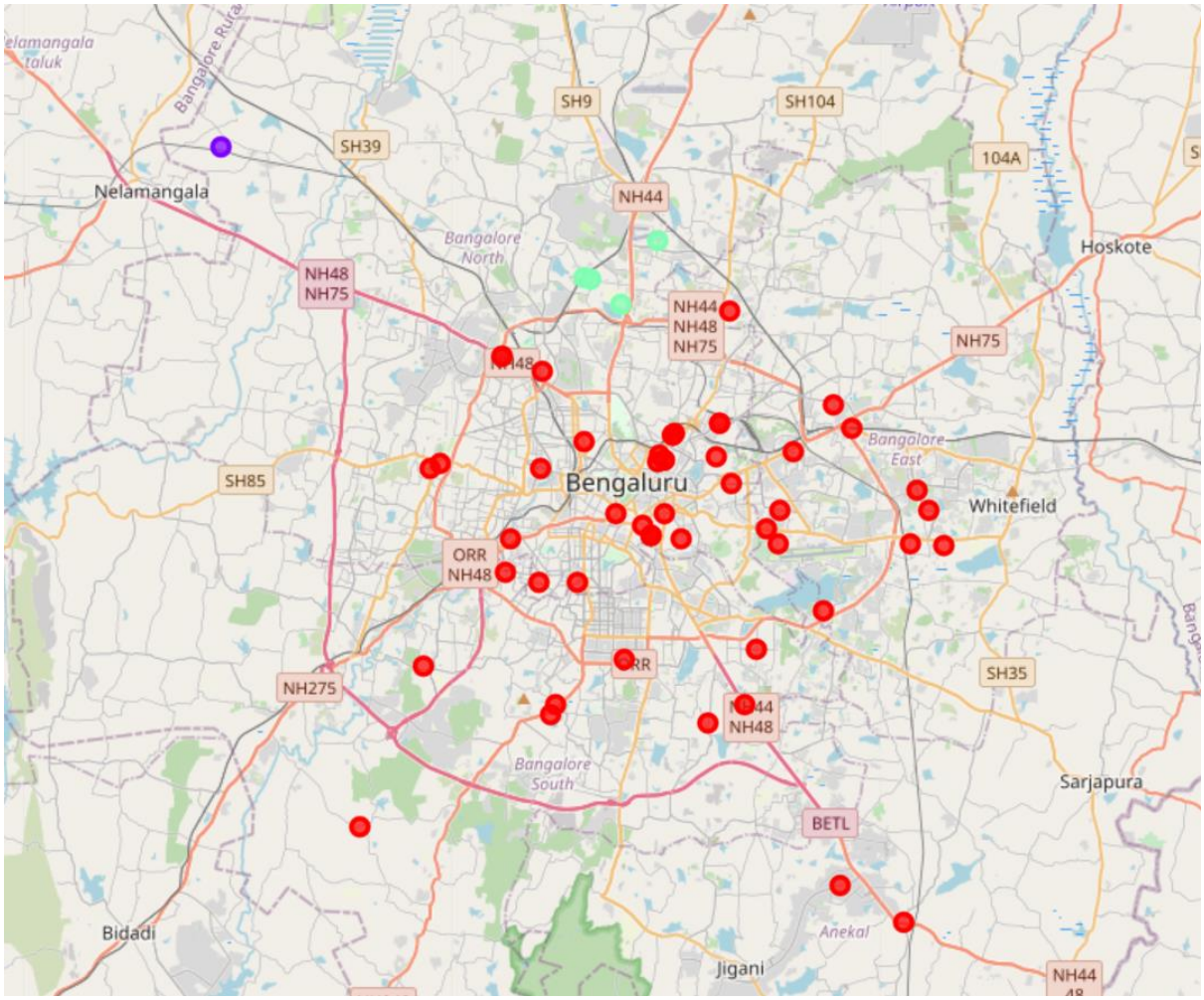

Figure 1: Visuals of the Clustered neighbourhoods in Bengaluru

Discussion

As observations noted from the visual in the results section, most of the bus stations are concentrated in the central area of Bengaluru city, with highest number in cluster 0 and moderate in cluster 2. The cluster 1 has only 1 bus station. This represents a great opportunity and high potential area to open bus stations as it may help us in connecting the city neighbourhoods. Therefore, this project recommends the city mayors to open up new bus

stations in cluster 1 and 2. Lastly, it is recommended from the project to avoid neighbourhoods in cluster 0 which has the most bus stations.

This project considered only the frequency of occurrence of bus stations, there are other factors such as population density in the cluster 1 and cluster 2, land availability, economic indulgence of the people living in the outer parts of the city and whether the people living in cluster 1 and 2 do really require a bus station for commute. It is to be known that bus stations are used for multiple buses. There must be actual need for these bus stations. Further projects could devise a methodology which involves usage of the above unused factors.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e., council of development in the city (BBMP) to open bus stations. To answer the business question raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 and 2 are the most preferred locations to open a new bus station.

References

Category: Suburbs of Bengaluru. Wikipedia. Retrieved from https://commons.wikimedia.org/wiki/Category:Suburbs_of_Bangalore

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs