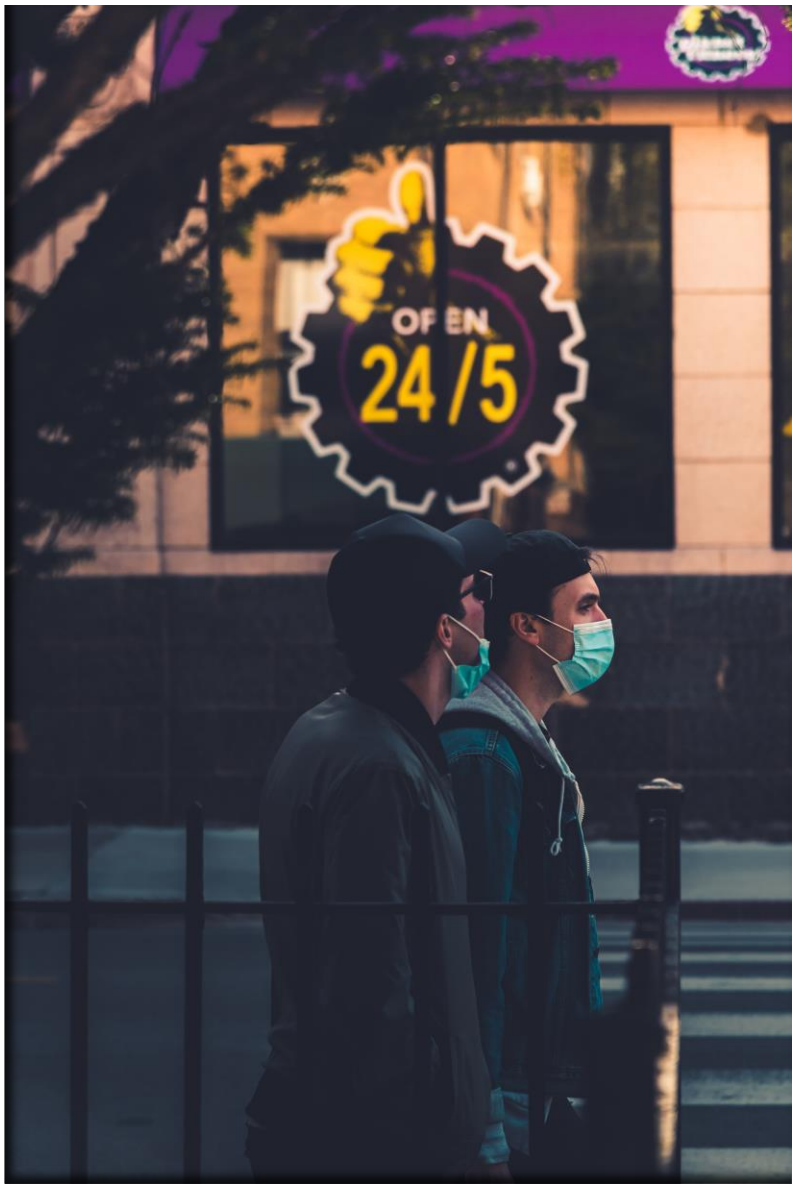


IBM DATA SCIENCE CAPSTONE PROJECT

COVID-19 TESTING CENTRES

MEHUL SAMPATH | IBM DATA SCIENCE PROFESSIONAL CERTIFICATE | July 1, 2020

INTRODUCTION

The global pandemic that began in the month of January has crippled the way the world functions. Most countries have come to a standstill and with months into it, this is being called the new normal. However, with life in many parts of the world returning to normalcy with the opening of public places, the risks are still profound and precautions at cost are extremely necessary.

This project is being carried out with an aim to ensure precautions and not let a loose stance with the lifting of lockdowns. Let's take a look at what the problem is and what's the proposed solution.

Problem Statement

As public spaces open up and people come out to have a good time after months of being trapped, popular venues and destinations are going to get crowded. This also opens up the possibility of another out break emerging from such hotspots.

In regards to this, a safe prospect would be to set up Covid-19 testing centers in the proximity of these venues and trending places to conduct randomize testing to ensure regularly ensure that the place is not turning into a Covid-19 hotspot.

By helping identify the perfect places to set up testing centers using Machine Learning is one of the most efficient ways to go about the process and will result in precise decision making.

TARGET AUDIENCE/STAKEHOLDERS

Th target audience is the medical lobbies and associations that approve the setting up of test centers. It will also involve private testing labs to give in their input as to where the best places of setup could be.

In case of remote testing centers in form of moving vehicles, popular places in and around a particular radius can be covered in a single day. Since this project is focused in the city of Chennai, India, it will mostly be concerned with authorities in this region.

DATA SOURCES

We will use the following data to come to come up with the best suggestions:

- List of neighborhoods in Chennai, Tamil Nadu and to be more specific, we can extract the neighborhoods in the western part of this large city. We will extract data from this Wikipedia page - https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai
- Geographical coordinates or latitudes and longitudes of the specific locations to plot the map for a visual outlook.
- Data of the popular places and venues in the western part of Chennai.
- We will also consider some commonly known facts about the Chennai to make the best possible model.

After cleaning up the data, we will use K-Means to create clusters of the neighborhoods and then we will figure out which cluster will need a testing center and what would be the most ideal spot.

DATA CLEANING

Scraping data from Wikipedia is not the most desirable thing but thankfully, with beautiful soup the process becomes very pleasing and then the desired list of areas is then converted to a neat little list. Out of the total areas, since our interest lies in just concentrating in one region of the large city, that is, West Chennai, we will only maintain the list of the required areas and get rid of the remaining area names and convert it into a neat little data frame.

The Data Cleaning process is quite standard and most of the work is actually taken care of by beautiful soup and we only need to focus on materialistic parameters like what needs to stay and what needs to go as per our understanding and basic knowledge.

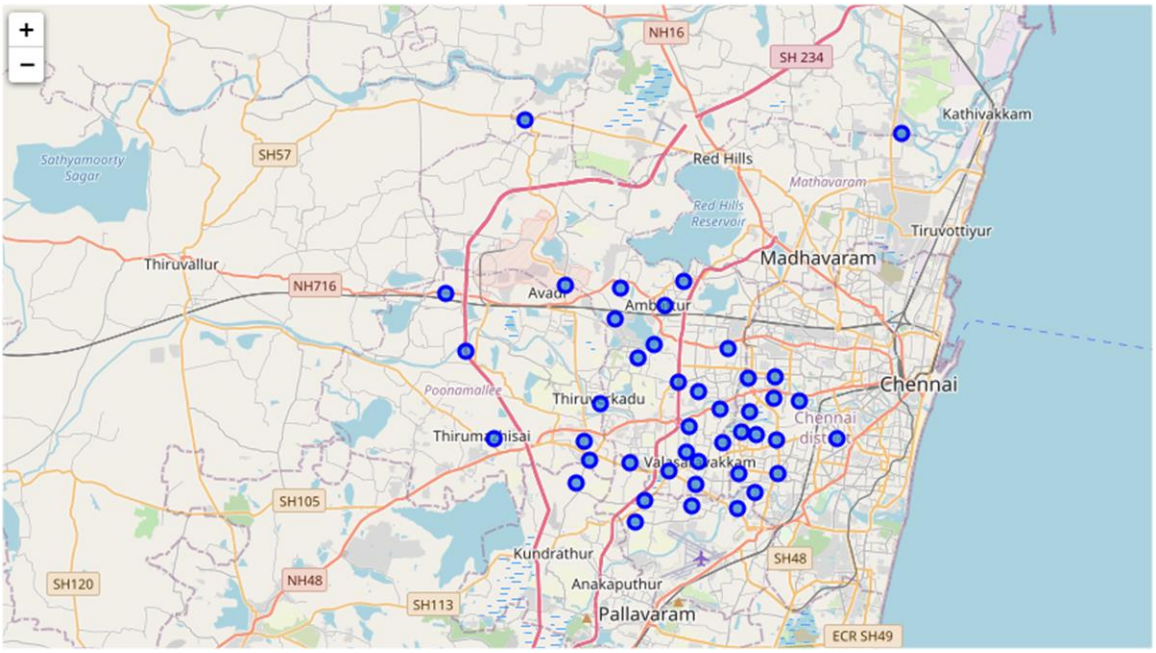
Data cleaning is a constant process and will be carried out in further steps as well as we import more data from various APIs.

Merging Data into one consolidated Data Frame

We also need to add the co-ordinates of the areas to our data frame in order to better visualize our data and also to be able to use FourSquare API to extract more relevant data

in terms of the popular venues in the area. So, we then add the latitude and longitude columns to our existing data frame.

We then plot the map using Folium to get a sense of the geographical orientation of our areas and a generic overview of the changes that will be incorporated in the procedures to follow.



FOURSQUARE API INTEGRATION

The next step in the methodology is to use the FourSquare API and explore the common places around the given coordinates or in the given areas. This will give us a sense of which areas are more popular among people for spending a day out once the lockdowns are lifted.

We set a maximum threshold of 100 venues in an area. In essence, if there are 100 or more venues or places of interest in an area, it would basically become a prime area of importance for setting up a new testing center.

We then develop a data frame with the types of places in the area. None of the fields are redundant because any type of public place is a potential crowd gatherer and should be treated as so to accommodate and be prepared for the worst-case scenario of an outbreak.

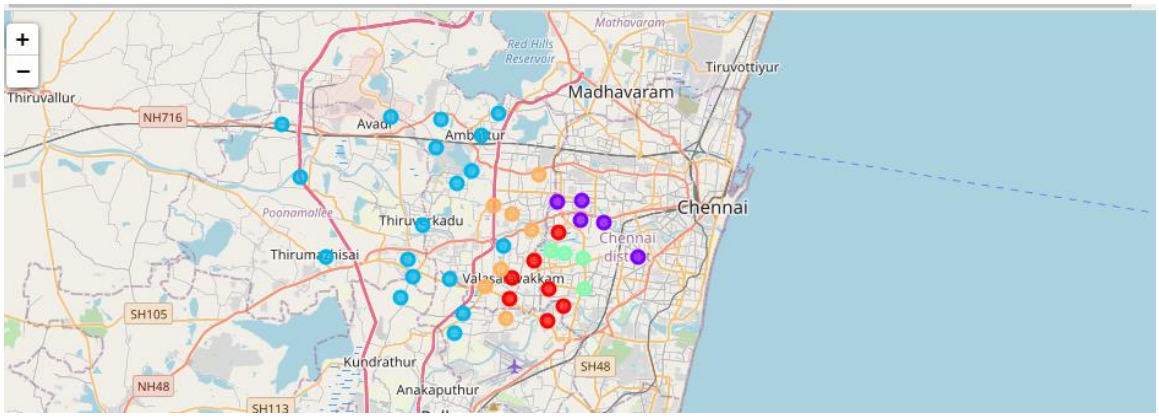
We primarily use the one hot technique to analyze the data and the various categories. A little bit of data cleaning to convert the numeric types to integers and then add the total number of venues in the areas.

#NOTE – We have capped the maximum venues at 100 but there can be more in an area. This is done due to the limited number of requests available. Moreover, this gives us a sense of understanding of areas with less than or more than 100 venues. Those with less than 100 are not as dangerous as those more than 100.

CLUSTERING WITH K-MEANS

Next step is to create clusters of similar areas and assign them priority orders to set up testing centers. We use the K-Means Machine Learning algorithm to divide West Chennai into five different clusters and then concentrate on each cluster setups to ensure efficiency of testing and trace an outbreak to a particular cluster in case of an unfortunate outbreak.

After creating the clusters, we can visualize it on the map with Folium to understand the geographical orientation of our newly formed clusters.



We have 5 different clusters represented by Red, Light Blue, Dark Blue, Orange and Green respectively. The clusters are divided on the severity of action needed in each of these clusters.

FINAL ANALYSIS

The final work includes an analysis of the clusters to figure out which cluster needs the most attention and is a possible Covid-19 hotspot when the lockdown is lifted. So, we find the sum of venues in each cluster and the cluster with the highest total should be clearly our top priority. An alternate would be finding the average number of venues in an area in each cluster separately but eventually they would still serve the same purpose as the number of areas in each cluster is almost the same.

RESULTS

After the final analysis, we are quite confident of the results and present the following output:

- Cluster 1 should be the highest priority for setting up testing centers as it is likely to get the most crowd
- Cluster C0 and C3 fall next in the priority order and should be the immediate priority after Cluster 1.
- Cluster C2 and C4 will receive lesser crowd as compared to the above clusters and therefore can be a little relaxed but close vigilance is still important.

	Cluster	Total Hotspots
0	C0	225
1	C1	479
2	C2	112
3	C3	252
4	C4	124

DISCUSSION

So, as we can see from the above obtained Data Frame that Cluster number 1, i.e. C1 has the highest amount of venues that people are likely to visit in the unlock phase and it is most likely to turn into Covid-19 hotspots. Hence, it is necessary to have testing centers in C1 at high priority followed by testing centers in C3, C0, C4 and C2 respectively. The target audience and stakeholders can take note of the following analysis before making a decision.

LIMITATIONS AND SCOPE FOR FUTURE RESEARCH

This is a preliminary analysis and the future scope is wide and extremely necessary because the more you informed you are, the better prepared you are. In the case of a pandemic, this statement becomes all the more important because people's lives are at stake. This project considers the parameter of only potential crowd gathering venues and does not take into account the existing Covid-19 hotspots in the area or around. It also does not count the containment zones that witness complete lockdown due to increased cases. Combining this with Covid-19 case database can give way to advanced research and a more wholistic point of view to make the best decisions.

CONCLUSION

Purpose of this project was to identify the neighborhoods in West Chennai which are more likely to receive a crowd once the lockdowns are lifted. This way, the government, hospitals, private labs or any other stakeholder involved in the testing phase of Covid-19 can be prepared by setting up appropriate number of permanent and mobile testing centers to avoid any possible outbreak due to sudden increase in movement of people.

Regular random testing in high risk neighborhood, C1 in this case can have multiple testing centers in and around to cater to the high needs. The remaining can have lesser number of centers or however the stakeholders propose the plan.

The findings presented will help the relevant stakeholders and if chosen to make public, can also help a common man avoid going to crowded places even after the lockdown is lifted. This is the new normal and we should be strong to face it.

REFERENCES

- https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai
- https://developer.foursquare.com/docs
- https://geocoder.readthedocs.io/
- https://python-visualization.github.io/folium/
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html