**1. Introduction**

**1.1 Background**

Chennai is the capital city of Tamil Nadu, India and is vastly spread. The Chennai Metropolitan Area spreads over 1,189 square kilometers and has a population of about 9 million people. The city is broadly divided into 15 zones. The local administration aims to provide good healthcare to the public and continues to improve the same.

For the sake of this project, I am a consulting Data Scientist, to help them provide the necessary guidance on the location of their upcoming healthcare facility. The administration is of the opinion that given the size of the city which continues to expand and the growing population, there is enough scope to increase the number of healthcare facilities. Of course, they have ear-marked in their budgets a specific sum for new / expansion of healthcare facilities. The local administration has approached me to help them prioritize where new facilities could be developed, to improve the reach of such facilities to the population.

**1.2 The Problem**

The location of a new healthcare facility will depend on multiple factors, such as the population, already existing facilities, air quality & water quality in the given area, number of households, accessibility & cost factors. This project aims to group the entire population with these characteristics and recommend possible locations where new factilities could be built.

**Stakeholders**

The Local administration is the primary stakeholder, who is interested in creating and maintaining new healthcare facilities. Apart from them, the general public at large is a key stakeholder, as their contributions by way of local taxes are being spent for such development activities. So, there is a keen interest by the general public and the local administration to improve the overall healthcare system of the city of Chennai, which will become a Model City for the rest of the state and the country.

**2. Data Section**

**2.1 Data sources**

I plan to use the Foursquare APIs for collecting the location and neighborhood data as well as categories of hospitals / clinics etc. There have been problems accessing Foursquare website for quite sometime and hence, as an alternative, I plan to carry out webscraping from various sites which contain Chennai's demographic data as well as population and other details. The Chennai Metropolitan Development Authority maintains data related to various parameters grouped by individual Pincodes.

**2.2 Data Cleansing**

There are different websites which maintain different sets of data, like the Pollution Control Board which maintains the air quality, water quality while the district administration maintains the population, healthcare statistics. I will need to map the different areas / pin codes / zone codes to the respective rows and arrive at a consolidated dataframe for meaningful analysis.

**2.3 Feature Selection**

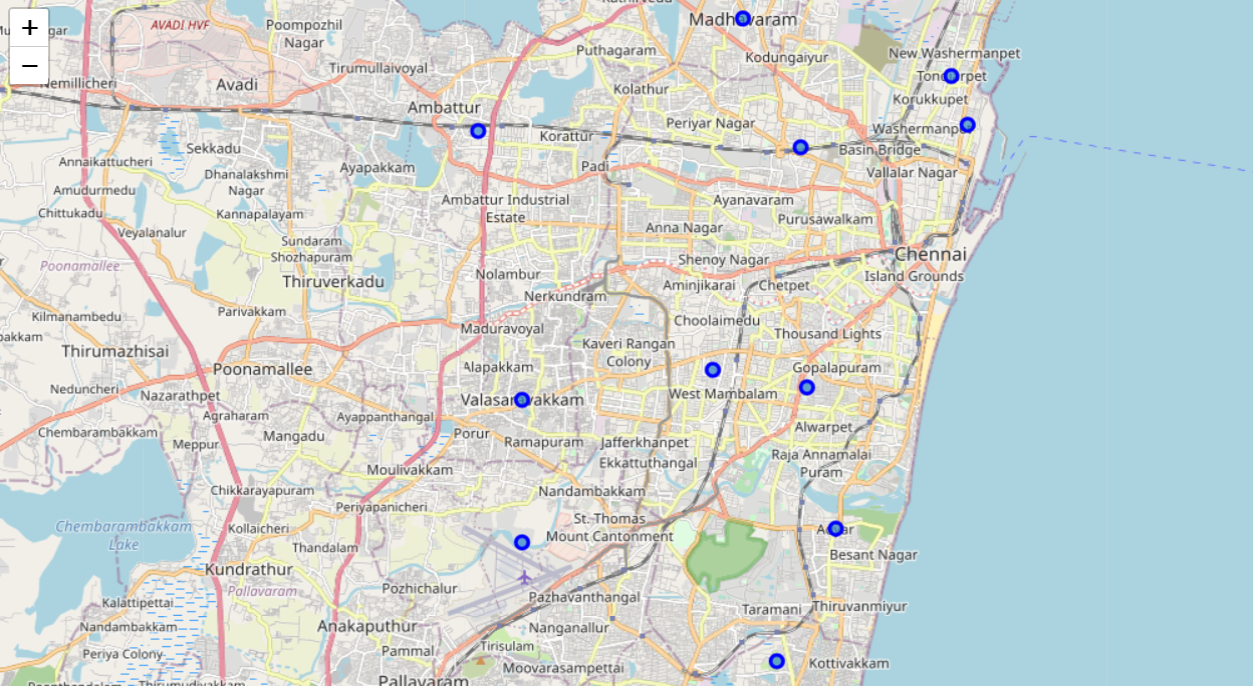
The primary features for this project will be the geographical division by way of pincodes, wards, zones with population, number of healthcare facilities, air quality, water quality data.

**3. Methodology and Analysis**

To analyse defined problem of the project a suitable analysis method need to be identified. The goal of analysis is to identify neighborhoods which are suitable for a new healthcare facility. Regression models are out of question as the final target is not of linear nature. Classification models are selected to be used as the problem requires neighborhoods to be clustered. Based on unlabeled data that is used in the problem solving of the project, it was decided to use k-Mean clustering model to segment all neighborhoods of Chennai in accordance of their similarity of the features selected, namely population, air quality, water quality, number of households.

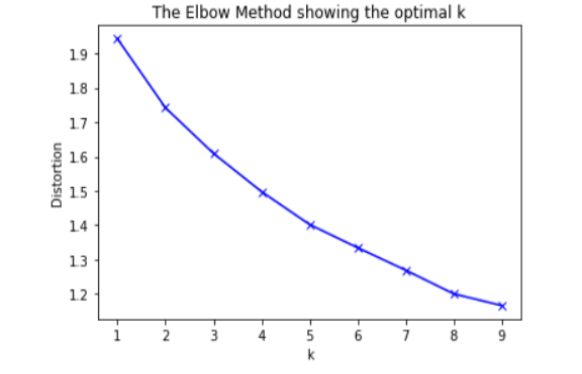
**3.1. Visual and Data Analysis**

Chennai is divided into multiple zones and to analyse data of neighborhoods, first visualization was created. Since Foursquare website has been down for a long time,. Geolocator API was used for identifying correct longitude and latitude coordinates of zones.

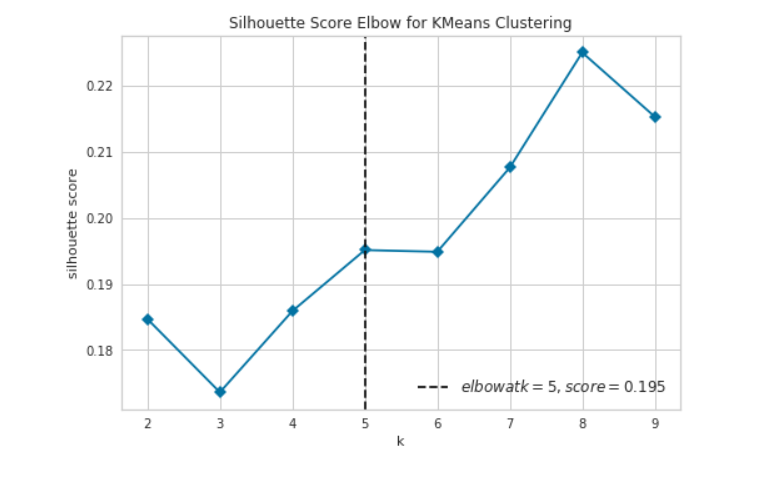


**3.2. Cluster Analysis**

For the sake of this assignment, since the population, air quality, water quality and households data were not directly available, I have used my learning of Python in this course, to generate random numbers and apply mathematical operations to mimic real-life data. As the consolidated data has different variations and scales, these features were Standardized to better fit clustering model. For standartisation purposes StandardScaler tool was used from Sci-Kit Learn library. After features of interest were standardized, optimal number of clusters for neighborhood clustering was identified. While fitting the k-Means clustering model with a cluster counts from 1 to 9, model distortions were calculated on every iterative step. Distortions were plotted to identify the “elbow” – the point in plotted line that represents the optimal number of clusters.



Visually inspecting the plot, a clear elbow point is not identified. To find optimal cluster number of our investigation, Silhouette score of k-Means model was used in the next step of modeling.



This time, highest Silhouette was achieved with 8 clusters of the k-Means model. Identified optimal number of clusters is used in next steps of clustering the neighborhoods. k-Means model was one more time fitted with standardized feature dataset.

Clustering analysis revealed neighborhoods which are relatively similar based on their population density and the features under consideration. The results of analysis provided answers to the defined problem for the project. Based on this, the ranking of the different zones has been arrived at, for the local administration to decide on where to come up with the next healthcare facility.

**4. Results and discussion**

Data acquisition and cleaning steps of the analysis provided the insights that many data sets that may be found on internet or might not be available in the template / format as needed. For a city like Chennai, it was very difficult to get a lot of information. But for the purpose of this project, I had simulated the data using the coursework that I had learnt, using various methods / functions, including random number generation. One would naturally think that having a thickly populated place calls for the next location for healthcare facility, but this analysis has given a clear scientific and statistical data on how to approach the problem and arrive at a ranking, which can be effectively used to upcoming healthcare facilities. In fact, a few other features that can be used in the next iteration are the cost of per sq. ft area, count of existing healthcare facilities etc. This analysis has given the top 3 zones to focus on, as Adyar, Annanagar and Royapuram.

**5. Conclusion**

With the city of Chennai being grouped administratively into different zones, the question on where should we focus our energies on coming up with the next healthcare facilities has been addressed by arriving at the top 5 zones. This will help the local administration to prioritize the facility and also collect additional data which can be used for further refining this solution, as and when newer facilities come up.