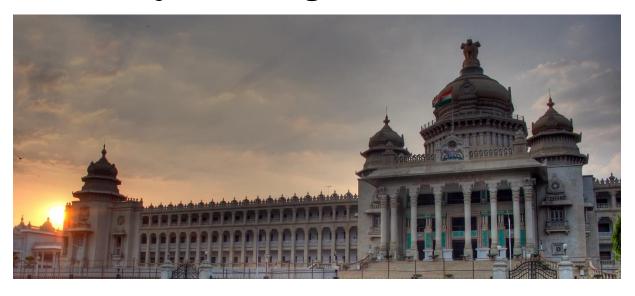# IBM Applied Data Science Capstone

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 database 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.