SEGMENTING AND CLUSTERING NEIGHBORHOODS WITH LEVEL OF RISK FROM COVID-19

1 ABSTRACT:

Coronavirus pandemic has taken toll across the world. Delhi (, India) being one of the most densely populated cities in the word is a huge concern. In this project, we will analyse India’s national capital Delhi. Starting with Delhi’s demographic and location data, we will be segmenting and clustering neighbourhood of Delhi with the goal of identifying high-risk neighbourhood.

2 INTRODUCTION:

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing pandemic of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first case of the COVID-19 pandemic in India was reported on 30 January 2020, originating from China. As of 15 May 2020, the Ministry of Health and Family Welfare have confirmed a total of 81,970 cases, 27,920 recoveries (including 1 migration) and 2,649 deaths in the country. On 22 March 2020, India observed a 14-hour voluntary public curfew. The government followed it up with 21 days nationwide lockdowns, which wad further until 17 May, 2020.

At a time when India is likely to end its third phase of Coronavirus lockdown on May 17, there are some points that need to be taken into account while planning the exit strategy (ending the lockdown). Across the world, 14 studies have been conducted with a similar conclusion that public enclosed places like restaurants and workplaces can act as COVID-19 super-spreading environments, the Indian Express reported. The studies were done after tracking index patients and the channels of contagion including religious gatherings, their homes, public transport and workplaces. According to the report, these studies have put enclosed public gatherings and mass living spaces under high-risk environments and suggested that 5-1 per cent of the transmission can be traced to dining, entertainment and transportation. The report citing a study in The Lancet said that COVID-19 is largely transmitted by closed contact especially when the contact is prolonged or in a closed congregation.

Meanwhile in India, density is still a huge concern and it can act as adding fuel to fire. Through this project, we will cluster and segment neighbourhood of India's capital - Delhi. It should give us insights like the area where we are most likely to get COVID-19. The results can be used by the state government to implement policies to curb coronavirus effectively while ending the lockdown. Further, this report can be used to raise awareness and help individuals to avert from high-risk locations.

3 DATA

Multiple sources of data were used considering the problem. The sources were listed below:

- Postal code of Delhi was obtained from geonames. This data would be used divide the city into neighbourhood. (http://download.geonames.org/export/zip/)
- Python Geocoding Toolbox was used to obtain the geo-coordinates of each postal codes. This data will help us analyse the area around neighbourhood.

- District wise population density was obtained from India census 2011. Population density was used as places with higher density could spread coronavirus further. (https://www.census2011.co.in/census/state/districtlist/delhi.html)
- District wise coronavirus containment zone was obtained from Jagran Josh. On top of population density, this will help us locate areas where we have observed more coronavirus cases. (https://www.jagranjosh.com/current-affairs/delhi-coronavirus-hotspots-manish-sisodia-arvind-kejriwal-coronavirus-masks-delhi-1586369305-1)
- Lastly, we use Foursqare API to get venue around each neighbourhood. It will help us identify high-risk places like restaurants and workplaces can act as COVID-19 super-spreading environments.

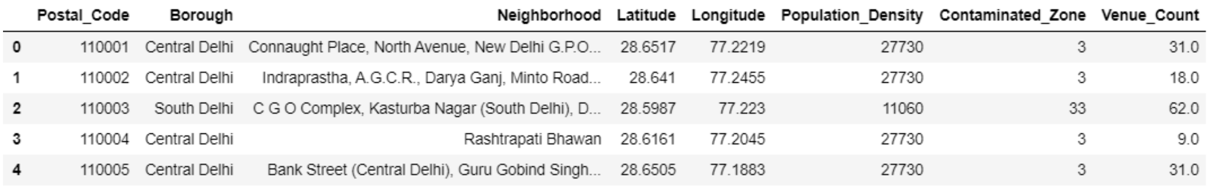
4 METHODOLOGY

4.1 DATA ACQUISITION, CLEANING AND PRE-PROCESSING

The data has been collected through multiple sources as mentioned in the above section. Three methods were mainly used while getting the data:

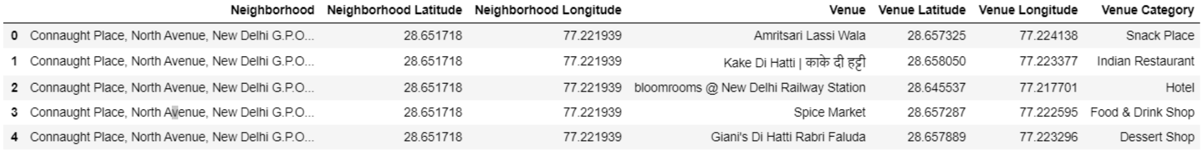
- Foursquare API – the places API offers real-time access to Foursquare's global database of rich venue data and user content to power your location-based experiences in your app or website.
- Web scrapping – I used Beautiful Soup, a Python library for pulling data out of HTML and XML files. It works with your favourite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.
- CSV Database – I used pandas to load a comma-separated values (CSV) file. It is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas.

After retrieving data from different sources, I created a DataFrame using Pandas library in python. Initially the geo-coordinated data obtained from geonames website was inaccurate so I used python's Geocoding Toolbox to get accurate geo-coordinates of postal codes (postal codes were used to divide the city into neighbourhood). A few of the postal codes did not return any geo-coordinates and I had to drop those records from the database. Please note since the last census was carried out in 2011, The population density and district classification were done with respect to 2011. Also, the contaminated zones were up to date during the time of analysis (May 11, 2020). The column Population_Density, Contaminated_Zone and Venue_Count were later pre-processed by the method StandardScaler from sklearn pre-processing library. The snapshot can be seen below.



	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Population_Density	Contaminated_Zone	Venue_Count
0	110001	Central Delhi	Connaught Place, North Avenue, New Delhi G.P.O...	28.6517	77.2219	27730	3	31.0
1	110002	Central Delhi	Indraprastha, A.G.C.R., Darya Ganj, Minto Road...	28.641	77.2455	27730	3	18.0
2	110003	South Delhi	C G O Complex, Kasturba Nagar (South Delhi), D...	28.5987	77.223	11060	33	62.0
3	110004	Central Delhi	Rashtrapati Bhawan	28.6161	77.2045	27730	3	9.0
4	110005	Central Delhi	Bank Street (Central Delhi), Guru Gobind Singh...	28.6505	77.1883	27730	3	31.0

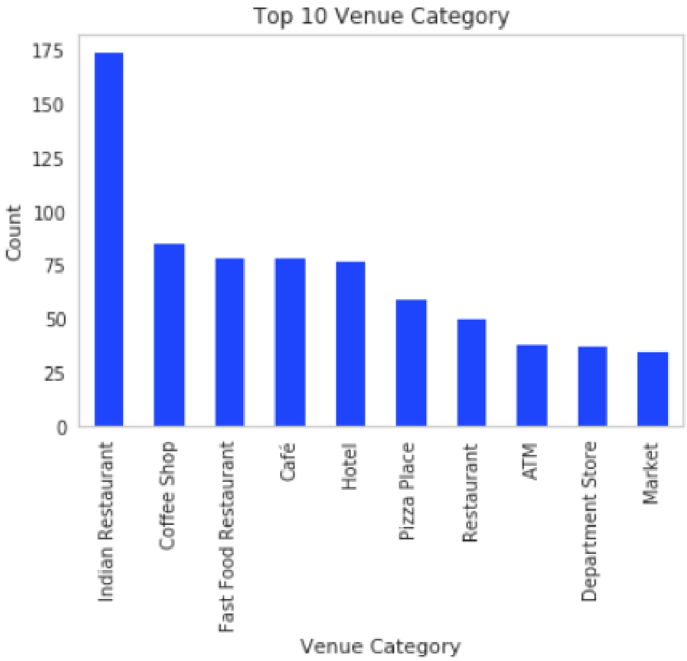
Next, we use Foursquare API to read the locations from above table and return all venue under the radius of 1000 meters in each neighbourhood. In the below image we can see few venues returned for the 1st Neighbourhood "Connaught Place,…"



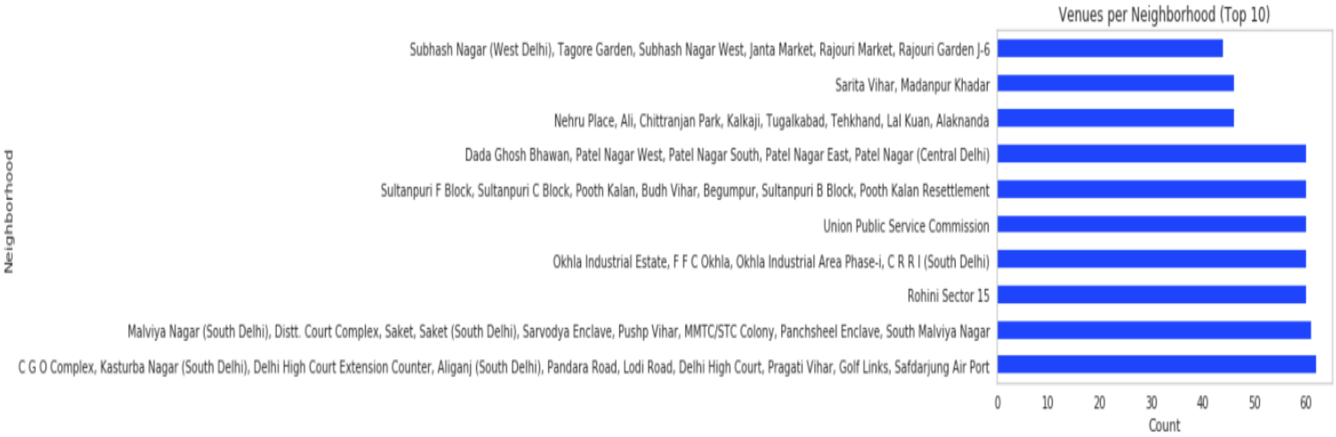
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Connaught Place, North Avenue, New Delhi G.P.O...	28.651718	77.221939	Amritsari Lassi Wala	28.657325	77.224138	Snack Place
1	Connaught Place, North Avenue, New Delhi G.P.O...	28.651718	77.221939	Kake Di Hatti | काके दी हट्टी	28.658050	77.223377	Indian Restaurant
2	Connaught Place, North Avenue, New Delhi G.P.O...	28.651718	77.221939	bloomrooms @ New Delhi Railway Station	28.645537	77.217701	Hotel
3	Connaught Place, North Avenue, New Delhi G.P.O...	28.651718	77.221939	Spice Market	28.657287	77.222595	Food & Drink Shop
4	Connaught Place, North Avenue, New Delhi G.P.O...	28.651718	77.221939	Giani's Di Hatti Rabri Faluda	28.657889	77.223296	Dessert Shop

4.2 Exploratory Analysis

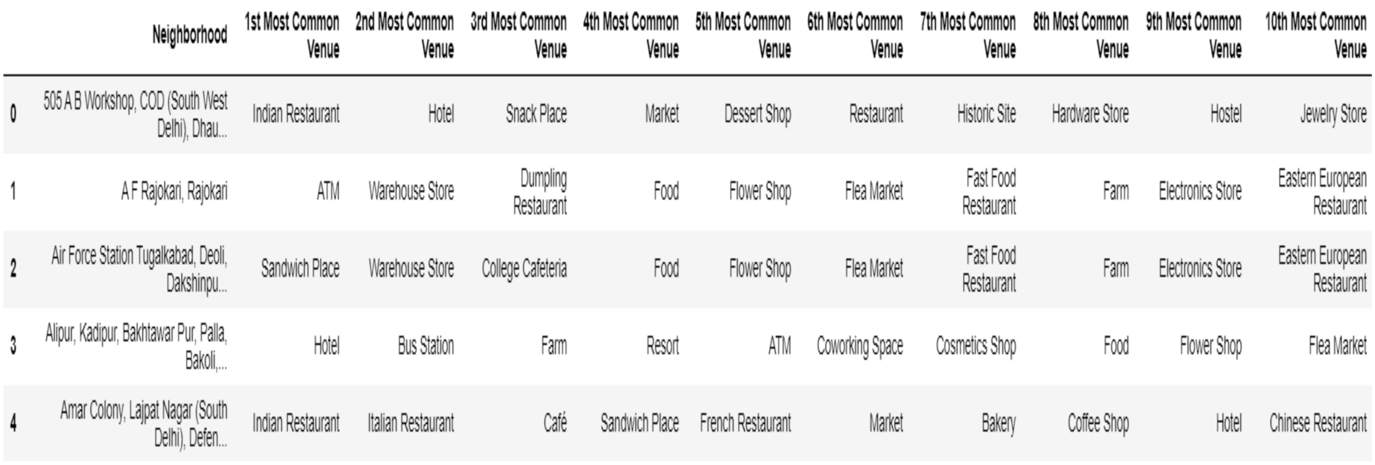
Foursquare API explored 90 neighbourhood (postal codes) and returned almost 1500 different venues across 160 categories (For example, Indian restaurant, Clothing store etc.).



In the graph above, we can see that the most common venues were restaurants, coffee shops/cafe and hotels. Also, in the graph below, we can see the top 10 neighbourhood that returned highest number of venues.



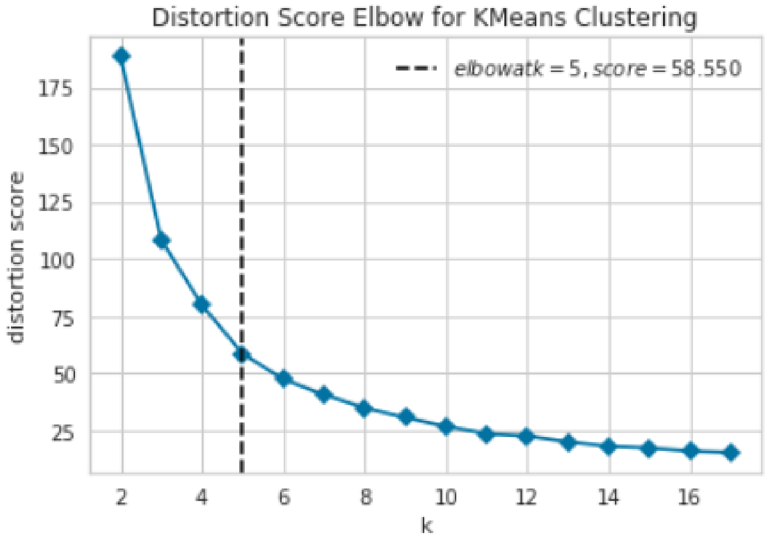
Further, now we can join the venue category and neighbourhood data columns to get top 10 venues in each neighbourhood (as shown in the table below).



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	505 A B Workshop, COD (South West Delhi), Dhau...	Indian Restaurant	Hotel	Snack Place	Market	Dessert Shop	Restaurant	Historic Site	Hardware Store	Hostel	Jewelry Store
1	A F Rajokari, Rajokari	ATM	Warehouse Store	Dumpling Restaurant	Food	Flower Shop	Flea Market	Fast Food Restaurant	Farm	Electronics Store	Eastern European Restaurant
2	Air Force Station Tugalkabad, Deoli, Dakshinpu...	Sandwich Place	Warehouse Store	College Cafeteria	Food	Flower Shop	Flea Market	Fast Food Restaurant	Farm	Electronics Store	Eastern European Restaurant
3	Alipur, Kadipur, Bakhtawar Pur, Palla, Bakoli,...	Hotel	Bus Station	Farm	Resort	ATM	Coworking Space	Cosmetics Shop	Food	Flower Shop	Flea Market
4	Amar Colony, Lajpat Nagar (South Delhi), Defen...	Indian Restaurant	Italian Restaurant	Café	Sandwich Place	French Restaurant	Market	Bakery	Coffee Shop	Hotel	Chinese Restaurant

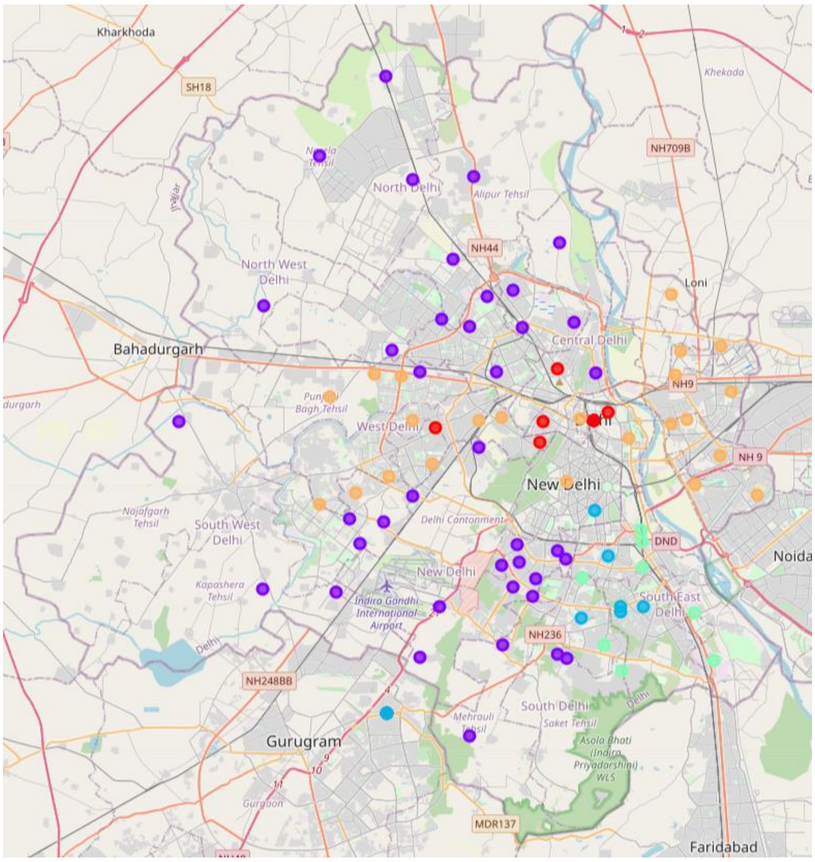
4.3 Modelling

The aim of our project is to identify neighbourhood having high risk of coronavirus spreading. I will be using K-Means clustering algorithm – a type of unsupervised machine learning. I am using clustering since we have to segment similar neighbourhoods having high risk of coronavirus. Features like population density, contaminated zones, number of venues and venue category have been used to identify high risk areas. Finally, we use the elbow method to find the most optimum value of "K". In our case, it comes out to be 5.

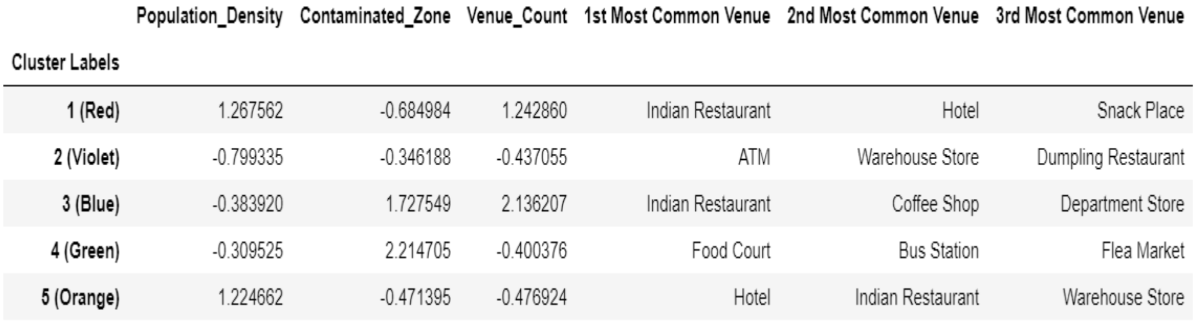


5 Results

We have segmented each neighbourhood to a cluster based on feature similarity like population density, venue category etc. I have used Folium library to visualize the clusters as seen in image below.



Now that we have clustered, lets see the properties of each cluster.



	Population_Density	Contaminated_Zone	Venue_Count	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Cluster Labels						
1 (Red)	1.267562	-0.684984	1.242860	Indian Restaurant	Hotel	Snack Place
2 (Violet)	-0.799335	-0.346188	-0.437055	ATM	Warehouse Store	Dumpling Restaurant
3 (Blue)	-0.383920	1.727549	2.136207	Indian Restaurant	Coffee Shop	Department Store
4 (Green)	-0.309525	2.214705	-0.400376	Food Court	Bus Station	Flea Market
5 (Orange)	1.224662	-0.471395	-0.476924	Hotel	Indian Restaurant	Warehouse Store

6 DISCUSSION

After successfully carrying out clustering using K-Means. We can see that neighbourhood in cluster 1 and cluster 3 can be considered as high-risk areas as these areas mostly contain places are densely occupied with restaurants/hotels and have either high population density or contaminated zones. Neighbourhood in cluster 2 seems to be the safest one as it has low value in all the metrics and also only contains less risky venues. The remaining neighbourhood falls in the cluster 4 and cluster 5. These clusters can be considered as areas of medium risk as it only has high value in one of the metrics.

7 CONCLUSION

We successfully used data points like neighbourhood venues, population density and contaminated zones to identify and cluster areas with high-risk of spreading coronavirus. Our model can be further extended to carry out the same analysis for more cities – leading to analysis for the complete country. The results of the analysis can be used. by the state government to implement policies to curb coronavirus effectively while ending the lockdown.