Predicting Covid Intensive Zones in Delhi

- By Anand Gangadharan

May 18, 2020

1. Introduction

1.1 Background

Covid-19 is an infectious disease that has caused a havoc in the modern world by disrupting the daily life of every civilian on earth. The virus which primarily affects the lungs, has infected 4,543,390 people globally and 303,711 people have lost their lives because of it (as of 15th May, 2020). The pandemic has forced world leaders to come up with stringent measures such as nation-wide lockdowns to curb the spread of the virus. But lockdowns are holding back civilians from working and earning a living, and its effect does not trouble the poor but every section of the society, even the governments themselves. In such a situation, where economies fall and countries will be heading to recessions, people might be forced to head out and work and live along with the virus. This causes a huge risk to countries like mine, India, with an enormous population density.

1.2 Problem

To foresee the effects of lifting lockdowns, my research and this notebook will help people understand which neighborhoods might see a surge in Covid-19 cases. I will be using population density data and location data of popular venues (like popular market places) to estimate the interaction occurring at a particular place. The neighborhoods will be clustered into groups so that common and effective strategies could be built to handle similar places. I will be focusing on the predictions mainly in my city (Delhi, the capital of India).

1.3 Interest

With this, I hope readers can better understand potentially risky areas and authorities can take measures to place restrictions beforehand to such areas to reduce the spread of the virus and thus the suffering to civilian life.
I hope and pray whoever reads this is safe.

2. Data

To solve the above problem, I have made use of the following data:

- **Neighborhoods of Delhi (and the basis of their segmentation)**
  - https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi
  There are 9 districts in Delhi and the neighborhoods are aggregated into one of these districts.

Please note this data is according to Census 2011 which declares 9 districts as opposed to 11 districts as per the current situation. Therefore, I will be adding places in SOUTH EAST DELHI under SOUTH DELHI and places in SHAHDARA under EAST DELHI.

This data would be used to define the neighborhoods when I will be plotting the results on the map. I used the help of the 'search' option in Google Maps to figure out the approximate coordinates of each neighborhood and created my own dataset.

```
In [6]: districts

Out[6]: ['North West Delhi',
         'North Delhi',
         'North East Delhi',
         'Central Delhi',
         'New Delhi',
         'East Delhi',
         'South Delhi',
         'South West Delhi',
         'West Delhi']
```

Fig1. Districts of Delhi

Out[8]:

	District	Neighborhood	Latitude	Longitude
0	North West Delhi	Adarsh Nagar	28.719390	77.173270
1	North West Delhi	Ashok Vihar	28.687260	77.177689
2	North West Delhi	Azadpur	28.712997	77.177360
3	North West Delhi	Bawana	28.797247	77.048331
4	North West Delhi	Begum Pur	28.726457	77.064246

Fig2. Data frame containing Neighborhood Coordinates

- **Location data of popular venues - provided by Foursquare API**
  The Foursquare API provides data related to the frequency of occurrence of different venues at a particular place.
  I will be using the location data to identify clusters where proximity of venues with high footfall (like markets) are present.
  These will be identified as hotspot heighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adarsh Nagar	28.71939	77.173270	My Idea Store	28.717487	77.170922	Mobile Phone Shop
1	Adarsh Nagar	28.71939	77.173270	Pahalwan Vaishno Dhaba	28.715881	77.173565	North Indian Restaurant
2	Adarsh Nagar	28.71939	77.173270	Adarsh Nagar Metro Station	28.716598	77.170436	Light Rail Station
3	Adarsh Nagar	28.71939	77.173270	Vishyavidyalaya Metro Station@Entry gate #1 n ...	28.715596	77.170981	Train Station
4	Ashok Vihar	28.68726	77.177689	Mirch te Masala	28.687374	77.173744	Mughlai Restaurant

Fig3. Data frame containing venue data

It is noted that Foursquare API doesn't have much details regarding venues in Delhi, therefore this is just an approximate picture of the real world. But we can surely say that it is a good approximation and will serve our need.

- **Population density of various districts - Census 2011 data**
  - https://www.census2011.co.in/census/state/districtlist/delhi.html
  The population density data would be used to mark districts with potentially high risk of community transmission due to the high proximity in living conditions.

	#	Density	District	Increase	Literacy	Population	Sex Ratio	Sub-Districts
0	1	8254	North West Delhi	27.81 %	84.45 %	3,656,539	865	List
1	2	11060	South Delhi	20.51 %	86.57 %	2,731,929	862	List
2	3	19563	West Delhi	19.46 %	86.98 %	2,543,243	875	List
3	4	5446	South West Delhi	30.65 %	88.28 %	2,292,958	840	List
4	5	36155	North East Delhi	26.78 %	83.09 %	2,241,624	886	List
5	6	27132	East Delhi	16.79 %	89.31 %	1,709,346	884	List
6	7	14557	North Delhi	13.62 %	86.85 %	887,978	869	List
7	(adsbygoogle = window.adsbygoogle || []).push(...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	8	27730	Central Delhi	-9.91 %	85.14 %	582,320	892	List
9	9	4057	New Delhi	-20.72 %	88.34 %	142,004	822	List

Fig4. Data frame after scraping the website

We will drop the row with the google advertisement info. Also, we just want the district-wise population density data so we drop the other columns.

As per our assumption lets add districts SOUTH EAST DELHI and SHAHDARA with density population equal to SOUTH DELHI and EAST DELHI respectively.

	District	Density
0	North West Delhi	8254
1	South Delhi	11060
2	West Delhi	19563
3	South West Delhi	5446
4	North East Delhi	36155
5	East Delhi	27132
6	North Delhi	14557
7	Central Delhi	27730
8	New Delhi	4057
9	South East Delhi	11060
10	Shahdara	27132

Fig5. Final Data frame containing district-wise population density

- **District boundaries data**
  I created a data set of my own having the polygon coordinates of each district in Delhi. I made use of two websites - **http://nominatim.openstreetmap.org/** and **http://polygons.openstreetmap.fr/index.py** to first extract the location of a place of the map and then extract its coordinates respectively.

3. Methodology

To predict the Covid Intensive Zones in Delhi, we shall perform the following steps:

- Find the popular venue categories of each neighborhood
- Form clusters of the neighborhoods according the prevalent venue categories using **KMeans Clustering**.
- Superimpose a layer of chloropleth map of the population densities in each district to rightly predict the severity of the possible situation

Initially, I start off by visualizing the neighborhoods and the venues present in each neighborhood.

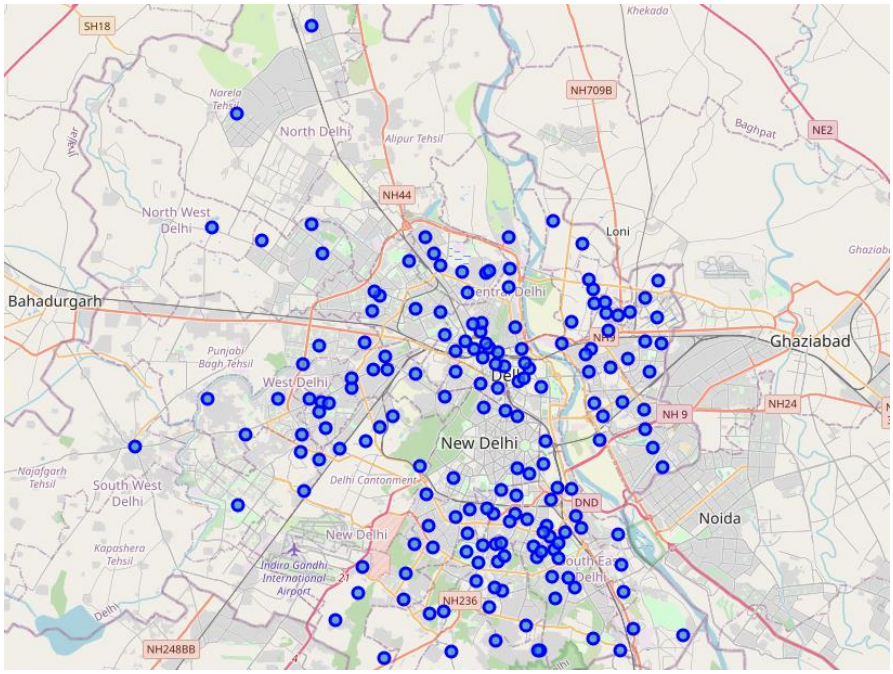

Fig6. Location of these neighborhoods (plotted using folium)

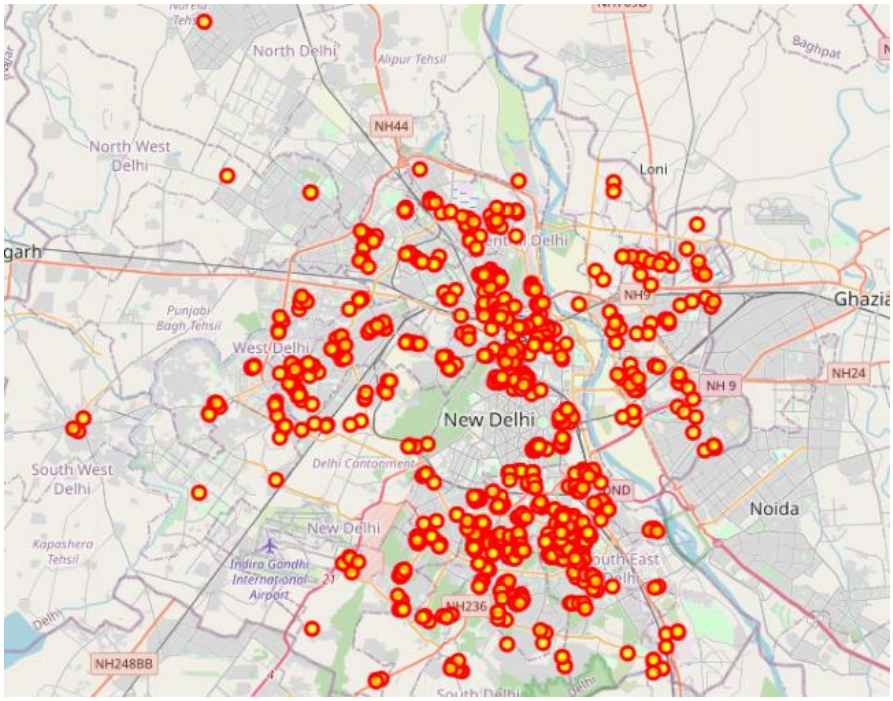

Fig7. Location of these venues (plotted using folium)

Upon grouping the venues according the neighborhoods, we see that only 161 neighborhoods were returned. This means that out of 177 neighborhoods, 16 had no venue results.

Now, I created a table containing the **Top 2** venue categories found in each neighborhood.

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue
0	Adarsh Nagar	Train Station	North Indian Restaurant
1	Alaknanda	BBQ Joint	Food & Drink Shop
2	Ashok Nagar	Fast Food Restaurant	North Indian Restaurant
3	Ashok Vihar	Indian Restaurant	Garden
4	Azadpur	Park	North Indian Restaurant

Fig8. Data Frame containing **Top 2** venue categories in each neighborhood

According to the top venues found, I used **KMeans Clustering** to group the similar neighborhoods together. I made use of **Silhouette Score Comparison** to find the optimum value of **K** (ie the number of clusters).

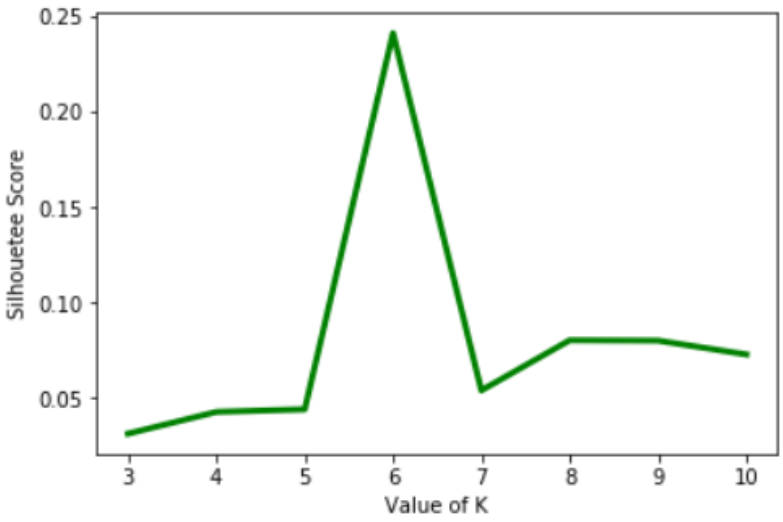

Fig9. Graph comparing Silhouette Score with values of K

The optimum value of K was found out to be 6, following which neighborhoods were clustered into 6 clusters using KMeans Clustering. The result was then merged into a single Data Frame.

	District	Neighborhoods	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
0	North West Delhi	Adarsh Nagar	28.719390	77.173270	2	Train Station	North Indian Restaurant
1	North West Delhi	Ashok Vihar	28.687260	77.177689	3	Indian Restaurant	Garden
2	North West Delhi	Azadpur	28.712997	77.177360	1	Park	North Indian Restaurant
3	North West Delhi	Bawana	28.797247	77.048331	5	Accessories Store	Women's Store
4	North West Delhi	Begum Pur	28.726457	77.064246	5	Accessories Store	Women's Store

Fig10. Data Frame containing neighborhood data and cluster labels

4. Result

Upon creating a Chloropleth map using the district-wise population density data, I arrived at this.

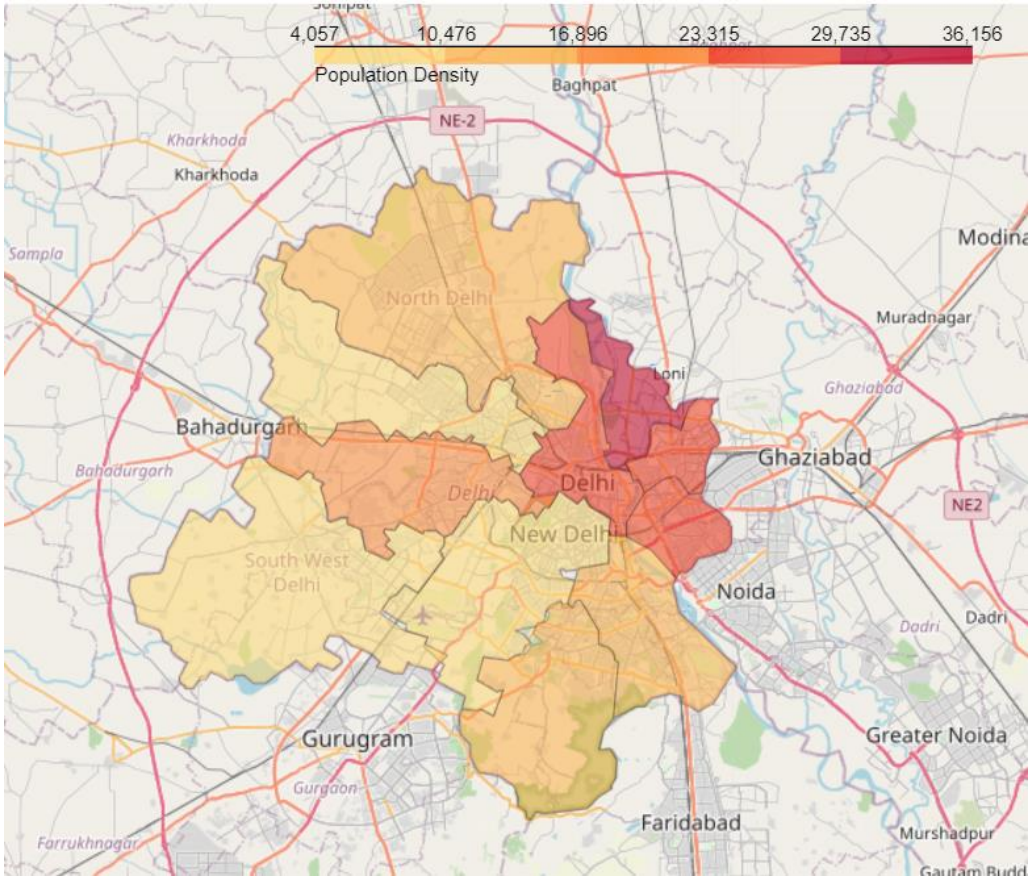

Fig11. Map showing district-wise population density in Delhi

Thereafter, I added the clusters of neighborhoods onto this map so we could know the neighborhood clusters and the degree of concern required for them according to the clustering attribute and population density at that place.

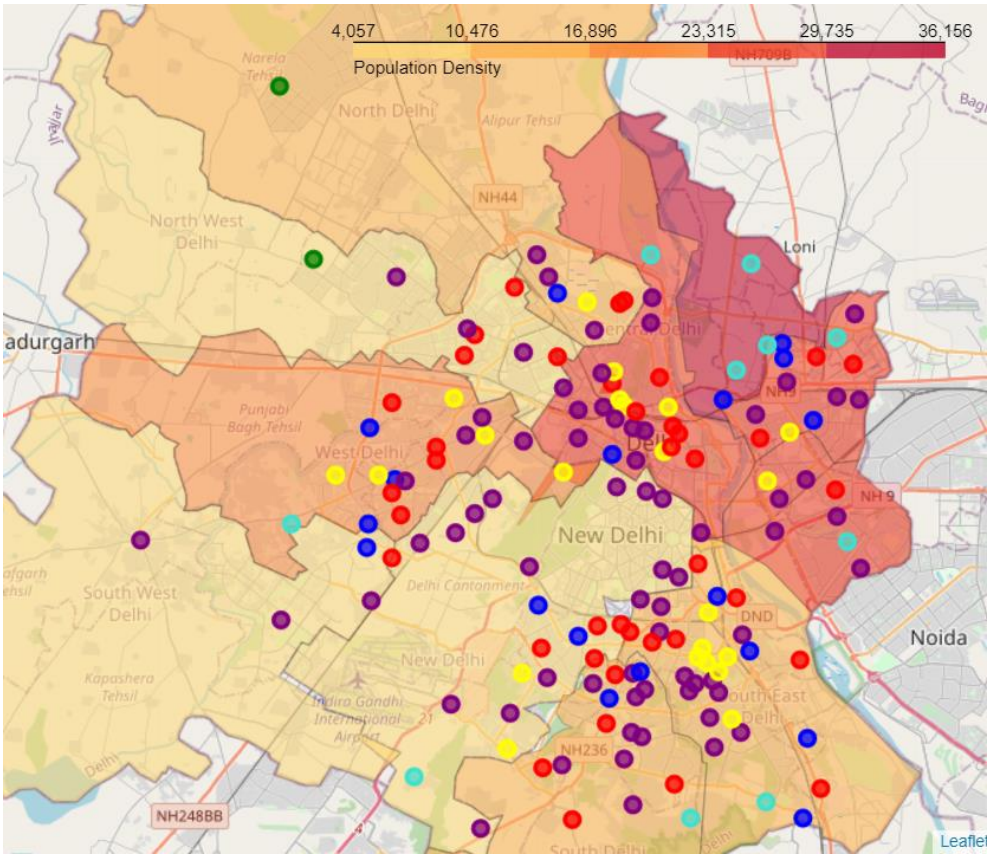

Fig12. Map showing neighborhood clustering

Legend:

- Most Concerning - **RED, YELLOW**
- Moderately Concerning - **PURPLE, BLUE**
- Least Concerning - **TURQUOISE**
- Not Concerning - **GREEN**

Neighborhood clustering attributes were:

1. Neighborhoods predominantly having 'Fast Food Restaurants' and other socially active places. (Most Concerning) - YELLOW MARKER

2. Neighborhoods with housing complexes and normal day-to-day civilian acitivies. (Moderately Concerning) - BLUE MARKER

3. Neighborhoods with many daily life associated venues. (Moderately Concerning) - PURPLE MARKER

4. Neighborhoods with 'Indian Restaurants' predominant. (Most Concerning) - RED MARKER

5. Neighborhoods with venues of essentials only. (Least Concerning) - TURQUOISE MARKER¶

6. Neighborhoods with very few venues only. (Not Concerning) - GREEN MARKER

5. Discussion

With the help of the final map, we can see which all districts are vulnerable to the spread of the virus and also which neighborhoods in the each district needs to be taken special take off.

Places with Red and Yellow markers (Most Concerning zones) especially in districts like North East Delhi, North Delhi and Central can aggravate the situation if timely action is not taken.
South West Delhi and North Delhi being a sparsely populated region should be the least concerning area in Delhi.

The following improvements could be worked upon in future studies:

- As mentioned earlier, Foursquare API doesn't provide very detailed results when used for collecting data about Delhi. Other Location Data providers could be utilized in further studies to result in more accuracy.
- Furthermore, rather than working with coordinates of neighborhoods, if neighborhood boundaries could be utilized and all venues within the boundary is extracted, it would produce near perfect results.
- Census 2011's data was used in this project. The current year is 2020 which we could mean the statistics could be changed a lot. So working with the latest Census, as and when it is released could better the results.

6. Conclusion

Covid-19 has wreaked havoc all around the globe and almost half of the year has been now spent by all the people on earth in just fighting off the pandemic. With better strategies and more efficient systems we should tackle this problem so as to be done with it as soon as possible for not just the health of all the people but also for the falling economies.
I hope this project of mine aids in the formation of mitigation strategies so that we all could win this battle and return to our normal lives at the earliest.

I would like to thank Coursera, IBM and all the faculties involved with the Data Science Professional Certificate Courses for introducing me to Data Science and enabling me with all the skills required to conduct the analysis of the problem at hand and reach my results. Last but not the least, I hope and pray my readers are safe.

Thank you for spending your valuable time in going through my work.