DELHI FOR EVERYBODY

By Navya Jain

17.06.2021

TABLE OF CONTENTS

1. INTRODUCTION

Being the capital of India, Delhi is the heart of international trades and multinational cooperates of India. The city is not only known for its diverse culture, but also for its immense sense of nationality due to the presence of major political institutions and historic monuments like the Red Fort.

1.1 BACKGROUND

There exists a lot of reasons people move to the capital.

Delhi has rich heritage and culture. It has beautiful architecture and eclectic mix of cultures.

The cosmopolitan city holds many multinational corporation's offices, and many embassies. Globalization in India can be best found in Delhi.

The metropolitan's public transport is improving rapidly, and new metro stations are making it fairly easy to commute from one place to another inside the city.

It has one of the best healthcare facilities in India, and holds major political places like the Parliament house. It has a dynamic economy.

The number of people projected to be living in Delhi by 2026 is around 30 million. Rapid urbanization, in conjunction with the intensified challenges of environmental degradation, has placed pressure on infrastructure, housing availability and the spread of slums.

1.2 BUSINESS PROBLEM

Known to be a city which has a place for everybody, moving to Delhi can be a life changing decision for anybody. It can, however, also be very overwhelming.

Throughout the project, every district of the capital is analysed to provide a person with a clear view of Delhi's infrastructure and facilities to make better decisions regarding location in the city.

Even though the capital has a place for everybody, everyone should be able to make informed decisions to find a place in the heart of India based on their needs.

1.3 WHO IS THIS ANALYSIS FOR?

For people who either aim to move to Delhi, or want to change their neighbourhood in Delhi, or want to understand the city's infrastructure better in order to look for opportunities, this report contains detailed analysis of all the areas of the union territory.

Delhi is for everybody – right from poets to technology enthusiasts, it has room for everybody to grow.

This report aims at assisting anybody looking for opportunities or life in the capital.

2. DATA DESCRIPTION

In order to analyse every aspect of Delhi, a list of all the postal offices is needed. This list is found at https://dmsouthwest.delhi.gov.in/std-pin-codes/ . It contains over 410 postal areas along with their PIN codes.

In order to access the FourSquare API, the latitudinal and longitudinal data of all the areas are collected using geocoder.

Venues belonging to various categories are also collected by using FourSquare and making API calls.

3. METHODOLOGY

3.1 IMPORTING LIBRARIES

This is a pre-requisite step, and it is done to import packages in Python which includes built-in functions required.

Libraries are imported for data scrapping, wrangling, pre-processing, analysis, and machine learning.

3.2 DATA COLLECTION AND WARNGLING

3.2.1 WEB SCRAPPING

In order to analyse the infrastructure of any city, it is vital to have a list of the various districts or neighbourhoods available.

This information is available on https://dmsouthwest.delhi.gov.in/std-pin-codes/ in the form of 2 tables.

Both the tables together constitute over 410 postal offices in Delhi along with their PIN codes.

The data is loaded into the data frame by using the technique of web scrapping. Since the information is directly available in the form of HTML tables, they can be directly combined and loaded into a data frame using the pandas library.

The obtained data frame consists of 410 rows and two columns – Area name and PIN code.

3.2.2 DATA GEOCODING

In order to request various places (referred to as venues) near a postal office, which is usually at the centre of the area, it is imperative to acquire the latitudinal and longitudinal data of all of the postal areas.

This can be simply done by using the geocoder library.

For this analysis, the geocoder library is used to enter the postal areas and PIN codes, and is used to return the given address into its coordinates.

The data obtained of the 410 neighbourhoods is then appended into the data frame.

For better results, the data frame is checked for empty values, just in case the geocoder library did not return coordinates of an area. Such empty cells are replaced with the NumPy NaN and deleted.

3.2.3 VENUES

FourSquare is a location data provider. API calls can be made to obtain the data of venues using FourSquare. In order to make API calls, a developer account is created to access credential details and access token.

In order to obtain the venues of an area, FourSquare API is used. But before using it, version and credentials are defined and printed.

API calls are then made using the FourSquare API in a loop to obtain 250 venues of all the postal areas. Files containing the top 250 venues in a 2 km radius of the centre of the neighbourhoods or the postal areas are obtained using the FourSquare API.

The FourSquare API returns the information for each area in the form of a .json file, which is then manipulated to extract information and appended in a data frame.

3.2.4 QUALITY VENUES

Not every type of venue or venue category is of interest to people moving in Delhi. Results of analysis are just as good as the quality of data used to obtain them, and in order to get quality results, it is important to use good quality data.

To make a clean data frame consisting of useful categories, first all the unique categories were found. There were 228 unique categories, which were grouped into umbrella categories to have a clean data frame.

Venues not corresponding to the quality categories were dropped from the data frame and was stored in a csv file.

	Area	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
1	Ajmeri Gate Extension	28.64366	77.22883	Haveli Dharampura	28.653247	77.232309	Hotel
8	Ajmeri Gate Extension	28.64366	77.22883	bloomrooms @ New Delhi Railway Station	28.645537	77.217701	Hotel
10	Ajmeri Gate Extension	28.64366	77.22883	Fabindia	28.632012	77.217729	Stores
12	Ajmeri Gate Extension	28.64366	77.22883	The Prime Balaji Deluxe @ New Delhi Railway St...	28.645247	77.217433	Hotel
20	Ajmeri Gate Extension	28.64366	77.22883	The Indian Grill Restaurant	28.646141	77.215133	Restaurant
...	...	...	...	...	...	...	...
15816	Zafrabad	28.67970	77.27151	Welcome Metro Station	28.671902	77.277772	Commute
15817	Zafrabad	28.67970	77.27151	Seelampur Metro Station	28.669805	77.266846	Commute
15818	Zafrabad	28.67970	77.27151	Shivaji park	28.682657	77.285503	Sports
15819	Zafrabad	28.67970	77.27151	yamuna vihar	28.689816	77.283876	Sports
15820	Zafrabad	28.67970	77.27151	Shahdara Metro Station	28.673344	77.289011	Commute

3.3 EXPLORATORY DATA ANALYSIS

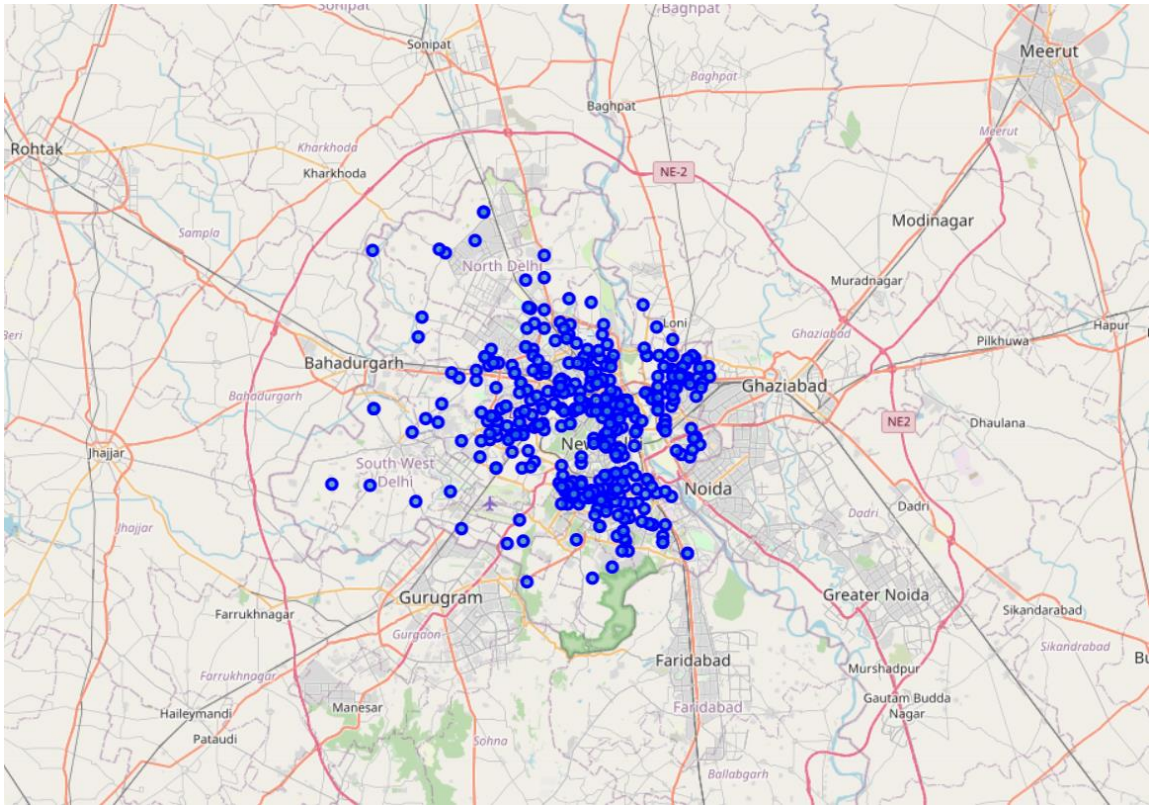
Exploratory data analysis is an approach of analysing data sets to summarize main characteristics of a dataset.

It is performed to understand the data better, and is often accompanied by data visualization and statistics.

3.3.1 MAP OF DELHI WITH AREAS MARKED

To understand the locational data and the relationship of infrastructure with location, it is imperative to understand how the different postal areas are placed in Delhi.

This is done by using the folium library to plot a map of Delhi with areas superimposed on top.



3.3.2 NUMBER OF VENUES IN EACH AREA

To know which area has the greatest number of good quality and important infrastructural buildings, the total number of final venues obtained are grouped and stored.

This is done to have an overview of the different areas in Delhi, and to find out the best places in the metropolitan.

3.3.3 FINDING OUT TOP 10 VENUES OF EVERY LOCATION

Every location contains a number of places, and it would be better to understand the data if the top 10 venues of each location is found out.

In order to obtain a table consisting of the top 10 venues of every location, first, one hot encoding is done on the data to replace categorical values with binary values.

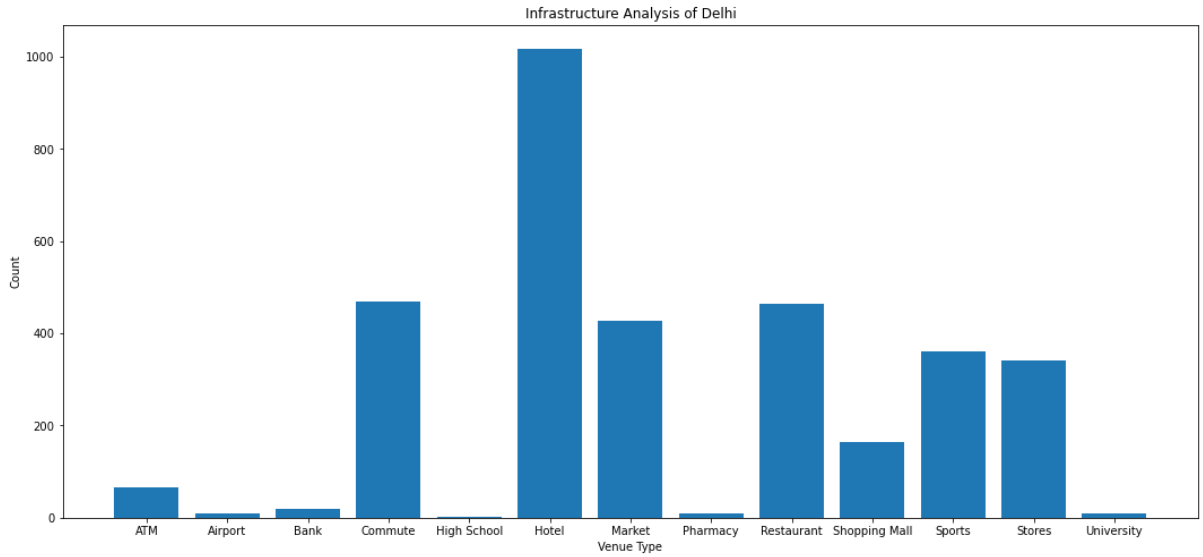
Then, the top types of venue that occur the most are found by calculating the frequencies of each type of venue in an area.

This data is stored in a data frame and then saved in a csv file.

3.3.4 ANALYSING THE TOTAL INFRASTRUCTURE OF DELHI

A good overview of Delhi is provided by first calculating the number of different types of venues available in Delhi irrespective of their postal areas.

They are then stored in a separate data frame, and plotted on a bar graph.



3.4 MODEL DEVELOPMENT AND DEPLOYMENT

For an end-to-end project, a machine learning algorithm is applied on the dataset.

For this project, clustering is applied to cluster neighbourhoods according to their venue attributes in order to help people make informed decisions by diving the areas into buckets or clusters.

To deploy clustering machine learning technique, KMeans clustering algorithm is used, which aims at increasing the inter-cluster distance and reducing the intra-cluster distance.

3.4.1 ONE HOT ENCODING

Machine learning algorithm like KMeans clustering requires numerical values to process data and yield results.

The data frame consists of categorical or labelled values. For example, the VenueCategory column in the data frame consists of categorical values like ‘Hotel’ or ‘Pharmacy’.

Categorical values are converted into numerical or binary values using the method of one hot encoding.

3.4.2 GROUPING THE TABLE

After replacing the table’s categorical values with numeric values, the table is grouped using the values obtained after performing one hot encoding.

3.4.3 K-MEANS CLUSTERING

The algorithm used in the project to perform clustering on the postal area dataset and to form clusters of different areas using the postal area dataset is K-Means clustering.

It aims at increasing the inter-cluster distance and reducing the intra-cluster distances.

The algorithm is deployed using the sci-kit learn library.

The obtained clusters are then appended to the dataset consisting of the top ten venues in all the 387 area locations.

To the data frame consisting cluster labels, latitudes and longitudes are appended for easier plotting and visualization using the folium library.

4. RESULTS

After performing clustering on the data frame, three clusters were obtained.

This means that all the areas in Delhi were grouped into 3 clusters consisting of similar attributes.

4.1 CLUSTER 1

The first cluster obtained looks like:

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Clusters	Pin Code	Latitude	L
3	A T Mills	Restaurant	Hotel	University	Stores	Sports	Shopping Mall	Pharmacy	Market	High School	Commute	0	110033	29.683990	7
5	Ajmeri Gate Extension	Hotel	Market	Stores	Restaurant	University	Sports	Shopping Mall	Pharmacy	High School	Commute	0	110002	28.643660	7
10	Ambrohi	Hotel	Restaurant	Stores	Commute	University	Sports	Shopping Mall	Pharmacy	Market	High School	0	110045	28.634100	7
11	Amrit Kaur Market	Hotel	Stores	Restaurant	Market	Commute	University	Sports	Shopping Mall	Pharmacy	High School	0	110055	28.643040	7
18	Ansari Road	Market	Hotel	Stores	University	Sports	Shopping Mall	Restaurant	Pharmacy	High School	Commute	0	110002	28.646000	7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
371	Tughlakabad	Restaurant	Hotel	University	Stores	Sports	Shopping Mall	Pharmacy	Market	High School	Commute	0	110044	28.518580	7
372	Tughlakabad A F Station	Restaurant	Hotel	University	Stores	Sports	Shopping Mall	Pharmacy	Market	High School	Commute	0	110044	28.518580	7
374	Udyog Bhawan	Hotel	Restaurant	Sports	Commute	Market	University	Stores	Shopping Mall	Pharmacy	High School	0	110011	28.611850	7
376	Vasant Vihar I	Hotel	University	Stores	Sports	Shopping Mall	Restaurant	Pharmacy	Market	High School	Commute	0	110057	27.906793	7
391	Zakhir Nagar	Hotel	Sports	University	Stores	Shopping Mall	Restaurant	Pharmacy	Market	High School	Commute	0	110025	28.567530	7

102 rows × 15 columns

The first most common venue type in the areas in this cluster are of hotels or restaurants, followed by universities are stores.

4.2 CLUSTER 2

The second cluster obtained looks like:

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Clusters	Pin Code	Latitude	Longitu
1	A G C R	Shopping Mall	Commute	University	Stores	Sports	Restaurant	Pharmacy	Market	Hotel	High School	1	110002	28.630798	77.275
4	Adarsh Nagar	Commute	Sports	University	Stores	Shopping Mall	Restaurant	Pharmacy	Market	Hotel	High School	1	110033	28.720350	77.172
7	Alipur	Restaurant	Commute	University	Stores	Sports	Shopping Mall	Pharmacy	Market	Hotel	High School	1	110036	28.798050	77.144
8	Amar Colony	Commute	Stores	Sports	University	Shopping Mall	Restaurant	Pharmacy	Market	Hotel	High School	1	110024	28.678440	77.049
12	Anand Nagar	Commute	Sports	Market	Hotel	University	Stores	Shopping Mall	Restaurant	Pharmacy	High School	1	110005	28.674490	77.167
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
363	Tatarpur	Commute	University	Stores	Sports	Shopping Mall	Restaurant	Pharmacy	Market	Hotel	High School	1	110027	28.688448	77.317
369	Timarpur	Commute	University	Stores	Sports	Shopping Mall	Restaurant	Pharmacy	Market	Hotel	High School	1	110007	28.700780	77.221
375	Uttam Nagar	Stores	Commute	University	Sports	Shopping Mall	Restaurant	Pharmacy	Market	Hotel	High School	1	110059	28.617390	77.052
390	Zafrabad	Commute	Sports	University	Stores	Shopping Mall	Restaurant	Pharmacy	Market	Hotel	High School	1	110053	28.679700	77.271
392	Zakhira	Commute	Hotel	University	Stores	Sports	Shopping Mall	Restaurant	Pharmacy	Market	High School	1	110015	28.669784	77.162

89 rows × 15 columns

The first most common type of venue is that of commute, including venues like metro stations or bus stations. It is followed by university and then stores.

4.3 CLUSTER 3

The third cluster obtained looks like:

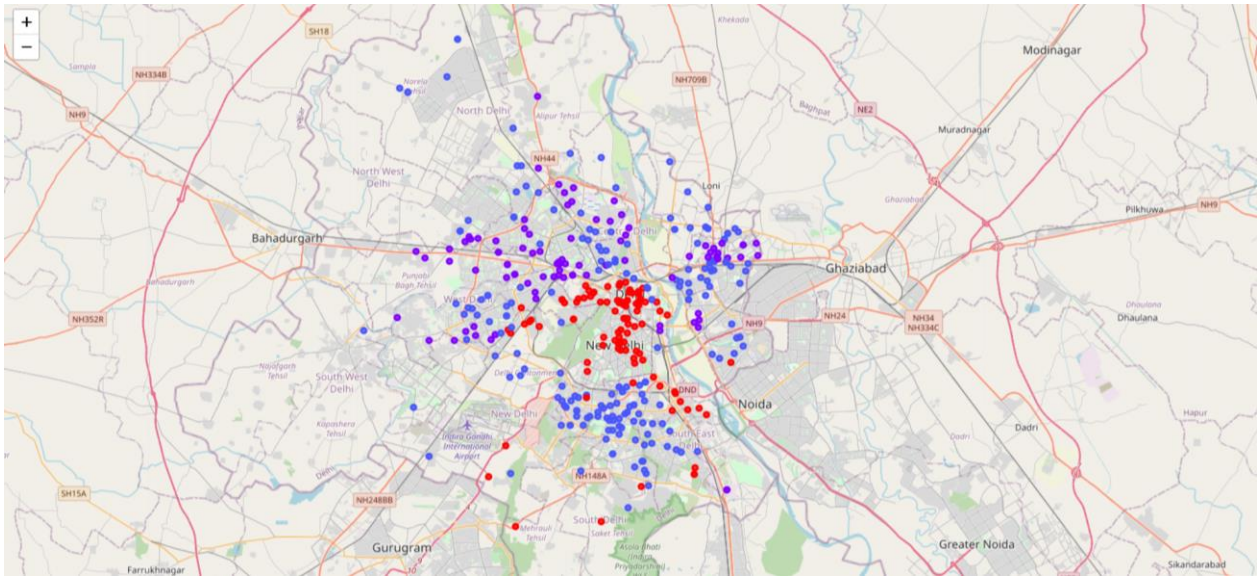
	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Clusters	Pin Code	Latitude	Lo
0	A F Rajokari	Stores	Shopping Mall	University	Sports	Restaurant	Pharmacy	Market	Hotel	High School	Commute	2	110038	28.51937	7
2	A P S Colony	Sports	Hotel	University	Shopping Mall	Pharmacy	Market	Stores	Restaurant	High School	Commute	2	110010	28.57523	7
6	Aliganj	Market	Sports	Restaurant	Commute	Hotel	University	Stores	Shopping Mall	Pharmacy	High School	2	110003	28.58354	7
9	Ambedkar Nagar	Stores	Hotel	Shopping Mall	Restaurant	Market	University	Sports	Pharmacy	High School	Commute	2	110062	28.51949	7
13	Anand Niketan	Sports	University	Stores	Restaurant	Market	Shopping Mall	Pharmacy	Hotel	High School	Commute	2	110021	28.57555	7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
385	Wazir Nagar	Market	Sports	Commute	Stores	Restaurant	Shopping Mall	Hotel	Airport	University	Pharmacy	2	110007	28.57029	7
386	Wazirpur Phase III	Sports	Restaurant	Stores	Market	Commute	University	Shopping Mall	Pharmacy	Hotel	High School	2	110052	28.69302	7
387	Work Shop	Stores	Commute	ATM	University	Sports	Shopping Mall	Restaurant	Pharmacy	Market	Hotel	2	110010	23.11945	8
388	Yamuna Vihar	ATM	Stores	Sports	University	Shopping Mall	Restaurant	Pharmacy	Market	Hotel	High School	2	110053	28.70061	7
389	Yusaf Sarai	Restaurant	Market	Sports	Commute	University	Stores	Shopping Mall	Pharmacy	Hotel	High School	2	110016	28.55863	7

202 rows × 15 columns

The first most common type of venue is of markets and stores, followed by commute, universities and sports.

5. VISUALIZATION

The following map shows the areas in 3 different clusters depicted with 3 different colours.



5. DISCUSSIONS

The data was collected by scrapping the data into a pandas data frame, geocoding it to obtain coordinates, and then making API calls using the FourSquare API.

It was pre-processed by looking for null values and forming a data frame consisting of specific attributes needed.

Exploratory data analysis was then performed to understand the dataset better.

After all this, clustering was done on the different postal areas. The algorithm used in the project to perform clustering on the postal area dataset and to form clusters of different areas using the postal area dataset is K-Means clustering. It aims at increasing the inter-cluster distance and reducing the intra-cluster distances.

Cluster one consists majorly of hotels and restaurants followed by universities and markets or malls. Areas in this cluster sounds ideal for working professionals living alone and working full time in the metropolitan, or foodies, or professionals such as salesmen whose jobs require them to attend business dinners.

In general, if a person wishes to live in an area containing a lot of hotels or restaurants, the areas in cluster 1 are ideal.

Cluster two consists majorly of infrastructural building made for easy transportation, like metro stations or bus stations. The next most common type of venue is that of universities and then stores. Areas in this cluster appears to be ideal for university students or people who have to travel a lot intra-city.

In general, if a person wishes to live in an area consisting of transportation facilities, areas in cluster 2 are ideal.

Cluster 3’s most common venue is markets and stores. It is ideal for small business owners or shopaholics wanting to explore the culture of the city.

Areas in cluster 3 are ideal are people wishing to live in the economic booming sector or those who wish to live near markets or stores.

6. CONCLUSION

Purpose of this project was to identify the postal areas with the best infrastructure in Delhi in order to help citizens make informed decisions about moving in Delhi.

By clustering the different areas based on the types of venues and their frequencies from the FourSquare data, first the top types of places in an area are analysed. Then, after further extensive collection of locations which satisfy some basic requirements of a person moving to Delhi, infrastructure of Delhi is analysed.

Clustering of all the neighbourhoods in Delhi was performed in order to create major zones of interests.

Final decision on optimal location differs from person to person and will be made by stakeholders based on specific characteristics of neighbourhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location, real estate availability, prices, social and economic dynamics of every neighbourhood etc.