2. DATA ANALYSIS - A DESCRIPTION OF THE DATA AND HOW IT WILL BE USED TO SOLVE THE PROBLEM

DATA REQUIREMENT AND COLLECTION

(i) List of neighborhoods in New Delhi, India have been collected for the instant Project. This is required to segregate neighborhoods into clusters and carry out further analysis.

(ii) Latitude and Longitude coordinates of identified neighborhoods have been collected for the instant Project. This has been used for plotting on maps as well as to arrive at the venue data of various neighborhoods.

(iii) Use of Foursquare API for fetching the nearest venue locations for defining clusters along with details of venues and their names. This has been used to make detailed analysis for arriving at the best locations for developing Multiplex.

(iv) Venue data, especially that related to 'Multiplex'. This has helped in clustering neighborhoods for arriving at the results of the project and making final recommendation.

3. METHODOLOGY

A) WEB SCRAPING AND BUILDING DATAFRAME

- In this step, list of neighborhoods were scared from Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi)

- The same was done using Python requests and beautifulsoup packages to extract the list of neighborhood data. With these packages, the list of names of neighborhoods of New Delhi, India were populated into a dataframe.

B) GEOGRAPHICAL COORDINATES OF IDENTIFIED NEIGHBORHOODS

- In this step, Geocoder package has been used to extract geographical coordinates of identified neighborhoods of New Delhi, India. After gathering the data, the same has been populated in pandas dataframe.

- The identified neighborhoods are then visualized in a map using Folium package.

- The process has helped in checking the output of geographical coordinates data returned by Geocoder and its correct plotting on the map.

C) FOURSQUARE API FOR EXPLORING IDENTIFIED NEIGHBORHOODS

- Foursquare API has been used to fetch nearest venue locations so as to use them in forming a cluster. In the present case, Foursquare API has been used to get top 100 venues that are within the radius of 2000 meters.

- For the said purpose, a new Foursquare Developer Account has been opened for obtaining Client ID and Client secret key.

- Thereafter, API calls are made to Foursquare API by passing the geographical coordinates of identified neighborhoods in a Python loop.

- Foursquare returns venue data in JSON format from which venue name, venue category, venue latitude and venue longitude are extracted.

- Each neighborhood is then analyzed by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.

- Data is accordingly prepared for clustering. Thereafter, the data is filtered for analyzing data related to 'Multiplex'.

D) K-MEANS CLUSTERING

- In this step, clustering on data is performed using k-means clustering. In this method, k number of centroids are identified and then allocated to nearest cluster, while keeping centroids as small as possible.

- This method is useful in the instant case considering the fact that clustering would help in narrowing down the target neighborhood by eliminating clusters and then coming at the final result of the problem at hand.

- In the instant case, neighborhoods have been clustered into three (3) clusters based on frequency of occurrence of 'Multiplex'. The clusters identify the level of concentration of multiplexes in each neighborhood and which is the best suited neighborhood with fewer number of multiplexes and providing opportunity to open a new multiplex.

- Thereafter, clusters are then visualized in a map using Folium package.

E) STATISCIAL ANALYSIS OF CLUSTERS

- Considering the first cluster has very sparse and wide neighborhood data related to multiplexes, it is considered appropriate to carry out statistical analysis on the other two clusters.

- The statistical plotting library of seaborn is then used on two clusters to analyze the results arrived in the analysis carried out while using k-means clustering.

- On identification of the cluster, as an illustrative case, one neighborhood is explored to arrive at the final decision of opening a multiplex.