COVID-19 Pandemic Pharmacies Alert In Tirupati

1. Introduction:

1.1 Background

Tirupati city, an attraction of millions of pilgrims & tourists every year, is a home for more than four lakh native people. It's a city with a long history & culture and also a centre for evolving business, technology, arts & culture. In Tirupati, people have a lot of first-aid centers & hospitals to get the help from.

1.2 Problem Statement

For someone who is looking to open a pharmacy or clinic, it is vital to choose the neighbourhood and retail location. The goal of this project is to figure out the places with less or no clinics nearby & where they should be set up to run successfully with the accurate data analysis. During the lockdown periods, this will help people find the pharmacy or first aid medicines in the detoured routes.

During lockdowns due to pandemic COVID-19, this will also help us find out whether there is any nearby emergency medical shops in a particular area.

1.3 Targeted Audience

1. People in the Neighbourhood.
2. Government of Andhra Pradesh.
3. Municipal Corporation of Tirupati.
4. Business People Interested in medical field.

2. Data acquisition and cleaning:

2.1 Data sources

- First of all, I listed out all the Neighbourhoods in the locality from the official website of Municipal Corporation of Tirupati. Click Here to go there.

- For the neighbourhoods in that list we have to generate the Latitude and Longitude of each neighbourhood. For this, I have used the "geopy" library. Even before we start our coding, first we will find out the latitude and longitude of Tirupati which will be helpful in drawing the maps by using folium library.

- I have used Google places API instead of Foursquare API, as it has more reliable & accurate information about places in Tirupati.

2.2 Data cleaning

As we have the list of neighbourhoods in Tirupati, Next task for us is to generate the latitude and longitude for each neighbourhood in the list. To achieve this I have used the pandas 'apply' method to iterate the process for all the vales in the list. The output from the method is a tuple with Latitude and Longitude of each neighbourhood. I am assigning the output to a new column named Lat_Long. As the Output from the above method is a tuple and we need it as two columns, for that I am again applying lambda function for the iterative process on the Lat_Long column to split it into two other columns. After we split the column and we have the required two columns we no longer need the Lat_Long column so dropping the column from that data frame. To save this iterative process and to avoid running it through geopy again I am saving it as intermediate data file. Finally the output looks like below.

	Location	latitude	longitude
0	TTD Quarters, Tirupati, Andhra Pradesh, IN	13.625078	79.423043
1	Reddigunta, Tirupati, Andhra Pradesh, IN	13.758038	79.812770
2	Tiruchanur Road, Tirupati, Andhra Pradesh, IN	13.631430	79.460467
3	Renigunta Road, Tirupati, Andhra Pradesh, IN	13.638732	79.510476
4	Karakambadi Road, Tirupati, Andhra Pradesh, IN	13.654813	79.469513

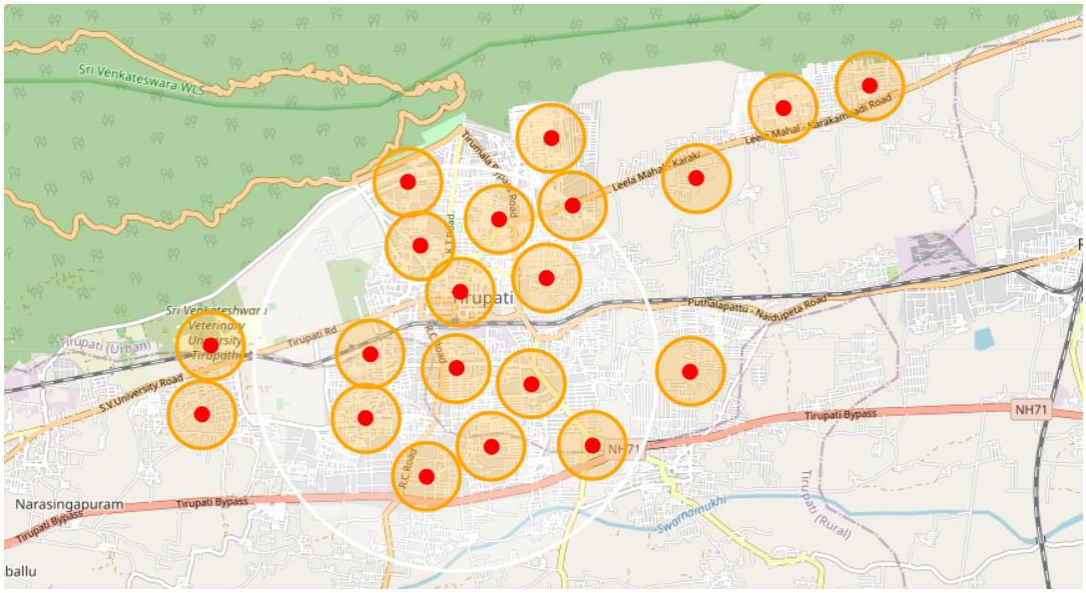
- As a part of data collection, I have to find the list of all the pharmacies in Tirupati with some other parameters like rating, total ratings, latitude, longitude. I tried using the Foursquare API ended up with less number of data points. So, I am left with no other option than using the Google Places API. To minimize the number of calls to Places API and to cover the entire locality with less number of overlaps. I have used the k-means clustering logic to divide the neighbourhoods with latitudes & longitudes into 'n' number of segments in such a way that it should full fill my requirement. I took the cluster centres to search nearest pharmacies around those centres so that it covers the entire geography of Tirupati with optimum number of call and less duplicity.

- To visualize the above concept, I have used folium library to prove my concept.

- From the below map, you can observe that there are circles with red dots in each circle. Each dot represent the centroid of that segment. So if you

see there are ‘n’ circles in the geography of Tirupati which shows us that we have covered most of the location.

- There is a white circle also in the middle showing like most of segments are inside the circle which means that the segments which are outside the white circle we can treat them as outskirts.



- Once we have the cluster centres from the above algorithm we are converting the cluster centres with labels into a data frame for our further data collection process and data analysis part. The converted data frame looks like below.
- In the below data frame, C_latitude, C_Longitude, Label the columns represent the centroid/segment latitude, longitude and label.

	C_Latitude	C_Longitude	Label
0	13.611688	79.424375	0
1	13.643783	79.435583	1
2	13.632251	79.420186	2
3	13.659750	79.476500	3
4	13.625100	79.385800	4

- Google Places API helps us in fetching the nearest pharmacies of each segment within a particular radius. The maximum limit of this API is 5km. It can fetch only the places around the latitude and longitude in the range of 5km but not more than that.

- There is less chance of the data duplicity. Still, we don't have to take a chance or introduce noise to the algorithm. To achieve that we are removing the duplicates from the above data frame.

- I have seen there are some places with no ratings available from the **Places** API. So I am filling them with Zeros instead of dropping because will be left with less number of data points we drop every point which doesn't match our criteria.
- Once we have filled the *NA*'s with Zeros now will clean the data frame to make it look self-explanatory.

- I have dropped the columns which are not required for our process column names are given by:

  ID: The place ID which adds no value to our algorithm except to check the duplicity.

  Place_Type: The type of the place as we are searching from specific keywords like *Pharmacy* It Adds No value.

  Place_Address: The address of the store this might be helpful but not now.

After doing all the above, the cleansed data frame which will be used for further analysis looks like below.

	Rating	Total_Ratings	P_Latitude	P_Longitude	Place_Name	Label	C_Latitude	C_Longitude
0	4.8	4.0	13.617692	79.422581	New Ganesh Pharmacy	0	13.611688	79.424375
1	0.0	0.0	13.617862	79.422495	Tejasree Medicals & Fancy Shop	0	13.611688	79.424375
2	0.0	0.0	13.618355	79.422433	City Pharmacy	0	13.611688	79.424375
3	3.8	5.0	13.647799	79.430087	Apollo Pharmacy	1	13.643783	79.435583
4	4.5	4.0	13.642131	79.428316	Apollo Pharmacy	1	13.643783	79.435583

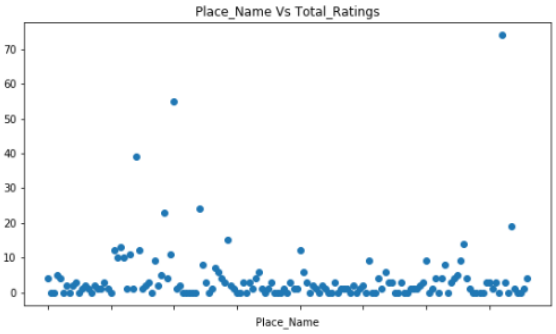
3. Exploratory Data Analysis:

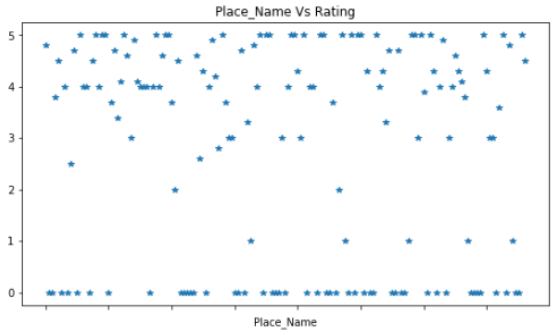
3.1. Relationship between place name & rating and the relationship between place name & total number of ratings

To add value to our solution instead of simply telling which segments have Medical stores available and Segments doesn't have medical stores. We can also segment the available Medical shops into different categories. To do that we need to analyse the ratings of each place. We are going to do the same in the below cell. Initially, I have drawn a bar plot between *Place_Name* and *Rating* I have seen most of the places have highest ratings. when I have drilled down the analysis if we see how most of the places have highest ratings is there are only one or two entries for certain number of stores and those are highest so why most of our stores have highest ratings.

If you see the below box plots of two different categories like rating and Total_Ratings:
The first plot tells us that most of the stores do not have more than 10 reviews per store.
But in the right, the second plot tells us that most of the stores will fall under ratings of 3.5 and above.

This is not as reliable as we should trust the rating alone of that store. So, we also have to consider the number of ratings to evaluate a store.





3.2 Feature selection

After data cleaning, we have 122 locations which are further divided as ‘n’ number of clusters. The main features that were used are locations, latitude & longitude of place, pharmacies, latitude & longitude of pharmacies & ratings. To get more reliable & accurate data after the analysis, we added he features of total number of ratings, place name which is pharmacy name & label.

4. Modelling:

Clustering & segmentation

K-means clustering is a method of vector quantization, originally from signal processing that aims to partition ‘n’ observations into ‘k’ clusters in which each observation belongs to the cluster with the nearest mean (center of clusters or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. K-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

Along with finding of segments with less number of Medical Stores, I wanted to give an add-on by segmenting the available medical shops into three categories good, average, can’t say. From the exploratory data analysis above I found that we have to give both total ratings and rating of the cluster as input to k-means to segment them into three categories.

So, from the cleansed data frame, we take only required columns and process them through the algorithm to get the labels for each store. Once we have the labels from the Algorithm, we can view the data and categorize them into our categories.

As the input for our algorithm is numerical data the output will also be in numerical type. If we see the above output we have three labels called **0**, **1, and 2.** Now, it’s time to convert them to Categorical labels by eye-balling the data.

Group-1

If we have a look at the data with labels **0** most of the stores ratings are high and the total number of ratings available are less. So we can say that **Cluster_Labels** with **0** cannot be determined as either **Good** or **bad** with the data that we have.

	Rating	Total_Ratings	P_Latitude	P_Longitude	Place_Name	Label	C_Latitude	C_Longitude	Distance	Cluster_Labels
0	4.8	4.0	13.617692	79.422581	New Ganesh Pharmacy	0	13.611688	79.424375	0.70	0.0
1	0.0	0.0	13.617862	79.422495	Tejasree Medicals & Fancy Shop	0	13.611688	79.424375	0.72	0.0
2	0.0	0.0	13.618355	79.422433	City Pharmacy	0	13.611688	79.424375	0.77	0.0
3	3.8	5.0	13.647799	79.430087	Apollo Pharmacy	1	13.643783	79.435583	0.74	0.0
4	4.5	4.0	13.642131	79.428316	Apollo Pharmacy	1	13.643783	79.435583	0.81	0.0
5	0.0	0.0	13.643108	79.432346	Sri Venkateswara Family Clinic	1	13.643783	79.435583	0.36	0.0
6	4.0	2.0	13.642236	79.430497	Gayathri Medical & Fancy	1	13.643783	79.435583	0.58	0.0
7	0.0	0.0	13.644531	79.430123	Sri guru pharmacy	1	13.643783	79.435583	0.60	0.0
8	2.5	2.0	13.642339	79.429211	New Gogula Medicals	1	13.643783	79.435583	0.71	0.0
9	4.7	3.0	13.642024	79.429295	Yuvan Medicals & Fancy	1	13.643783	79.435583	0.71	0.0

Group-2

If we have a look at the data with labels **1** the stores ratings and the total number of ratings available are good. So we can say that **Cluster_Labels** with **0** can be determined as **good** with the data that we have.

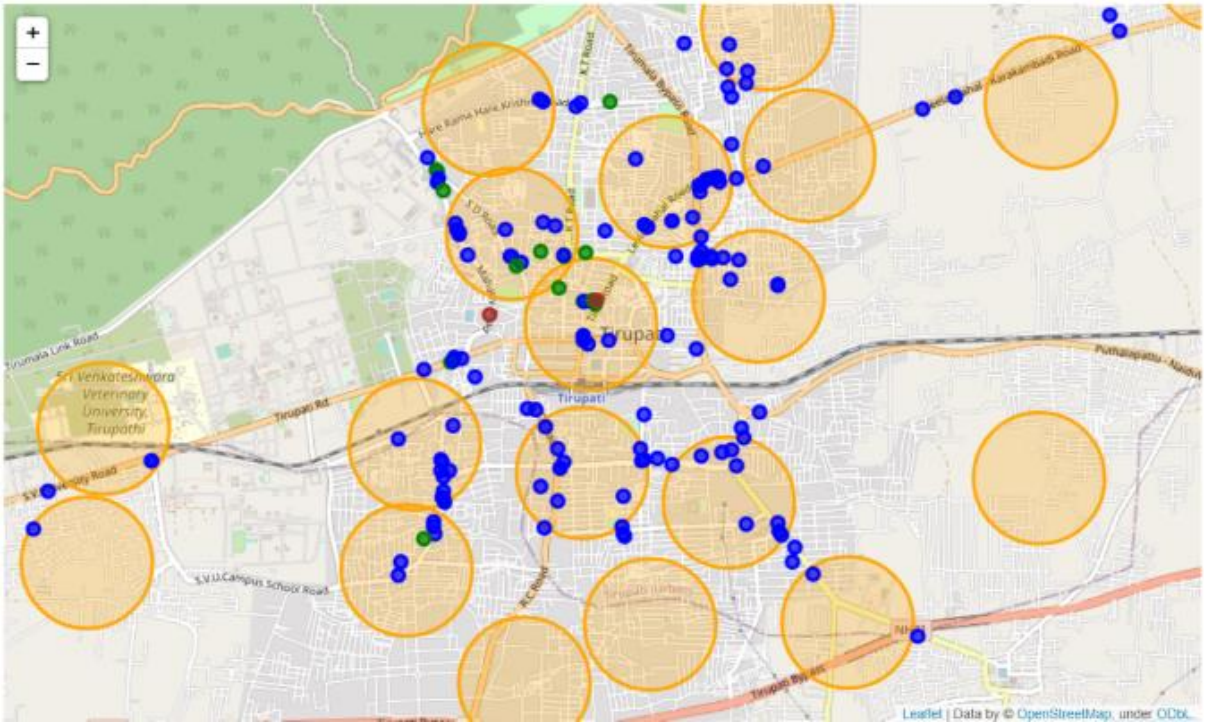
	Rating	Total_Ratings	P_Latitude	P_Longitude	Place_Name	Label	C_Latitude	C_Longitude	Distance	Cluster_Labels
28	4.9	39.0	13.632915	79.413075	AR Medicals	2	13.632251	79.420186	0.77	1.0
40	3.7	55.0	13.633862	79.420609	Hema vet & Poultry Medicals	2	13.632251	79.420186	0.18	1.0

Group-3

If we have a look at the data with labels **2** the stores ratings and the total number of ratings available are average. So we can say that **Cluster_Labels** with **2** can be determined as **Average** with the data that we have.

	Rating	Total_Ratings	P_Latitude	P_Longitude	Place_Name	Label	C_Latitude	C_Longitude	Distance	Cluster_Labels
21	3.7	12.0	13.633820	79.419643	Apollo Pharmacy	2	13.632251	79.420186	0.18	2.0
22	4.7	10.0	13.634697	79.417964	Sai Balaji Medical & Surgical Distributors	2	13.632251	79.420186	0.36	2.0
23	3.4	13.0	13.637172	79.419852	Apollo Pharmacy	2	13.632251	79.420186	0.55	2.0
24	4.1	10.0	13.636879	79.418317	MedPlus Thyagaraja Nagar	2	13.632251	79.420186	0.55	2.0
26	4.6	11.0	13.637238	79.416628	Galaxy Medical Distribution	2	13.632251	79.420186	0.67	2.0
29	4.1	12.0	13.636763	79.427966	Apollo Pharmacy	2	13.632251	79.420186	0.98	2.0
34	4.0	9.0	13.633667	79.420516	Sri Venkatasai Homoeo Medical Stores	2	13.632251	79.420186	0.16	2.0
37	4.6	23.0	13.633832	79.420441	Pavan's Pet's Mart	2	13.632251	79.420186	0.18	2.0
39	5.0	11.0	13.633889	79.420362	Vani Medicals	2	13.632251	79.420186	0.18	2.0
48	4.6	24.0	13.642759	79.409320	Adithya Pharmacy	5	13.646931	79.412958	0.61	2.0

5. Conclusion:



If we see the above map, things that we can conclude are given by:

1. Categories Of Stores
2. Segments With Density of Stores

As we have categorised the stores into three.

1. Brown colour indicates stores that are good.
2. Green colour indicates stores that are Average.
3. Blue colour indicates stores that cannot be determined.

If you see most of the medical stores are in Blue colour which says us that either they are not explored by the people or those stores are not that popular or they don't have enough supplies for the people. So it’s better if we can explore them more before even we start anything further.

And the Segmentation part the stores are not covering the entire geography of Tirupati location. There are locations to be explored for Business holders before they start their business in that particular location and to the people and Government, it’s a kind of warning kind of signal to alert those regions about the availability of medicals during this pandemic period.
To have specific analysis on each segment, find the below output.

The header Label tells us about the segment ID and count is number of stores in that segment.