

# THE CAPSTONE PROJECT

By - G.S.Sriram



## Introduction-

India is the second most populated country in the world. For many people who live here providing healthcare for such a highly populated country is a huge and critical task. We will have to set up a pharmaceutical enterprise at the right place in order to help the society and to gain business and earn profits too. We will have to find the place which has the least pharmaceutical enterprises to provide medical support to the people who are inaccessible to such facilities. This way we could have maximum profit and benefit the citizens in maximum possible ways.

## Business Problem-

The objective of this capstone is to analyze and select the best location in the city of Chennai, India to open a new pharmacy. Using Data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question, if a Pharmaceutical business developer wants to open a pharmacy which is the right place to open?

## Data-

- List of neighbourhood in Chennai
- Latitude and longitude coordinates of the neighbourhoods.
- Venue data particularly the ones related to Pharmaceutical enterprises.

## Approach-

Now let me take a metropolitan city Chennai in India. With the help of web scrapping we could get the list of all the areas in chennnai([https://en.wikipedia.org/wiki/List\\_of\\_neighbourhoods\\_of\\_Chennai](https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai)). Using geocoder we could get the coordinates of all the areas. Now the only thing left is to find out is the number of pharmacy in the neighbourhood. This could be done using FourSquare API and we can check which area has the lowest Pharmacy and that will be the right place to put up to have the best business.

## Methodology-

Firstly we find the right website for scrapping the data. So here the right website is

[https://en.wikipedia.org/wiki/List\\_of\\_neighbourhoods\\_of\\_Chennai](https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai). Now we must use python requests and BeautifulSoup packages to Convert the following data into proper table format. After the conversion we need to get the coordinates of

each of the places in the form of longitudes and latitudes. This could be done using Geolocation and get the latitude and longitude for each location. Now we populate the data into pandas dataframe and then visualize the neighbourhoods in a map using folium package. This allows us to visualize the coordinates data returned by geocoder.

Now the next task is to find the neighbourhood in that area like what all malls, hospitals are available in the neighbourhood using the Foursquare API. Before this we need to register a Foursquare developer account in order to obtain the Foursquare ID and secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in the python loop. The foursquare will return venue data in JSON format and we will extract necessary information.

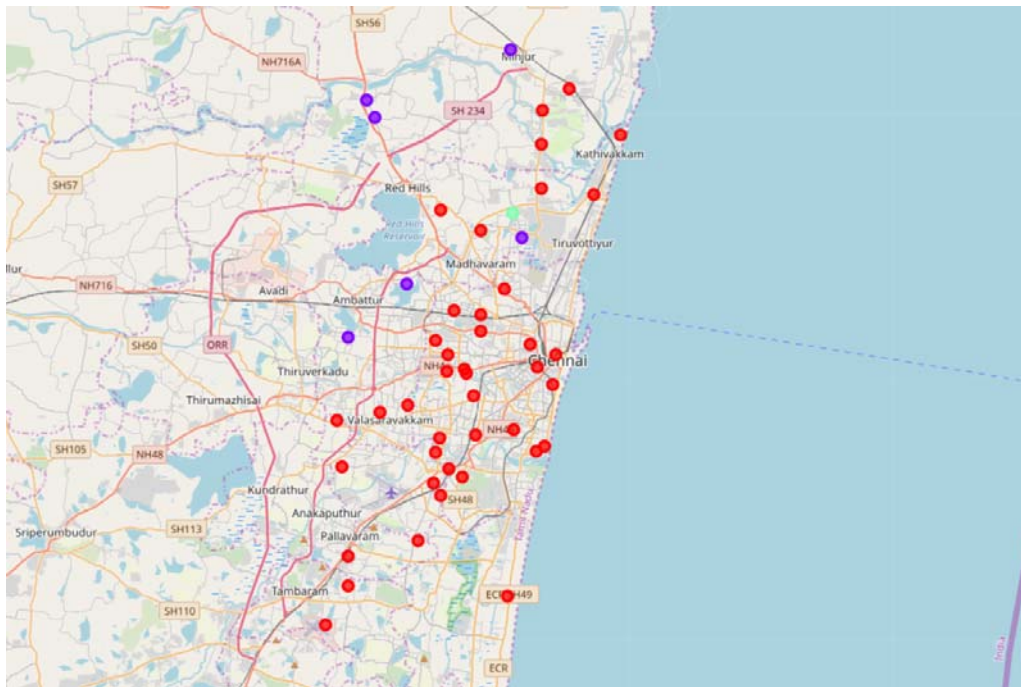
Lastly we perform clustering on the data by k-means clustering algorithm. We cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for Pharmacy. The results will allow us to identify which neighbourhoods have higher concentration of pharmacy while which neighbourhoods have fewer number of pharmacy. Now this will help us decide which location is suitable for opening new pharmacies.

## Results:

The results from the k-means clustering show that we can categorize the neighbors into 3 clusters based on the frequency of occurrence for “Pharmacy”:

- Cluster 0: Neighborhoods with low number to no pharmacies
- Cluster 1: Neighborhoods with moderate number of pharmacies
- Cluster 2: Neighborhoods with high concentration of pharmacies

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and Cluster 2 in mint green colour.



## Discussion

As observations noted from the map in the results section, most of the pharmacies are concentrated in the central area of Chennai city, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to totally no pharmacies in the neighborhoods. This represents a great opportunity and high potential areas to open new pharmacies as there is very little to no Pharmacies in that area. Meanwhile, pharmacies in cluster 2 are likely suffering from intense competition. From another perspective, this also shows that the oversupply of pharmacies mostly happened in the central area of the city, with the suburb area still have very few pharmacies. Therefore, this project recommends pharmaceutical developers to capitalize on these findings to open Pharmacies in neighborhoods in cluster 0 with little to no competition.

## Limitations

In this project, we only consider one factor, frequency of occurrence of pharmacies, there are other factors such as population and number of hospitals that could influence the location decision of a new pharmaceutical enterprise. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by the project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the

preferred locations to open a new shopping mall. In addition, this project made use of the free Foursquare API that came with limitations as the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

## Conclusions

In this project, we have identified the business problem, extracted the data and performed machine learning by clustering the data into 3 clusters based on the similarities, and lastly providing recommendation to the enterprise developers to give the best location to open a new pharmacy. The findings of the project will help us improve the rate of success in establishing a pharmaceutical enterprise.