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.

Methodology:

Firstly, we need to get the list of neighbourhoods in the city of Delhi, India. The list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. We also 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 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 pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Delhi.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 20000 meters as Delhi is a big city. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and 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 the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the 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 “Parks” data, we will filter the “Park” as venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm 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 for “Park”. The results will allow us to identify which neighbourhoods have higher concentration of Parks and which neighbourhoods have fewer number of Parks. Based on the occurrence of Parks in different neighbourhoods, it will help us to answer the question as to which neighbourhoods Government should build the Parks.

.

Results:

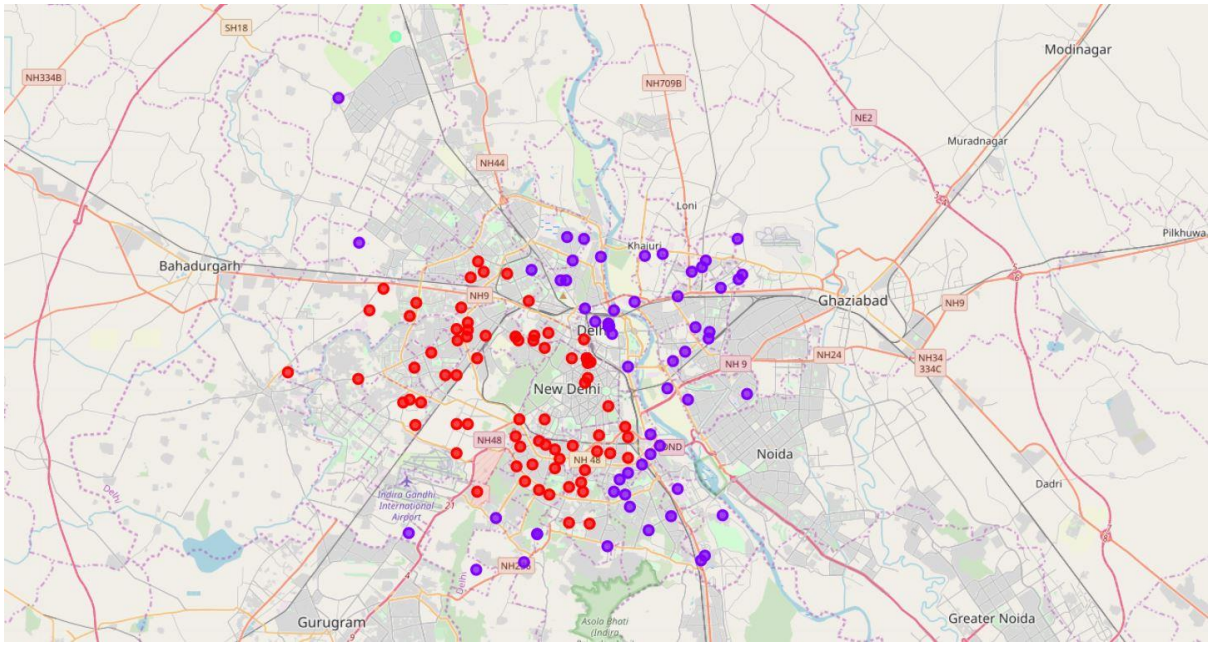
The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Park”:

• Cluster 0: Neighbourhoods with High number of Parks

• Cluster 1: Neighbourhoods with moderate number to no existence of Parks

• Cluster 2: Neighbourhoods with low concentration of Parks

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



Discussion:

As observations noted from the map in the Results section, most of the parks are concentrated in the south, west and centre area of Delhi, with the highest number in cluster 0 and moderate number in cluster 1. On the other hand, cluster 2 has very low number to no parks which includes rural, congested and outer areas from the city. This represents need of building Public parks in these neighbourhoods. Also, there are need to develop more parks in Cluster2 which will help in clean environment. This way Government can tackle the rising pollution level and it will also help the citizens for their daily workouts, play and jogging areas.

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. Government and citizen regarding the locations to develop Public Parks. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 and 1 are the most preferred locations to open new Parks. The findings of this project will help the relevant stakeholders to capitalize on the opportunities to build parks and give a safer, non-polluted environment to its citizens.