IBM Applied Data Science Capstone

Development of New Public Parks in Delhi, India

By: Ashish Tiwari

February 2020

Introduction:

For residents of any city, public parks are one of the basic needs. In these parks, people of every age like old age people, youngsters, ladies and children come for their different needs. Old age people come for walk and yoga; Youngsters and ladies for running and exercises; & children for playing their games. Apart from these benefits, Trees and plants in Parks play beneficial role in environment. Delhi is capital of India and since last few years rising pollution in this city is a cause of concern. Government is taking different measures to lower down the pollution. In addition to these measures, planting trees and plants will help the environment and lower down the pollution. Also, if it can be done through developing public parks, then apart from environment it can be so much useful for the residents of the Delhi city. We will analyse the data and figure out the areas in which Government should build the Parks.

Business Problem:

The objective of this capstone project is to analyse and select the best locations in the city of Delhi, India to develop new Public Parks. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Delhi, what are the locations where public parks need to be developed for the benefit of residents and lower down the pollution of the city.

Target Audience of Project:

This project is particularly useful for residents of the city Delhi which will help them to get a place for yoga, walk, exercise, and play. Also, this project is useful for the Government also, as the pollution level is going very high year by year in the city and development of parks, plantation of trees will help in lowering down the pollution level to very much extent.

Data:

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

• List of neighbourhoods in Delhi. This defines the scope of this project, which is confined to the city of Delhi, capital of India.

• Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and to get the venue data.

• Venue data, particularly data related to Parks. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them:

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi) contains a list of neighbourhoods in Delhi. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use 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 many categories of the venue data, we are particularly interested in the Parks category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.