IBM Data Science Professional Certificate

Course 9 – Applied Data Science Capstone

Final Report

Analysis on availability of Parks in Delhi (the capital city of India)

By - Gaurav Vajpayee

Introduction-

Delhi is a city which is situated in northern part of India . It is basically divided into two parts New Delhi and Old Delhi . New Delhi is the capital city of India . In this project we will be taking about whole Delhi city which includes the area of both new and old Delhi city.

Delhi has got a rich historical heritage. It has always been the center of power for entire India for more than 12 centuries. Through its architecture which includes forts ,palaces, gates, arcs , government buildings and parks we can clearly observe the historical magnificent of this city . The parks in Delhi have always been the place to relax , enjoy time with family and most importantly being close to nature and roots for the people of Delhi . These parks play a significant role in not just beautifying the city but also maintaining the pollution level of the city which unfortunately is very high if we talk about Delhi. Some part of the Delhi has got good number of parks but still there are many areas of the city which needs more parks . This problem is usually not addressed by government or real estate development companies which exploit the land in just making concrete floors but do not consider the suffering of city or ultimately people . So in this project we will be analyzing this problem in a more analytical way.

Business or Environmental Problem-

It is moreover an environmental problem which has link to business as well. Real Estate projects with parks in the areas which have less parks can be an attractive deal for customer. It is important to handle this problem in a more analytical way and the problem statement for this problem would be how we can decide which part of the Delhi has got good number of parks and which are facing the scarcity of them?

Beneficiaries of the Project-

1) This project could be beneficial for government while developing infrastructure .

2) This project could be beneficial for Real Estate Companies making new projects .

3) This project could be beneficial for people of Delhi for raising their issues in front of authorities.

Data Required-

1) List of all the neighborhoods of Delhi city .
2) Latitudinal and Longitudinal data of these neighborhoods .
3) Venue data of these neighborhood particularly availability of parks in these neighborhood.

Data Sources –

1. (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi) From this Wikipedia page we are taking out list of neighborhood of Delhi city. We will be using web scraping technique to extract this list with the help of python package called Beautifulsoup.

2. To get the geographical coordinates of these neighborhood we will be using Python package called Geocoder .

3. To get the venues data related to the neighborhood we will be using Foursquare API . Foursquare is widely used among developers for geospatial data. In venues we will particularly focus on parks present in the neighborhood.

Methodology-

1) Firstly we will be taking out the list of all neighborhoods present in Delhi city from a Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi) by using the technique called web scraping with the help of a Python package called Beautifulsoup. This package helps to extract html data and store it in our program for further usage. This extracted list of neighborhoods is stored in a python datatype of “list”.

2) Then with the help of Python package of Geocoder we get geographical coordinates of these neighborhoods stored in a list.

3) Now we merge the coordinates list with our neighborhood list and plot these neighborhood on map with the help of Python library called Folium .

4) Now we call Foursquare API using our credentials to get top 100 venues data about these neighborhood within 2000m radius of distances. Foursquare is widely used among developers to get geospatial data which is authenticate and updated on the regular basis.

5) With the help of one hot encoding we will convert our venue data into 0 and 1 for easy application of machine learning techniques on this data.

6)Now we take mean of the frequency of occurrence of different venue category. By this we are preparing our data for K-means clustering .

7) Lastly , we will perform clustering on the data by using K-means clustering. This algorithm identifies k number of centroids and then allocates every data point to the nearest cluster whole keeping the centroid as small as possible .It is one of the simplest and popular unsupervised machine learning algorithm and is particularly suited to solve the problem for this project. Here we will be clustering our neighborhoods into three clusters based on their frequency of occurrence of Parks .

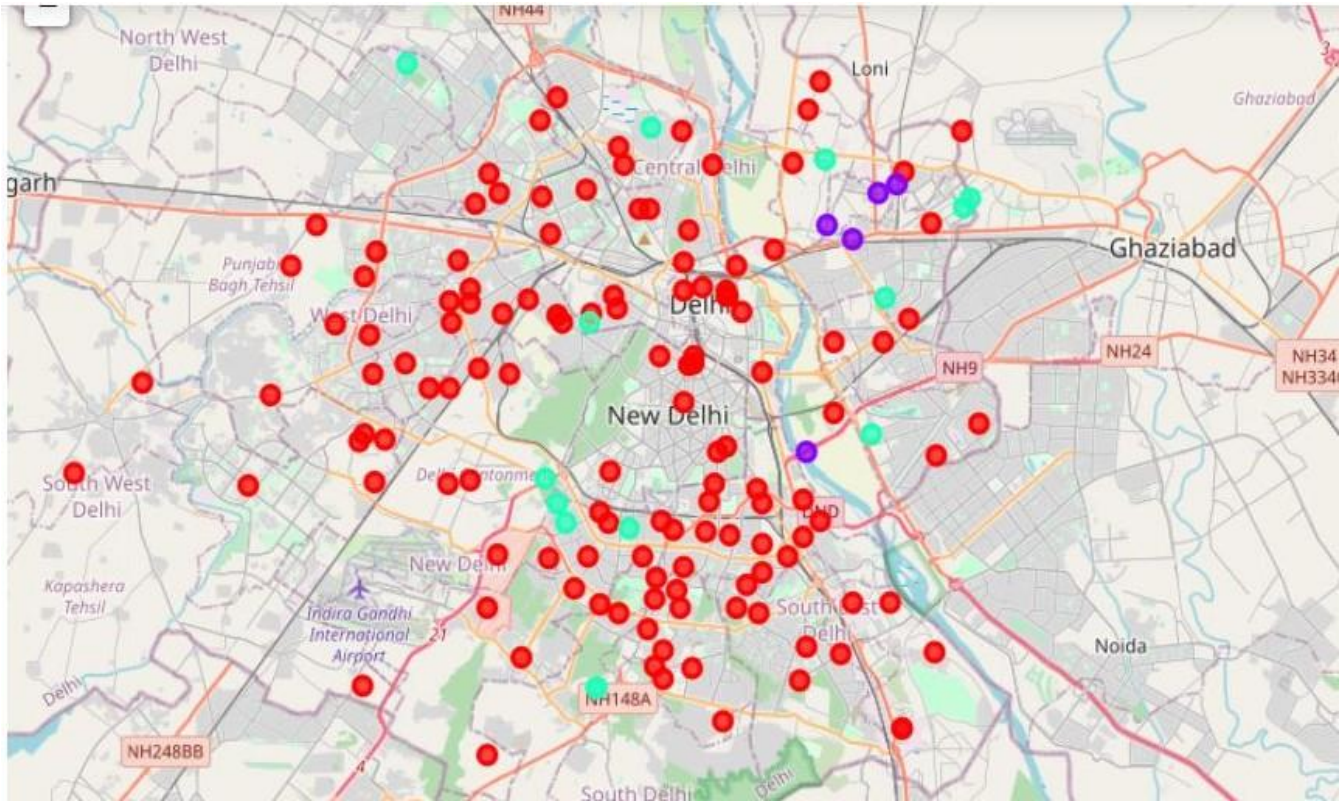
8)By this we will be able to identify the areas having higher number of parks, the areas having low parks and areas have moderate number of parks.

Results –

By clustering the neighborhoods into 3 categories based on the frequency of occurrence of “ Parks” we got following resulting clusters.

- **Cluster 0** - neighborhoods falling in this category are have very less frequency of occurrence of Parks.
- **Clusters 1** - neighborhood falling in this category have high frequency of occurrence of Parks.
- **Clusters2-**neighborhood falling in this category have moderate(between high and low) frequency of occurrence of Parks.

- The map plotting shows results in more visualized representation of different clusters in which purple dots showing cluster category 1 , green dots showing cluster category 2 and red dots showing cluster category 0.



Plot of neighborhood clusters

Conclusions-

By doing this project we came to a conclusion that most of the neighborhoods which have high or good number of parks are located in eastern part of Delhi whereas some neighborhoods of south and east Delhi have moderate number of parks. Whereas neighborhoods with low number of Parks are maximum in number and are located in almost entire area of Delhi but mostly the neighborhoods of western Delhi are facing this problem.

This data analysis can be important to locate the places with less number of parks and can be beneficial for future development purposes of Delhi.

Limitations-

This project has limitations like if talk about the different parameters on which it is tested . Parameters like population , terrain , geographical constraints , financial factors are not included in this analytical study. For more refined analysis including these parameters will be beneficial.

References-

- Category: Neighborhoods of Delhi. Wikipedia . Retrieved from https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi

- Foursquare Developers Documentation . Foursquare . Retrieved from https://developer.foursquare.com/docs

- Famous Parks in Delhi. Retrieved from https://www.treebo.com/blog/famous-parks-in-delhi/

- Delhi Parks and garden. Retrieved from http://www.maverickbird.com/india/north/delhi/delhi-parks-and-gardens-in-spring-can-this-be-the-most-polluted-in-the-world/