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 is done in order to fetch the neighbourhoods data of Delhi. The Python requests and beautifulsoup package is used in order to fulfil the need. A total of 220 items were scraped out of which only 187 were needed. First 12 data and the last 21 items were not relevant and thus removed from the list. Now after cleaning we had the correct number of neighbourhoods. Geocoder provided with the coordinates of neighbourhoods and then the Foursquare API fetched the venues data and nearby the neighbourhoods of Delhi. A total of 7312 data was fetched. A specific to Parks dataset was created which was finally used to cluster and find the appropriate results.

3. Methodology

The first and foremost step is to import and install all the required python libraries.

3.1 The Data Gathering

Gather all the data from the Data Sources we have and place it into the respective data frames. We have scraped data from Wikipedia which provides the neighbourhoods data in Delhi . With the list of neighbourhoods we now need the latitude and longitude of all the neighbourhoods whose list we have maintained. Using Geocoder we maintained all the coordinates and merged them with the neighbourhoods data frame. Now Gathering of Venues data is required, which with the Help of Foursquare API we easily fetched 100 venues within a range of 2000m from our every neighbourhood. Total Venues came out to be 7312 with 222 unique categories (like Parks, Pharmacy, Hotel etc.). Now combining all the data in one data frame , provided a master data file with neighbourhoods, their coordinates and list of venues found out in each location.

3.2 The Analysing of Data

Now we have all the data in place and we can start analysing. But before analysis we can visualize the data we have. Fig 1 shows the plot of neighbourhoods in Delhi on the map using Folium.

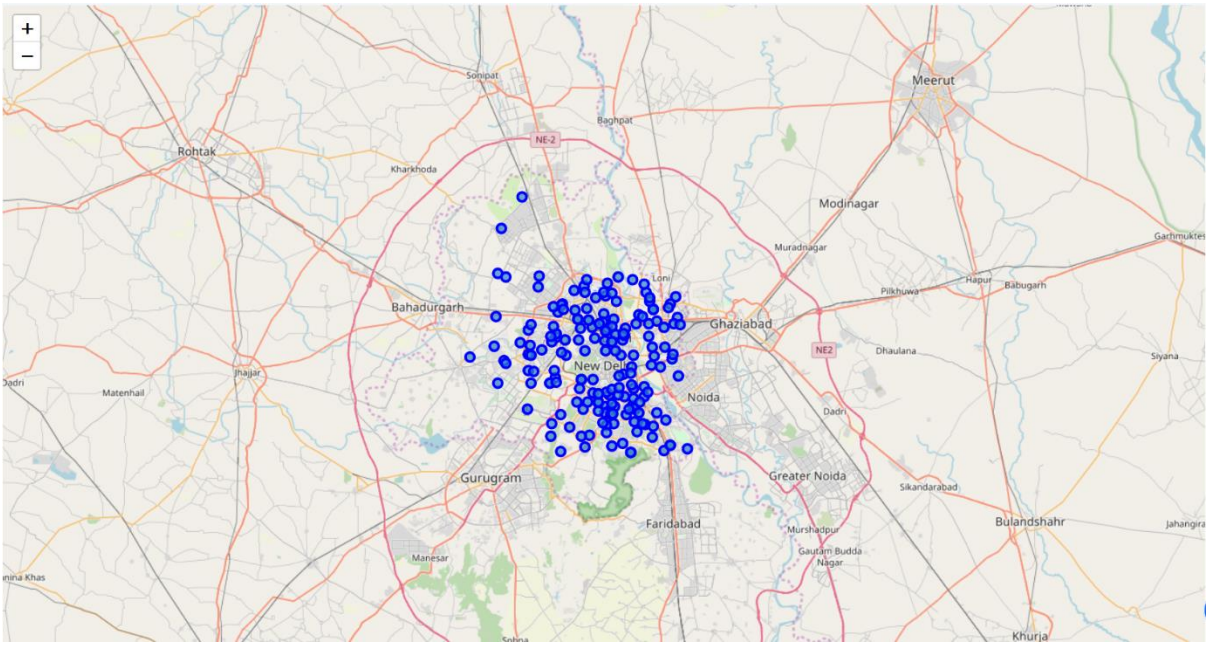

Fig 1: The neighbourhoods in Delhi

Analysis starts with one hot encoding of the data we have. We provide Boolean values to the venues respective to their areas. The value 1 is given to the present venues and 0 to the non-present venues in the neighbourhoods. Further we group them all together on basis of neighbourhoods and apply mean of frequencies of the occurrences of venues categories. This would provide a mean value between 0 to 1 about the every venue category respective to neighbourhoods. More the occurrences if it tends towards 1 and less if it tends toward 0. Now getting specific to the park venue category only, we notice that there are 58 neighbourhoods out of 187 with parks frequency greater than 0. As we need to analyse parks and other data is irrelevant so we create a data frame that's specific to parks data only.

3.3 The Clustering of Data

We are all set to cluster the data and see the outcomes. The data is clustered into 5 sets. K-means clustering algorithm is used to cluster the data into 5 sets. The Cluster labels from 0 to 4 would be applied on the data and we can merge the data altogether to get neighbourhood, parks and cluster label fields in the data frame. We can further merge the coordinates data with our data frame for providing more specific details. Finally when the clustering ends we get the data divided into 5 clusters varied on their frequencies. We can view these clusters on the basis of their labels separately to get full study. Fig 2 shows visualization of the clusters on the map plot.

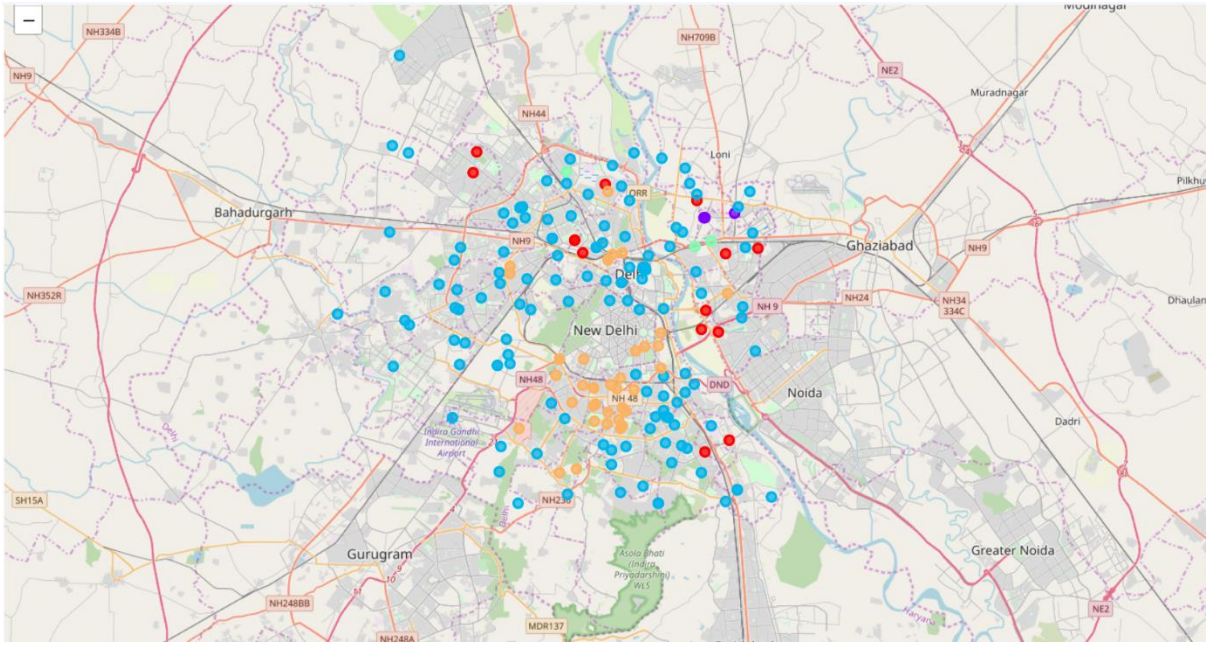

Fig 2 : The 5 clusters with park frequencies in neighbourhoods.

4. Results

The results of the K-means clustering shows that we have our data clustered into 5 sets based on the frequencies of the occurrence of the Parks.

- **Cluster 0** : Consists of average level park frequencies.
- **Cluster 1** : Best cluster as it shows park frequencies from 0.4 – 0.5.
- **Cluster 2** : This is the cluster with vast number of neighbourhoods in it and least level of park frequencies ranging from 0.00 – 0.01.
- **Cluster 3** : Consists of above average park frequencies.
- **Cluster 4** : Consists of below average park frequencies.

As we can see there are Good amount of parks in Cluster with Label 1 which is reflected as Violet colour on Map. With frequencies between 0.4 to 0.5. This is the best Cluster. Then we have Cluster with Label 3 which is Light Green coloured. It shows 2nd best areas in term of Parks. Then we have Clusters with Label 0 with Red Coloured Dots. These are Average level Parks frequencies that one area should have. And then Cluster Label 4 with Skin Colour dots lies below average with frequencies as 0.01 to 0.04.

At last Cluster with Label 2 coloured as Light blue which consists major neighbourhood areas are at the bottom of the parks frequencies list which shows values between 0.00 and 0.01. With maximum 0.00 values it shows why Delhi is one of the highest polluted cities in the world. Observations Clearly states that maximum regions in Delhi are a below average in rating for Parks. Having maintained an average level of parks in all regions would never give rise to High pollution atmosphere in places like capital of India.

Also, according to Economic Times articles from October-November 2019 the highly polluted regions in Delhi were Anand Vihar, Punjabi Bagh, Najafgarh, Chandni Chawk, Ashok Vihar, Mundka, Bawana. All these regions lies in the Cluster 2 from our analysis, which shows least park frequencies in these areas. The link to these articles are:

- https://economictimes.indiatimes.com/news/politics-and-nation/air-quality-slips-into-severe-category-in-several-parts-of-delhi/articleshow/71813979.cms?from=mdr
- https://economictimes.indiatimes.com/news/politics-and-nation/thick-layer-of-smog-envelopes-delhi-light-rains-expected/articleshow/71874950.cms

As October- November is the festival of Diwali and firecrackers go loud in these times, the people experience high pollutions levels other than normal days. But without Parks in these areas the pollution level goes really high as compared to other regions in Delhi.

Figure 3.1, 3.2, 3.3 shows the latest pollution levels in the these regions, Shahdara which lies in the Cluster 1 of our analysis is having Good air quality, whereas Najafgarh and Punjabi Bagh in cluster 2 shows moderate and severe air quality as of now also. Cluster 2 according to the results were having least number of parks, hence more pollution is expected from these regions. ***Hence the actual outcomes is almost equal to the expected outcome, therefore the regions with better park frequencies have good hold on air quality as compared to the least one's.***

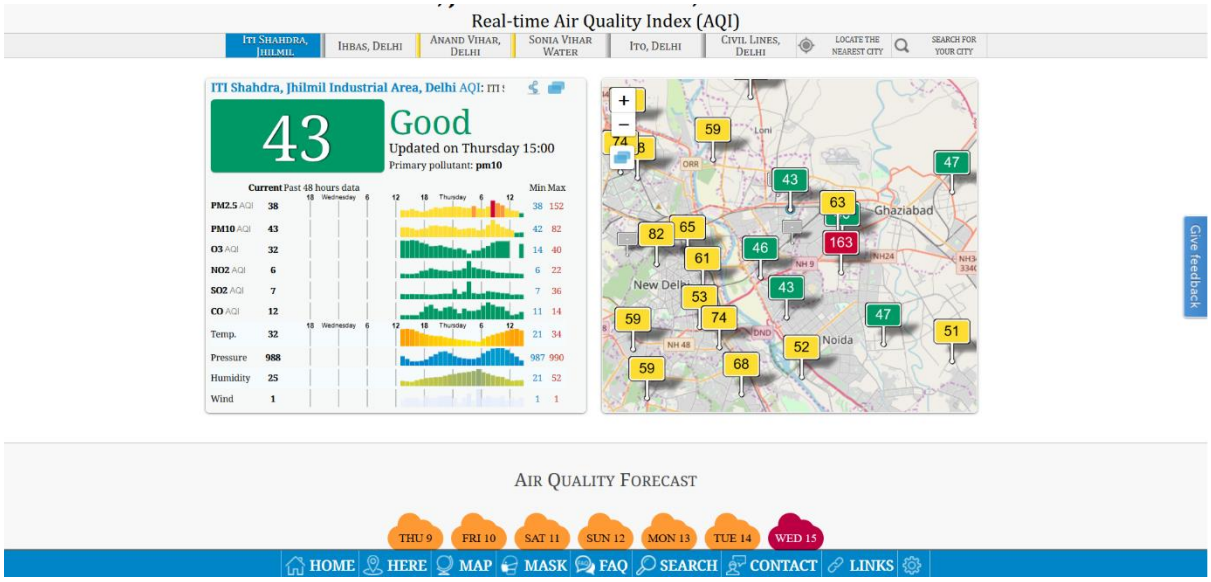

Fig 3.1: Shahdra Air quality Index

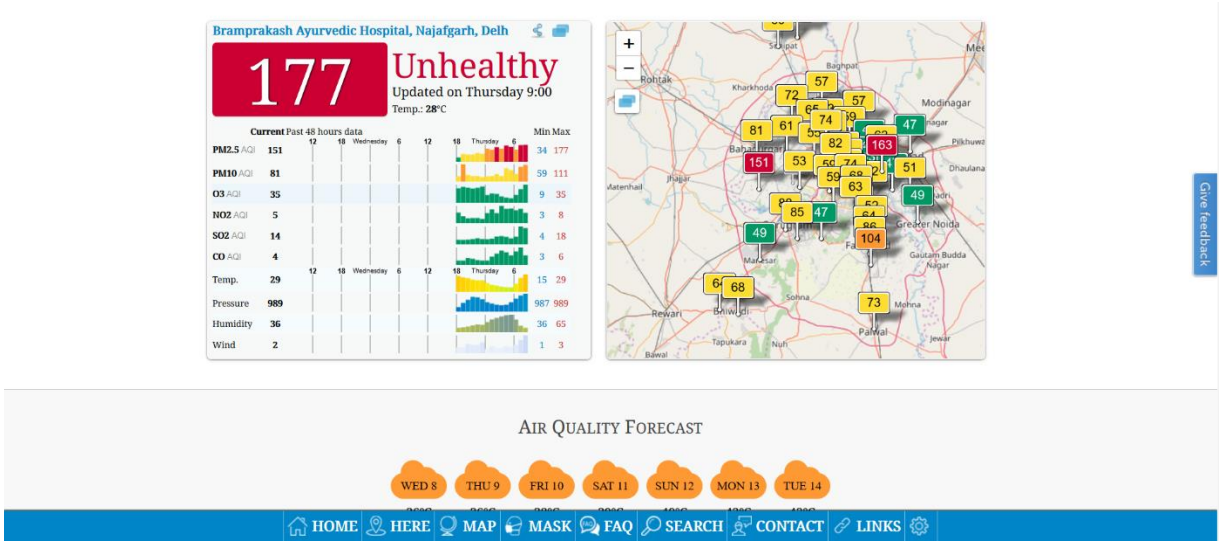

Fig 3.2: Najafgarh Air quality Index

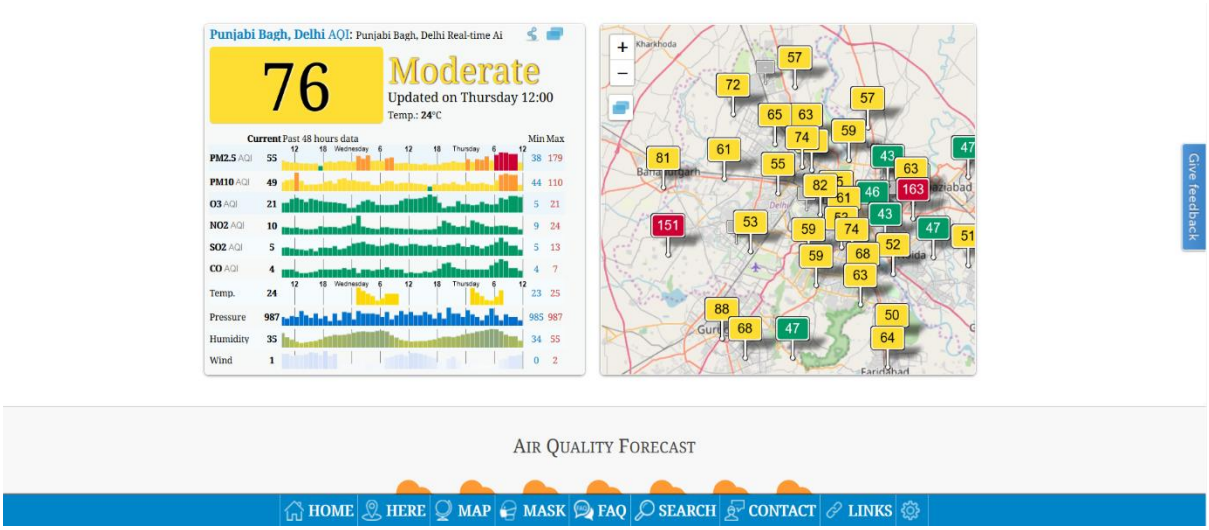

Fig 3.3: Punjabi Bagh Air quality Index

The air quality statistics for Oct-Nov 2019 in Delhi regions can be seen from national site of air quality Delhi (3rd bullet point). The statistics reflect severely affected regions from Delhi were mainly from our analysed cluster 2 where the park frequencies were minimal. Other regions were also highly polluted but they were lower in comparison to them. ***Hence building parks more in these areas would not clear all the pollution single handed but it plays significant role in decreasing the pollution of that particular area. Delhi can be seen as cluster of all these neighbourhoods and if the park frequencies are maintain in all the region, or in maximum of the regions then the pollution level will not rise to such an extent and will be under control.***

The current air quality of Delhi regions are been fetched from :

- https://aqicn.org/city/delhi/
- https://www.iqair.com/india/delhi
- https://app.cpcbccr.com/AQI_India/

5. Discussion

If we create a diagonal symmetrical line passing from the centre of the New Delhi from Northwest to Southeast, we can see Delhi in 2 halves where the upper half is concentrated with more parks and the lower half as very low amount of parks. On the map, the red dots , violet dots, green dots are all on the upper half. Which means, average and above average park frequencies are all located in the upper half of Delhi and on the lower half is the below average and least concentrated park frequencies. So to maintain a balance in the air pollution levels in Delhi as a whole, the organisations can start creating parks from the lower half of the Delhi. Therefore the balance would be created and restored.

6. Future Research

In this project we have only considered parks within a range of 2000m. In future we can check results for more wider range and hence which can provide more better understanding of impact of parks on the pollution of a city. Also, the same procedure can be applied with different neighbourhoods from different cities and compared results can be generated for better understanding. Also we have considered only parks as a single factor, in future studies we can add more factors in the analysis which can contribute to this project for example, number of industries in the neighbourhood areas. So with more factors better results with accurate insights can be generated. This project also used Foursquare API's sandbox account only, which has limitations of 9500 calls per day only. Fetching larger data can be possible if paid accounts with higher privileges are available.

7. Conclusion

The conclusion of the project is, Cluster number 2 with least amount of park frequencies are more vulnerable to severe air pollution. Whereas in comparison to Cluster number 1, Cluster 1 shows better air quality standards. There were 5 sets of clusters which defined the park frequency as best, above average, average, below average, least. Cluster 1 was the best, 3 was above average, 0 was averaged, 4 was below average and cluster 2 had the least amount of park frequencies. This data provides some insights to the NGO's and government of India so as where to create more parks and in which areas the park frequencies are good. Creating more greenery can be started from the lower half of Delhi which has low amount parks in it. Therefore targeted approach towards the goal gives getter outcome always. More the trees, less the pollution!

8. References

- The Neighbourhoods in Delhi:
  https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi

- Foursquare Developers Document:
  https://developer.foursquare.com/docs

- Economic Times Articles :
  https://economictimes.indiatimes.com/news/politics-and-nation/air-quality-slips-into-severe-category-in-several-parts-of-delhi/articleshow/71813979.cms?from=mdr

  https://economictimes.indiatimes.com/news/politics-and-nation/thick-layer-of-smog-envelopes-delhi-light-rains-expected/articleshow/71874950.cms

- The air quality index check:
  https://aqicn.org/city/delhi/

  https://www.iqair.com/india/delhi

  https://app.cpcbccr.com/AQI_India/