Coursera Capstone

IBM Applied Data Science Capstone

Clustering Bus Stations in Bengaluru, India

BY: HIMANSHU BASWAL

JUNE 2021

Business Problem

- Location of Bus station is important decisions because that will determine whether it was effective to create it in the location.
- Objective : To analyze and select the best location in the city of Bengaluru, India to open multiple bus station.
- This project is important as the city has expanded its area to 741 km2. This has resulted in need for connecting the area at the border of the city.
- Business Question : In Bengaluru, after the expansion of the city to 741km2, the newer area of the city needs to connect to the mainland, where would you recommend that they set up bus stations?

Data Description

- Data required:
  - List of neighborhoods in Bengaluru
  - Latitude and Longitude of the neighborhoods
  - Venue data, particularly data related to bus stations

- Sources of data:
  - Wikipedia page for neighbourhoods (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Bangalore)
  - Geocoder package for latitude and longitude coordinates
  - Foursquare API for venue data

Methodology

- Web scraping Wikipedia page for neighborhoods list.
- Get latitude and longitude coordinates using Geocoder
- Use Foursquare API to get venue data
- Group data by neighborhood and taking the mean of the frequency of occurrence of each venue category
- Filter venue category by Bus Station
- Perform clustering on the data by using k-means clustering
- Visualize the clusters in a map using Folium

Results

- Categorized the neighborhoods into 3 clusters:

➢ Cluster 0 : Neighborhoods with highest concentration of Bus Stations

➢ Cluster 1 : Neighborhoods with lowest concentration of Bus Stations

➢ Cluster 2 : Neighborhoods with moderate concentration of Bus Stations

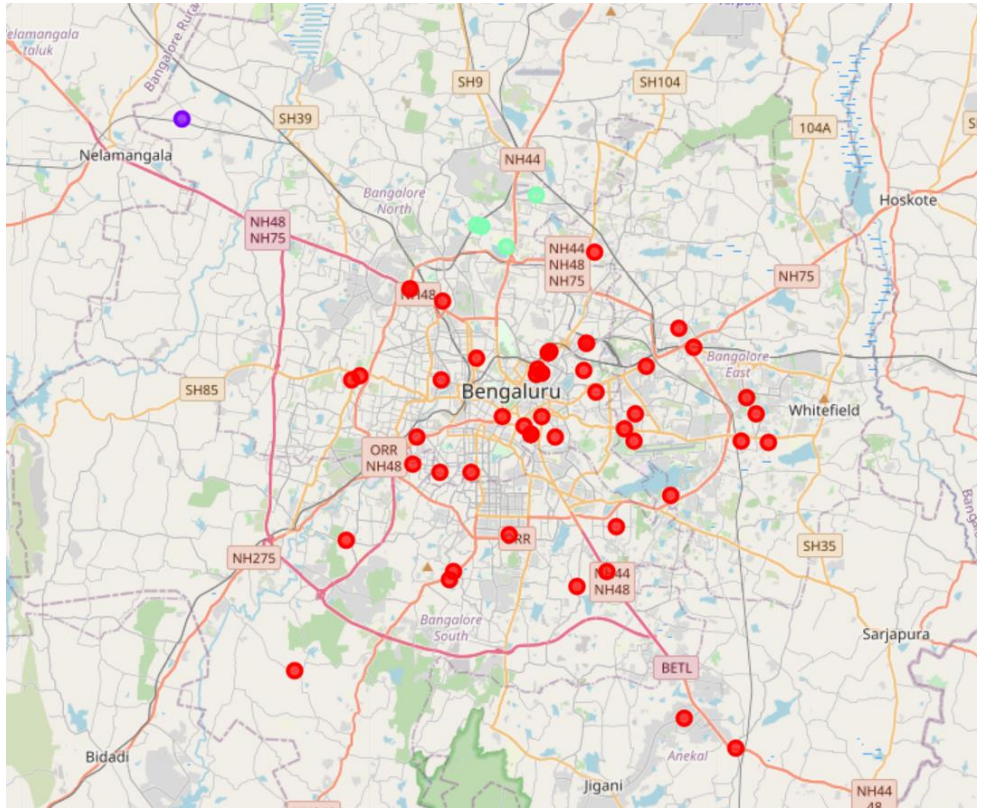

Figure 1: Visuals of the Clustered neighborhoods in Bengaluru

Discussion

- Most of the bus stations are concentrated in the central area of the city.
- Highest number in cluster 0 and moderate number in cluster 2.
- Cluster 1 has very low number to no bus stations in the neighborhood.
- 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.

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.
- Answer to the business question : The neighborhoods in cluster 1 and 2 are the most preferred locations to open a new bus station.