Analysing the impact of Parks on the Pollution Levels of Delhi, India

IBM Applied Data Science Capstone Course

By :-

Harshit Agarwal

9th April, 2020

1. Introduction

1.1 Background

New Delhi is the capital of India. Capital of any country reflects the country itself . Delhi , officially the National Capital Territory of Delhi (NCT), is a city and a union territory of India containing New Delhi, the capital of India. It is bordered by the state of Haryana on three sides and by Uttar Pradesh to the east. The NCT covers an area of 1,484 square kilometres (573 sq. mi). According to the 2011 census, Delhi's city proper population was over 11 million, the second-highest in India after Mumbai, while the whole NCT's population was about 16.8 million. As of 2016, recent estimates of the metro economy of its urban area have ranked Delhi either the most or second-most productive metro area of India. Delhi is the second-wealthiest city in India after Mumbai and is home to 18 billionaires and 23,000 millionaires. Delhi ranks fifth among the Indian states and union territories in human development index. Delhi has the second-highest GDP per capita in India. And yet the only thing we co-relate with Delhi the most is the Air Pollution. In recent years Delhi is climbing up in the list of most polluted cities and had gained a bad reputation in terms of pollution and air quality.

1.2 Business Problems

The main objective of this project is to analyse the amount or frequencies of Parks in different regions of Delhi. Delhi being one of the most polluted city, has many reasons for its pollution level to rise. Industries, emission of Carbon mono oxides from vehicles as well as Badarpur Thermal Power Station, overpopulation, and specially increased levels of pollution in the time of Diwali Festival when excess of firecrackers are burnt. All these contribute to the pollution that reaches up to severe conditions at sometimes. To curb all these issues an alternative solution is to plant trees and create parks full of trees. *So the idea here is to analyse and check which regions in Delhi has good frequencies of parks, where not, and then comparing the pollution levels in recent times of those areas*. It would help us to study the impact of Parks on the pollution level in Delhi. *Using this study we can create more parks in the regions where there is very low number of parks*. This would definitely help to curb pollution to some extent.

1.3 Targeted Audiences

The government of India, the state government of Delhi, the NGO's which work for planting more and more trees, or NGO's which work to curb pollution will be highly interested in this project. As the citizens of Delhi are the one's who suffer a lot due to pollution, they might also be interested in this project.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

- List of Neighbourhoods in Delhi, India. This data is provided from Wikipedia site: https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi . It contains all the regions that comprises inside Delhi which sums up to 187 in number. This is the scope of the project all the data is relevant to these regions only.

- The Latitude and Longitude data of the neighbourhood in Delhi. The geocoder library of python is used in order to fetch the coordinates of neighbourhood in Delhi.

- The nearby Venues to all the Neighbourhood's in Delhi. The Foursquare API is used in order to fetch 100 venues in the radius of 2000m of every neighbourhood in Delhi.

2.2 Data Cleaning

Firstly the scraping of data from Wikipedia site can be done in order to fetch the neighbourhoods data of Delhi. The Python requests and beautifulsoup package is used in order to fulfil the need. These packages can help us to get the neighbourhoods in Delhi. After fetching the neighbourhoods we need to get the coordinates of the neighbourhoods. The geocoder library in python can help to provide the coordinates of each neighbourhoods in Delhi, India. After having the coordinates we can use the Foursquare API to fetch the near by venues of all neighbourhoods. 100 venues can be selected from 2000m of radius. These venues can be merged with the main data and then Park's data can be specifically picked out. After picking the Park's data we can use one hot encoding's and create the frequencies grouped on the basis of the neighbourhoods. Then clustering of data can be done to provide the right clusters to the parks and the regions can be identified with more or less park frequencies.