Housing and Venue Data Analysis of Delhi

1. Introduction

This report discusses the project which analyses the Venues data and Housing prices data of Delhi. Delhi is the Capital of India. It is among one of most populated cities in the world. Due to a large no. of career and business opportunities in this region, people move to this region every day. Delhi has a population of over 3 crores which makes it one of the highly crowded cities of India. Therefore, it is necessary for anyone to do the analysis of Delhi in terms of economic wellness and locality status. Analysis like this is very crucial for anyone who wants to open a new store, restaurant, shopping mall, or setup any other business. It is also helpful for those who wants to buy a property or house or even if they just want to take a look at financial wellness of a region, it can be easily done using the results of this analysis. The idea is to create maps in which economic rating can be placed on Delhi, and most of the area clustered according to the venue density.

This project creates two maps representing Economical status of an area and Category of area. The Economical status of an area is created on the basis of average housing prices per sq. ft. in that area. These maps are capable of giving you insights about the economical conditions of different parts of Delhi and category of areas in Delhi.

The problem statement is to categorize different areas in Delhi based on their economic wellness and venues in their vicinity. The venues in the vicinity of an area can be anything from an ATM to Airport. The data generated as the end product of this project must be easily understandable by non-specialists also.

2. Data Description

To consider the problem we can list the data as below:

- I found Housing Prices for different areas in Delhi on **makaan.in** website. It is a website that is used for selling and purchasing of properties in Delhi. I scraped it from their website using crawlers. I then cleaned and reduced it to get average housing prices per sq. ft. for different areas in Delhi.
- I used geocoding and Oauth API of **mapmyindia** to get latitude and longitude of areas.
- I used **Foursquare's Places API** to get venues near a particular location.

3. Methodology

Through data mining I scraped Housing Prices for different areas in Delhi from **makaan.in** website. It is a website that is used for selling and purchasing of properties in Delhi. I scraped it from their website using crawlers.

	Name	Price range per sqft	Avg price per sqft	Price rise	Trend	View Properties
0	Uttam Nagar	843 - 13,333 / sqft	7,486 / sqft	5.4%	See trend	View 468 properties
1	Uttam Nagar west	996 - 11,548 / sqft	6,762.32 / sqft	-25.7%	See trend	View 98 properties
2	Malviya Nagar	2,310 - 14,222 / sqft	11,703.7 / sqft	41.1%	See trend	View 12 properties
3	laxmi nagar	2,353 - 13,889 / sqft	11,858.86 / sqft	-16.3%	See trend	View 588 properties
4	Greater kailash 1	2,900 - 43,333 / sqft	17,797.51 / sqft	-15%	See trend	View 8 properties
5	Saket	3,435 - 18,621 / sqft	13,995.16 / sqft	-7%	See trend	View 53 properties
6	Defence Colony	758 - 35,484 / sqft	20,939.86 / sqft	21.8%	See trend	View 25 properties
7	Safdarjung Enclave	6,500 - 29,423 / sqft	21,743.02 / sqft	-20.9%	See trend	View 12 properties
8	Vasant Kunj	2,165 - 1,85,714 / sqft	68,576.87 / sqft	7%	See trend	View 248 properties
9	Greater Kailash II	3,343 - 18,765 / sqft	16,788.9 / sqft	53.4%	See trend	View 9 properties
10	Vasant Vihar	935 - 1,48,093 / sqft	48,856.77 / sqft	-24.8%	See trend	View 19 properties
11	Dwarka More	1,000 - 84,881 / sqft	50,359.58 / sqft	46.3%	See trend	View 374 properties
12	Panchsheel Park	24,385 - 92,908 / sqft	60,449.23 / sqft	39.2%	See trend	View 4 properties
13	Panchsheel Enclave	14,675 - 24,401 / sqft	19,538.06 / sqft	26.2%	See trend	View 2 properties
14	Shivalik	9,706 - 19,444 / sqft	19,444.44 / sqft	24.8%	See trend	View 2 properties
15	Sarita Vihar	841 - 96,604 / sqft	34,203.51 / sqft	16.5%	See trend	View 282 properties

After looking at the data I found that I only need the column that has name of the locality and avg price per sq. ft. Therefore, I dropped other unnecessary columns like, price range per sq. ft., trends, price rise, and view properties.

	Name	Avg price per sqft
0	Uttam Nagar	7,486 / sqft
1	Uttam Nagar west	6,762.32 / sqft
2	Malviya Nagar	11,703.7 / sqft
3	laxmi nagar	11,858.86 / sqft
4	Greater kailash 1	17,797.51 / sqft
5	Saket	13,995.16 / sqft
6	Defence Colony	20,939.86 / sqft
7	Safdarjung Enclave	21,743.02 / sqft
8	Vasant Kunj	68,576.87 / sqft
9	Greater Kailash II	16,788.9 / sqft
10	Vasant Vihar	48,856.77 / sqft

Now, I filtered our dataframe and deleted rows that does not make any sense like,

'near' will not lead us to any specific location, instead it will lead us to vicinity of a location.

Then, I am renaming the column names of the dataframe, converting 'AreaName' column into lowercase so that different cases do not cause any error in later steps and converted the column 'AvgPrice' into floating point integers for calculations because previously it was in string.

I removed more rows, those which has null values, metro, road, highway, expressway, NH, Noida, Gurgaon, Gurugram, peripheral and railway in its name because locations like metro line, highways, expressways and railways are not an area instead they can lie anywhere like some metro lines in Delhi are more than 50Km long and locations like Noida, Gurgaon and Gurugram are not in Delhi.

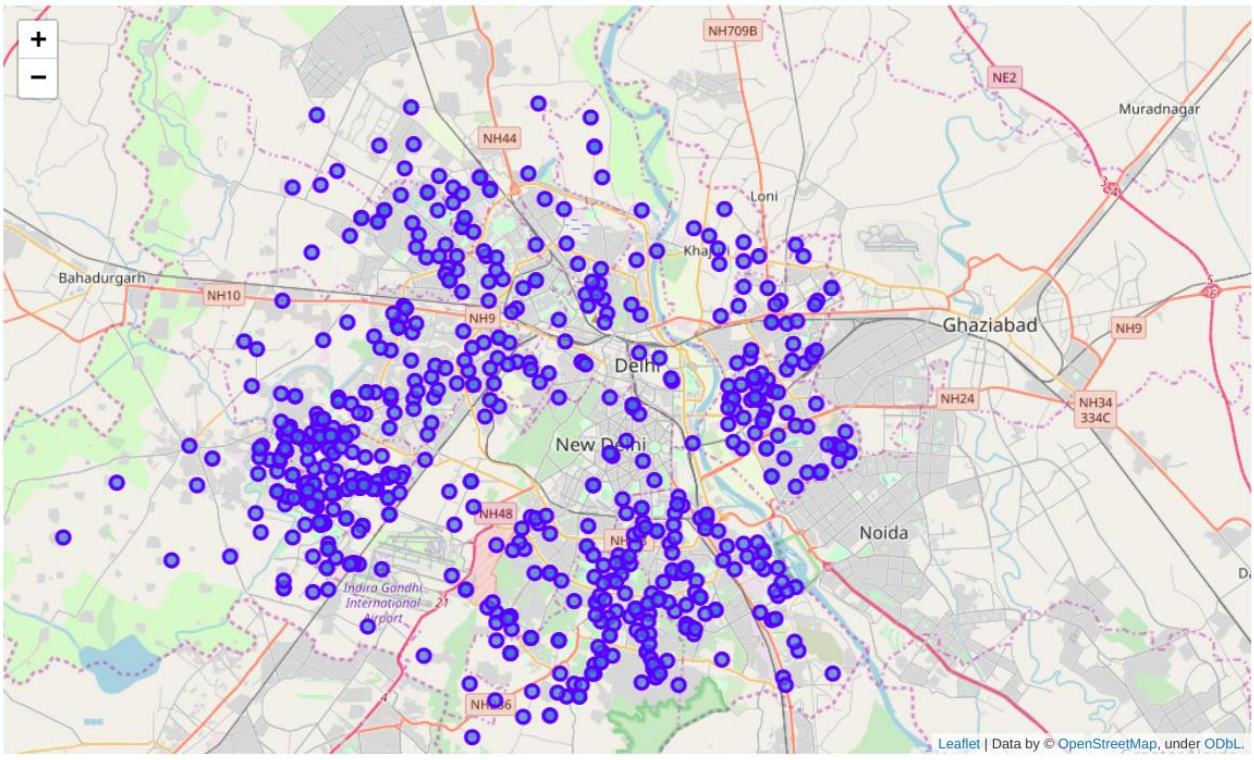
After completion of this cleaning process I received the data in following way:

	AreaName	AvgPrice
0	uttam nagar	7486
1	uttam nagar west	6762.32
2	malviya nagar	11703.7
3	laxmi nagar	11858.86
4	greater kailash 1	17797.51
5	saket	13995.16
6	defence colony	20939.86
7	safdarjung enclave	21743.02
8	vasant kunj	68576.87
9	greater kailash ii	16788.9
10	vasant vihar	48856.77

First, I was required to generate authorization header for accessing mapmyindia geocoding API. After this, I passed this authorization header to the geocoding API with location to get Latitudes, Longitudes, PIN codes and District names of locations. Then, I added this column to the existing dataframe which is as follows:

	AreaName	AvgPrice	Latitude	Longitude	PinCode	District
0	uttam nagar	7486	28.618535	77.056782	110059	West District
1	uttam nagar west	6762.32	28.618535	77.056782	110059	West District
2	malviya nagar	11703.7	28.534883	77.210245	110017	South District
3	laxmi nagar	11858.86	28.635202	77.283208	110092	East District
4	greater kailash 1	17797.51	28.548714	77.236102	110048	South East Delhi District
5	saket	13995.16	28.523848	77.206773	110017	South District
6	defence colony	20939.86	28.573199	77.232795	110024	South East Delhi District
7	safdarjung enclave	21743.02	28.565741	77.194877	110029	South District
8	vasant kunj	68576.87	28.514980	77.153245	110070	South District
9	greater kailash ii	16788.9	28.534883	77.241845	110048	South East Delhi District
10	vasant vihar	48856.77	28.564479	77.161155	110057	New Delhi District

Then, I plotted these co-ordinates fetched on the map of Delhi using folium geolocation and maps visualization library. The resulting map is accurate and also interacting, so user can zoom in and out and get info about the places.



After this, I fetched maximum 100 venues like ATM, Restaurant, etc. that lies within 1000m radius from the longitude and latitude of center of location. This data helped me in analyzing what kind of area that particular location is. To get an idea of what I have fetched I checked no. of venues fetched for each area in our database.

AreaName	
10 Dwaraka Marg	7
10 Sector Dwarka	9
A1 Block Paschim Vihar Delhi	9
Abul Fazal Enclave	6
Abul Fazal Enclave Jamia Nagar	4
Abul Fazal Enclave Part 2 New Delhi	5
Aerocity	44
Akshardham	13
Alaknanda	29
Alaknanda Gangotri Enclave	8
Ambica Vihar	4
Anand Lok	24
Anand Niketan	14
Anand Vihar	6
Antriksh Bhawan	100
Anupam Garden 10Th Lane	13
Arjun Nagar	83
Ashirwad	8
Ashok Nagar	31
Ashok Vihar	11
Ashok Vihar Phase-1	11

The fig has only few data points because it is not possible to display all the data points which are more than 700.

One Hot Encoding is a process through which we convert categorical values into numerical values. Here, in this case we are converting 100 most venues into their categorial binary value i.e. , either 0 or 1 which I used to convert venues category into binary data. Then, I calculated the frequency and sorted the data.

	AreaName	District	PinCode	Latitude	Longitude	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
0	10 Dwaraka Marg	South West District	110045	28.598530	77.087924	Department Store	Bank	Science Museum	Gym	Fast Food Restaurant
1	10 Sector Dwarka	South West District	110075	28.590611	77.057569	Fast Food Restaurant	Café	Department Store	Indian Restaurant	Electronics Store
2	A1 Block Paschim Vihar Delhi	West District	110063	28.669306	77.093265	Coffee Shop	Hotel	Food Truck	Indian Restaurant	Chinese Restaurant
3	Abul Fazal Enclave	South East Delhi District	110025	28.546370	77.307199	Fast Food Restaurant	Market	Diner	Park	Hotel
4	Abul Fazal Enclave Jamia Nagar	South East Delhi District	110025	28.555496	77.297802	Indian Restaurant	Playground	Arcade	Yoga Studio	Food
5	Abul Fazal Enclave Part 2 New Delhi	South East Delhi District	110025	28.549208	77.299861	Indian Restaurant	Park	Market	Fast Food Restaurant	Flower Shop
6	Aerocity	New Delhi District	110037	28.552016	77.121662	Hotel	Indian Restaurant	Spa	Bed & Breakfast	Hotel Bar
7	Akshardham	East District	110092	28.612332	77.278330	Dessert Shop	Restaurant	Food & Drink Shop	Bus Station	North Indian Restaurant
8	Alaknanda	South East Delhi District	110019	28.527939	77.250154	Restaurant	Coffee Shop	Convenience Store	BBQ Joint	Market
9	Alaknanda Gangotri Enclave	South East Delhi District	110019	28.524453	77.252141	BBQ Joint	Restaurant	Market	Thai Restaurant	Gym
10	Ambica Vihar	West District	110087	28.665092	77.080354	Furniture / Home Store	Asian Restaurant	Indian Restaurant	Mobile Phone Shop	Food

As data preparation for this part is completed. So, now I can cluster these areas using KMeans clustering. KMeans is a clustering algorithm that works on minimizing the intra-cluster distances and maximizing the inter-cluster distances. It is an unsupervised algorithm.

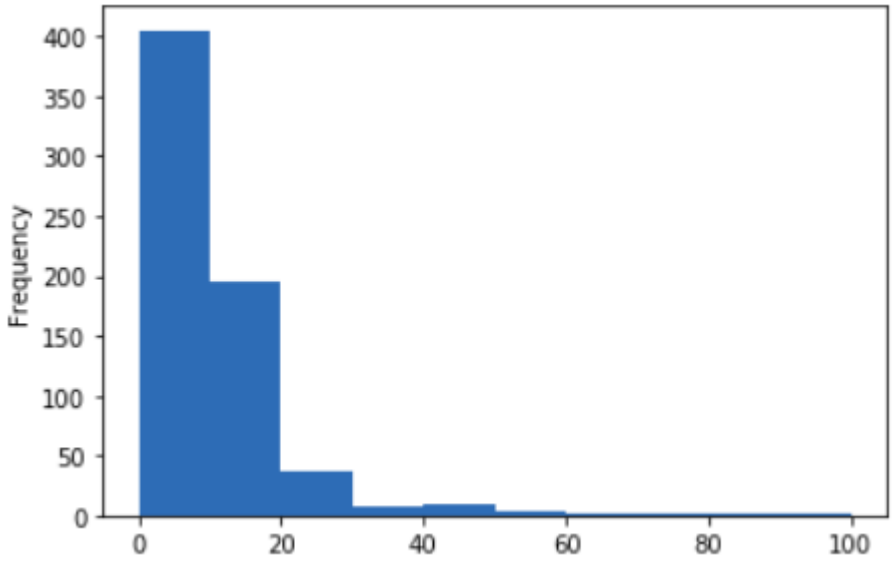
	AreaName	District	PinCode	Longitude	Latitude	Cluster	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
0	Sadar Bazaar	New Delhi District	110010	77.120554	28.591759	0	ATM	Convenience Store	Mobile Phone Shop	Café	Shoe Store
1	Tigri	South District	110062	77.238326	28.511275	0	ATM	Indian Restaurant	Electronics Store	Plaza	Fast Food Restaurant
2	Deoli Gaon Nai Basti	South District	110062	77.234013	28.505528	0	ATM	Indian Restaurant	Electronics Store	Trail	Athletics & Sports
3	Devli Extention Deoli Gaon Nai Basti	South District	110062	77.234013	28.505528	0	ATM	Indian Restaurant	Electronics Store	Trail	Athletics & Sports
4	Devli Nai Basti	South District	110062	77.234013	28.505528	0	ATM	Indian Restaurant	Electronics Store	Trail	Athletics & Sports
5	Rajinder Nagar	North East District	110094	77.287433	28.703218	0	ATM	Indian Restaurant	Pizza Place	Fabric Shop	Fried Chicken Joint
6	Bindapur	South West District	110059	77.068586	28.610709	0	ATM	Indian Restaurant	Pool	Mobile Phone Shop	Flower Shop
7	Sitapuri	South West District	110059	77.075381	28.607997	0	ATM	Pool	Market	Business Service	Health & Beauty Service
8	Gulabhi Bagh Om Vihar	West District	110059	77.052640	28.626044	0	ATM	Train Station	Bakery	Mobile Phone Shop	Department Store
9	Om Vihar	West District	110059	77.052640	28.626044	0	ATM	Train Station	Bakery	Mobile Phone Shop	Department Store

Now, I have normalized housing dataframe to create a rating scale from (0-100) for housing prices.

I normalized the dataset using Min-Max Method that scales values from 0 to 1 making it easier to understand and work on data.

	AreaName	AvgPrice	Latitude	Longitude	PinCode	District	NormPrice
0	Uttam Nagar	7486.00	28.618535	77.056782	110059	West District	7.169510
1	Uttam Nagar West	6762.32	28.618535	77.056782	110059	West District	6.470488
2	Malviya Nagar	11703.70	28.534883	77.210245	110017	South District	11.243497
3	Laxmi Nagar	11858.86	28.635202	77.283208	110092	East District	11.393370
4	Greater Kailash 1	17797.51	28.548714	77.236102	110048	South East Delhi District	17.129667
5	Saket	13995.16	28.523848	77.206773	110017	South District	13.456878
6	Defence Colony	20939.86	28.573199	77.232795	110024	South East Delhi District	20.164945
7	Safdarjung Enclave	21743.02	28.565741	77.194877	110029	South District	20.940739
8	Vasant Kunj	68576.87	28.514980	77.153245	110070	South District	66.178781
9	Greater Kailash Ii	16788.90	28.534883	77.241845	110048	South East Delhi District	16.155424
10	Vasant Vihar	48856.77	28.564479	77.161155	110057	New Delhi District	47.130620
11	Dwarka More	50359.58	28.618973	77.031514	110059	South West District	48.582223
12	Panchsheel Park	60449.23	28.543272	77.215076	110017	South District	58.328080
13	Panchsheel Enclave	19538.06	28.543794	77.229476	110017	South District	18.810910
14	Shivalik	19444.44	28.532298	77.206525	110017	South District	18.720480
15	Sarita Vihar	34203.51	28.530896	77.294073	110076	South East Delhi District	32.976652

After this, I created bins and generated a histogram.



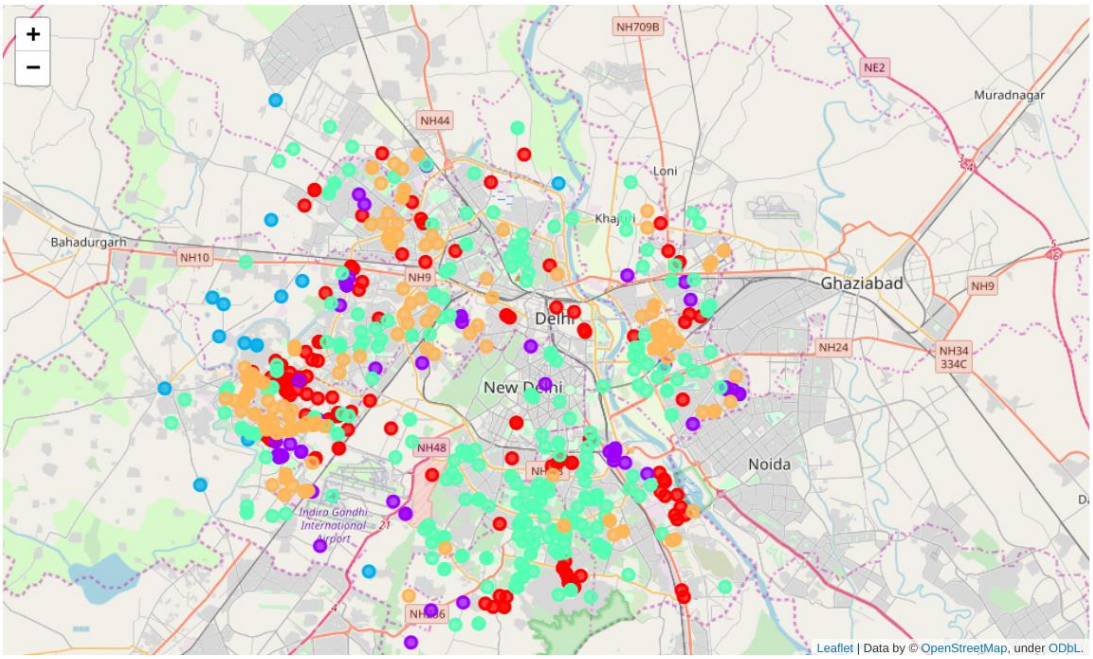
So I assigned categories to the areas based on their price ratings.

Categories:

- Rating(0-20)=> Category I
- Rating(21-40)=> Category II
- Rating(41-60)=> Category III
- Rating(61-80)=> Category IV
- Rating(81-100)=> Category V

4.Results

It is now time to see what I have obtained and which areas are better for which purpose. Here, I am using folium which is python library for visualizing geo-spatial data.



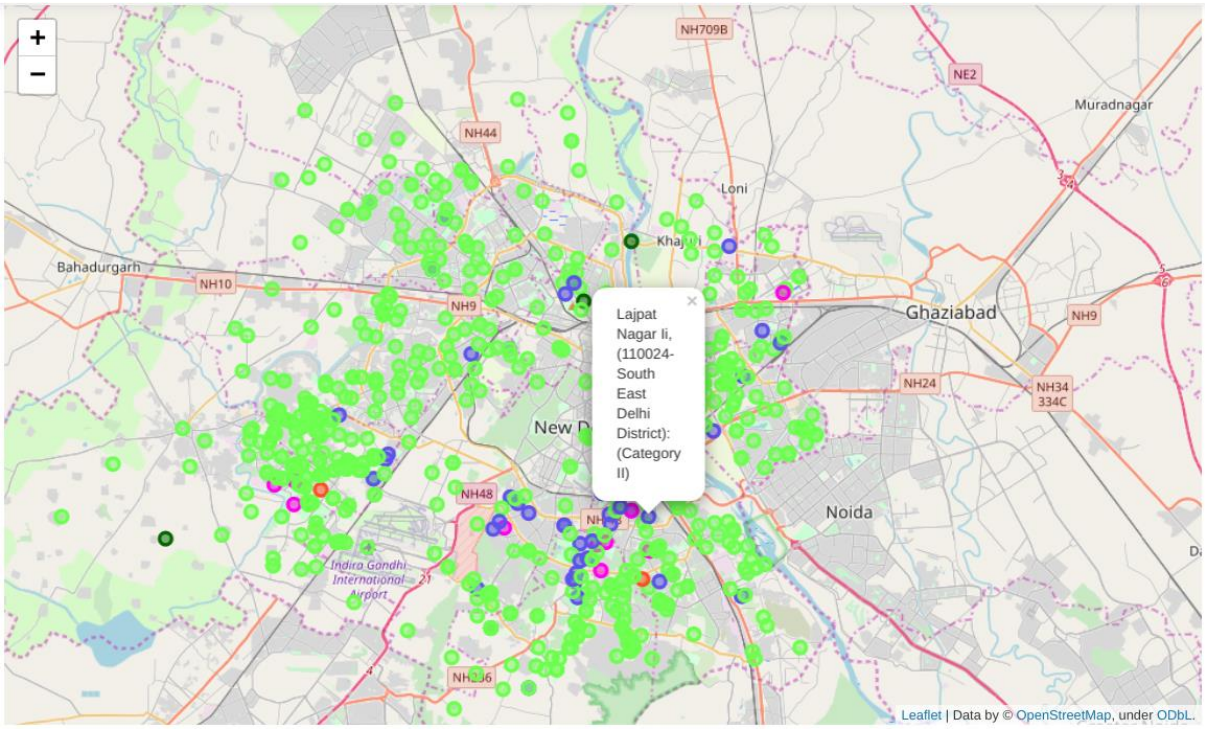
The Results of this clustering are as follows:

- Cluster-0 (tomato) has higher no. of ATMs, Indian Restaurants, Markets, Shops. These are the locations where people go for shopping or hangout.
- Cluster-1 (mediumpurple) has higher no. of Hotels, Asian Restaurants, Gyms and Neighborhoods. So these are residental areas.
- Cluster-2 (mediumtorquoise) has higher no. of Gardens and parks. These areas are more green than other areas.
- Cluster-3 (aquamarine) has airports and also some high end facility providers. These areas are downtown and posh areas.
- Cluster-4 (burlywood) has higher no. of metro stations, pizza places and other businesses. These areas are well connected areas by metros.

Results from the housing price analysis is that:

Higher the prices of houses higher is the chance that people with better economic conditions live in that area.

People living in area of Category V has highest economic strength and people living in area of Category I has lowest economic strength.



5. Discussion

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

6.Conclusion

The maps generated through this project are useful for the vast majority of people. Many countries and companies generate these kinds of maps, which help them in deeply analyzing the present conditions of the state. There are so many companies that are democratizing data like this, which gives insights for businesses. Geographical and Economic insights are a boon for everyone, whether it is a big company or an individual.

For more detailed and accurate guidance, the data set can be expanded and the details of the area or street can also be drilled.

I also performed data analysis through this information by adding the coordinates of districts and home sales price averages as static data on makaan.in. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data and clustering information on the Delhi map. In future studies, web or telephone applications can be carried out to direct investors.