Coursera Data Science Capstone Project

Clustering and Neighbourhood Analysis – Hyderabad

By: Angam Praveen(Aug 2019)

1. Introduction

1.1 Background

Hyderabad is the capital city of the newly formed state of Telangana (India) after its separation from Andhra Pradesh. Enclosing over 650 square kilometres along the banks of the Musi River, Hyderabad is home for about 9.7 million people, 6th most populous urban area in the southern region of the Indian sub-continent. Over the last two decades, there has been significant increase in the economic activity with the rise in numerous service industries especially IT sector. Other types of employment include state and central government organizations. This growth has boosted economic activities in other sectors like trade and commerce, transport, storage, communication, real estate and retail. One such sector which is very dynamic by its nature and also a booming sector is real estate. Every year large numbers of people invest considerable amount of money in real estate for residential as well as an investment with the hope of making a good ROI. Investing in real estate, especially in a residential property, needs a great deal of analysis of the different areas of the city by taking into consideration a variety of factors like Supermarkets, Pharmacy, Hospitals, Transportation, Water facilities, Environment, Parks and other recreational venues.

1.2 Problem

Considering all the different factors that contribute to the identification of areas with best facilities and good ROI a data driven analysis of the areas within the city would greatly enhance in making insightful decisions in real estate investment. Primarily this particular project aims at analysing various localities which are similar in nature with respect to different venues in the area and their prices. Secondly we cluster neighbourhoods in the city based on similarities in different aspects and identify the best places of investment in a residential property.

1.3 Interest

The following analysis would be valuable for individual buyers and investing agents in residential property in the city. This would also help builders in understanding potential market areas for a profitable investment in residential category. Other group of people/institutions who would benefit from the analysis are transport agents, traders, local vendors and retail etc.

2. Data

2.1 Data Sets

Following data sets are used to perform the analysis and build the model:

- **Residential property prices**: Data related to the prices of residential property in the city are obtained from 99acres.com (an online application for buying/selling properties) which contains list of different areas in a metropolitan city and their corresponding price ranges.
- **Nominatim API**: It is a tool used to search through OpenStreetMap data by name. This data is used to fetch the coordinates (Latitude and Longitude) of different neighborhoods of the city which are used for analysis of the prices.
- **Foursquare API**: To get the details of various venues and facilities in the neighbourhoods of the city, data from Foursquare API is used. Using Folium and Foursquare data, areas are explored on maps and the venues are analysed and used to group areas into different categories and best areas of interest are identified.

2.2 Data Usage

Firstly, data from the online real estate site 99acres.com (https://www.99acres.com/property-rates-and-price-trends-in-hyderabad) is used to get the locality Name, Buying Price and rental prices of different areas of the city. For example one of the areas in the list, *Banjara Hills* is a highly developed area with prices of the place ranging from Rs. 110,160 – 137,318 per Square Yard. Secondly, pricing data loaded from previous step is used to visualize which areas of the city are more costly compared to others. Using the Nominatim API (https://nominatim.openstreetmap.org), geo-coordinates of both city and various areas within the city are fetched and mapped on to Hyderabad map to visualise the spread and look for similarities between different areas to form clusters. For example, Hyderabad has co-ordinates of latitude and longitude with 17.389 and 78.461 respectively. Similarly details of other areas in the city are extracted using the API. Thereafter, various venue details from Foursquare API are used to explore the neighbourhoods of the city. Foursquare helps in exploring venues in a place with detailed listing of the venue categories (restaurants, cafes, parks, studios etc), its location, reviews, users, user details etc. Areas with similar characteristics are used to group together using clustering algorithms and then labelled under different categories. Pricing data is also considered as feature for clustering the areas into different groups. These clusters are plotted on to Maps using python package Folium for better visual understanding.

3. Methodology

3.1 Property Prices Data

From the website it can be observed that the webpage includes both Residential Apartments as well as Residential Land under Localities in Hyderabad, we will be focusing on Residential Land for investment. Details of the data consist of area wise maximum, minimum prices, trends over each Quarter, percentage increase or decrease in the costs. We will only focus on the max and min prices of the property due to unavailability of complete data for other categories of information in the webpage.

From the webpage, columns of Locality, Price range are copied to local file and saved in .csv format. Data is cleaned by separating the minimum and maximum prices from the price range column. Saved data is then uploaded to the workspace in the cloud under datasets category. Sample data is shown below:

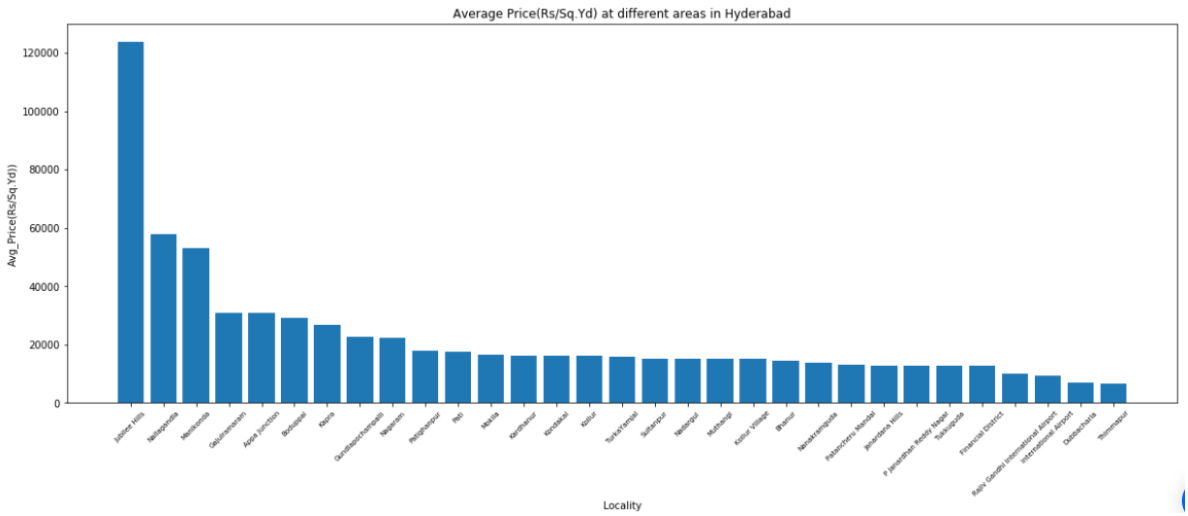
	Locality	Price_Min_Rs_per_SqYd	Price_Max_Rs_per_SqYd
0	Appa Junction	26392	35190
1	Bhanur	13005	16065
2	Boduppal	26775	31365
3	Dubbacharla	6885	6885
4	Financial District	12622	12622

We then calculate the average/mean price of property for ease of analysing the price in different areas. Mean price is calculated using mean function and the values are assigned to a new column 'Avg_Price_Rs_per_SqYd' as shown below:

	Locality	Price_Min_Rs_per_SqYd	Price_Max_Rs_per_SqYd	Avg_Price_Rs_per_SqYd
0	Appa Junction	26392	35190	30791.0
1	Bhanur	13005	16065	14535.0
2	Boduppal	26775	31365	29070.0
3	Dubbacharla	6885	6885	6885.0
4	Financial District	12622	12622	12622.0

3.2 Property Price Analysis

For better understanding of the data visually we plot a bar chart showing the average price against different areas of the city. The prices are arranged in the decreasing order of the average price as displayed below:



It can be seen that Jubliee Hills has the highest average price per SqYd. While this is known fact that Jubliee Hills is a high quality posh area within the city with maximum prices reaching unusually high, the statement is further reinforced with this data driven display. Furthermore interestingly Nallagandla is at the 2nd place in terms of average price which was not so in the recent past.

3.3 Coordinates Data

In order to analyse the localities using information of the venues in these areas, we need Latitude and Longitude values of the areas. While some of the coordinates are fetch from Nominatim Open Street Map(https://nominatim.openstreetmap.org) few of them are fetch from Google Maps due to non-availability of the data in Nominatim webpage.

Data from these sites are fetched manually and written to a .csv file. The file is then uploaded into the IBM cloud workspace under datasets category. Coordinates from the file are extracted locality wise and loaded into a dataframe. Below table displays the sample coordinates.

	Locality	Latitude	Longitude
0	Appa Junction	17.349710	78.360158
1	Bhanur	17.478825	78.169449
2	Boduppal	17.416539	78.580225
3	Dubbacharla	17.107689	78.370912
4	Financial District	17.412001	78.343023

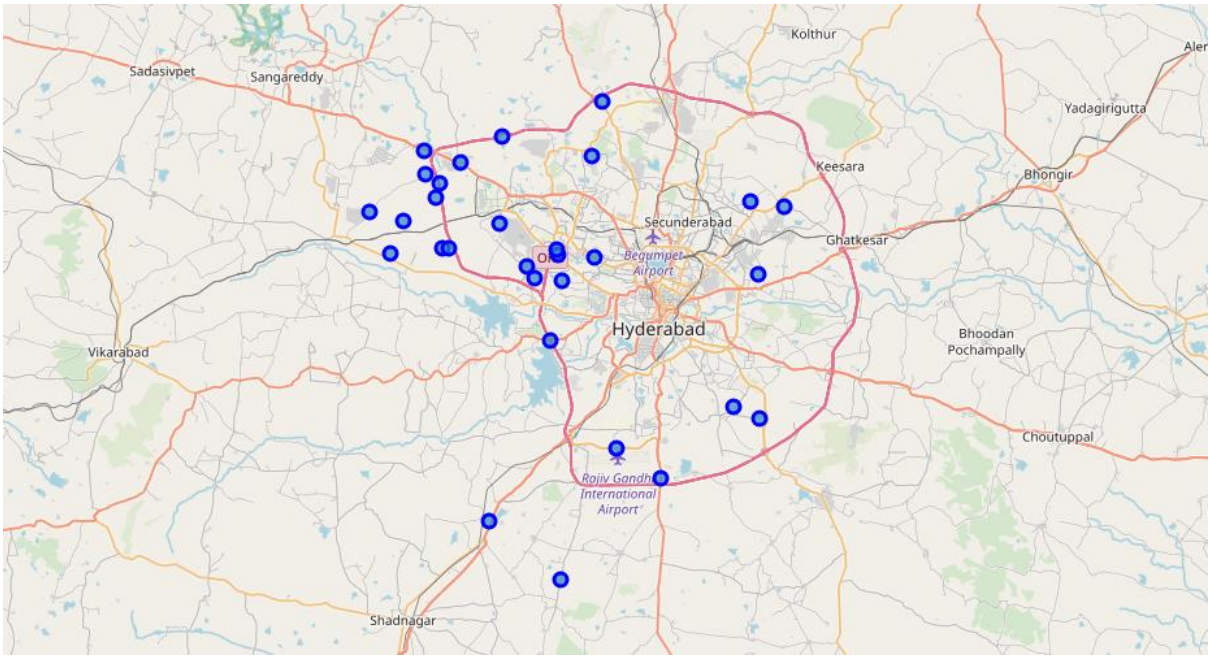
3.4 Data Cleaning

It is observed that there are duplicate entries of the Locality due to different names being used for the same Locality like 'International Airport' has different entry though there is

similar Locality which is 'Ragiv Gandhi International Airport' which are one and the same. Hence we remove one entry from the data.

3.5 Localities on Map

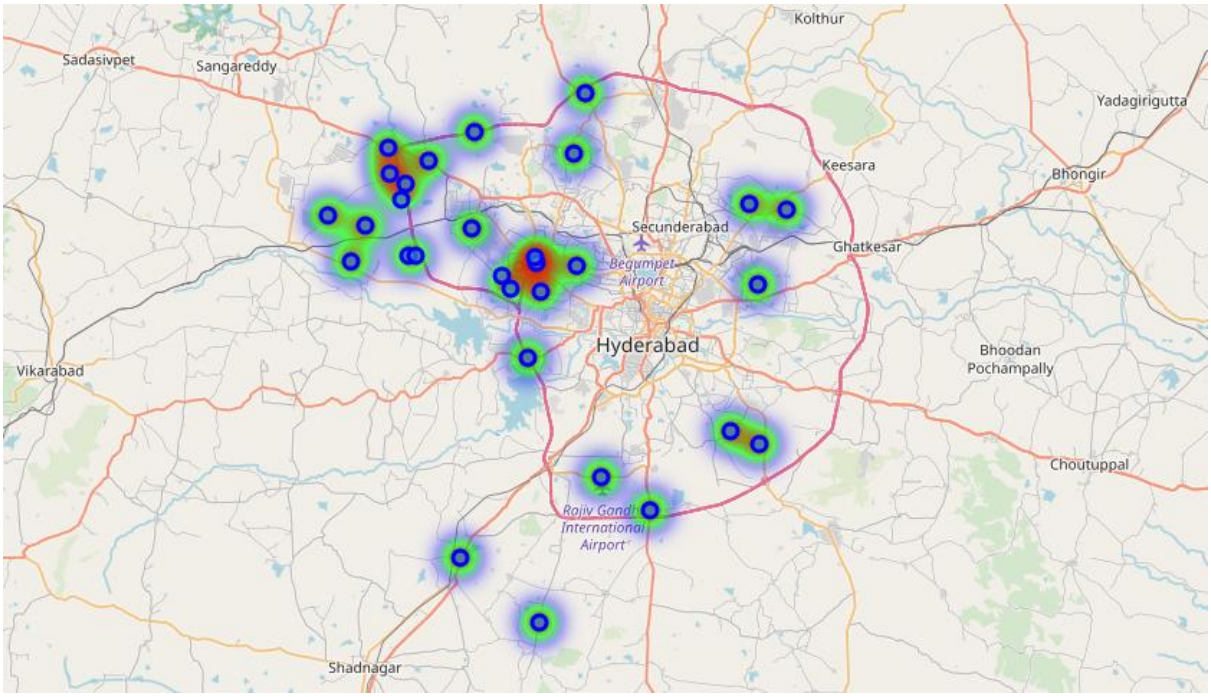
Using the latitude and longitude details of different areas in the dataframe we plot the coordinates using a Folium Map centred at Hyderabad coordinates. We mark the points using different attributes of Folium.Circlemarker. Using Folium.popup we provide hints which display the area name and the city name upon hovering over the data points in the map as below.



From the previous plot it is clearly evident that much of the investor interest is in the West and North-West part of the city with majority of the price listings from the Localities concentrated in this region of the city. Possible reason for this could be due to numerous Industries which cover the North Western region and large number of IT companies in the Western region. Yet another interesting observation is that much of the points are adjacent to and within the Outer Ring Road which is an expressway encircling the city.

3.6 Average Price Heat Map

In order to visualize the price ranges geographically we now embed the average prices in these areas over the map in the form of a Heat map which best explains the weight/intensity of a variable on a map. Below displays the heat map of the average prices in different areas.



Complimenting to our earlier observation it is notable that the prices of the land in the West/North-West areas are high. Thus it is understood that West/North-West regions are the most happening regions in the city for investment.

3.7 Venues Extraction

Each area around the Localities within the radius of 1000 meters is explored for fetching the top 100 venues using Fourquare API. Inputs for the API include radius, number of venues, Latitude and Longitude. The fetched venues are merged with the original dataframe. Below table lists the sample data.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Jubilee Hills	17.432707	78.407197	Cream Stone Concepts	17.430823	78.408039	Ice Cream Shop
1	Jubilee Hills	17.432707	78.407197	Conçu Cakes	17.431740	78.408065	Cupcake Shop
2	Jubilee Hills	17.432707	78.407197	Lasermaxx	17.431772	78.406630	Gaming Cafe
3	Jubilee Hills	17.432707	78.407197	Kavanah Patisserie & More	17.433271	78.408364	Deli / Bodega
4	Jubilee Hills	17.432707	78.407197	Di Bella Coffee	17.430678	78.408683	Coffee Shop

3.8 Top 10 most common venues extraction

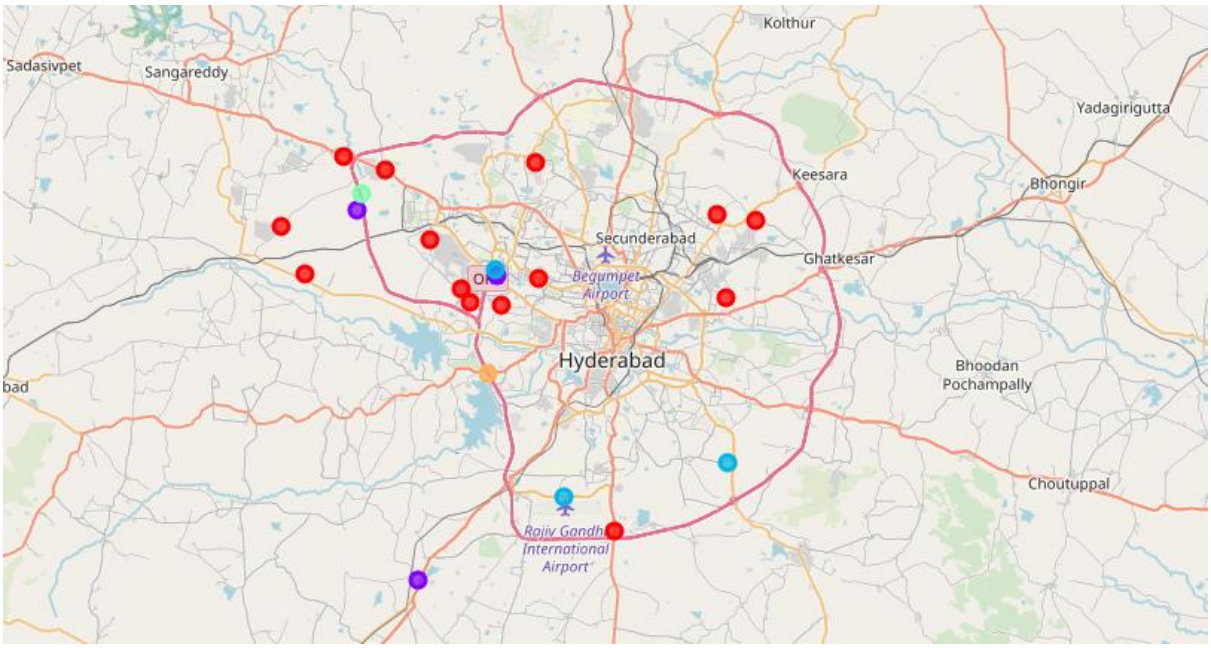
After venues are extracted from the previous step, different venues are group by each Locality to check for the most common venues surrounding each area. These details are captured in a dataframe constituting the top 10 most common venues in each area sample of which is displayed below.

	Neighborhoods	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
0	Appa Junction	Athletics & Sports	Vegetarian / Vegan Restaurant	Deli / Bodega	Clothing Store	Cocktail Bar	Coffee Shop	Comfort Food Restaurant	Convenience Store	Cupcake Shop	Dairy Store
1	Bhanur	Indie Movie Theater	Vegetarian / Vegan Restaurant	Furniture / Home Store	Clothing Store	Cocktail Bar	Coffee Shop	Comfort Food Restaurant	Convenience Store	Cupcake Shop	Dairy Store
2	Boduppal	Dairy Store	Pharmacy	Pizza Place	Vegetarian / Vegan Restaurant	Chinese Restaurant	Clothing Store	Cocktail Bar	Coffee Shop	Comfort Food Restaurant	Convenience Store
3	Financial District	Café	Food Court	Hotel	Spa	South Indian Restaurant	Coffee Shop	Pizza Place	Dairy Store	Chinese Restaurant	Clothing Store
4	Gajulramaram	Multiplex	Café	Chinese Restaurant	Clothing Store	Cocktail Bar	Coffee Shop	Comfort Food Restaurant	Convenience Store	Cupcake Shop	Dairy Store

3.9 K-Means Clustering

In order to understand the areas which are similar in nature with respect to venues we will be clustering neighbourhoods using K-Means Clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. Using the clustering process, one can choose areas which are similar in nature for investing taking into consideration the prices of the property in different areas. In this scenario post investigating different sizes, size of 5 best categorizes the areas in 5 groups.

In this process one hot encoding of the venues in each of the areas is performed, frequency is calculated and normalized in order to feed numeric data to train the model. Post clustering details of the different groups are displayed with different colour in the map as below.



4. Results and Discussions

We now try to understand the clusters and come up with suggestions based on individual choice of the surroundings of each area.

Cluster 0 as seen from the Cluster Map is distributed across the city. Relating these observations with the prices map, property prices of these areas are in the normal range. Areas under this cluster have many of the amenities typically required for urban living like ATM, Restaurant, Supermarket, Pharmacy and Cloth Stores. These areas could be an ideal choice if someone wants to buy a property and immediately move in. Except for few areas in this cluster most of the areas are within the Outer Ring Road thus ensuring minimal distance from the heart of the city. Below displays some of the areas in Cluster 0.

	Locality	Price_Min_Rs_per_SqYd	Price_Max_Rs_per_SqYd	Avg_Price_Rs_per_SqYd	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Jubilee Hills	110160	137318	123739.0	17.432707	78.407197	0	Indian Restaurant	Café	Coffee Shop	Brewery	Lounge
19	Financial District	12622	12622	12622.0	17.412001	78.343023	0	Café	Food Court	Hotel	Spa	South Indian Restaurant
18	Tukkuguda	12622	12622	12622.0	17.210545	78.476297	0	Movie Theater	Pharmacy	Restaurant	Cafeteria	Café
15	Patancheru Mandal	11858	14535	13196.5	17.528824	78.265740	0	ATM	Shipping Store	Breakfast Spot	Clothing Store	Cocktail Bar
14	Nanakramguda	12622	14918	13770.0	17.423959	78.334966	0	Hotel	Sandwich Place	Coffee Shop	Gym	Cafeteria
13	Bhanur	13005	16065	14535.0	17.478825	78.169449	0	Indie Movie Theater	Vegetarian / Vegan Restaurant	Furniture / Home Store	Clothing Store	Cocktail Bar
12	Muthangi	13388	16830	15109.0	17.540302	78.227444	0	Snack Place	Café	Clothing Store	Cocktail Bar	Coffee Shop

Cluster 1 has areas which are covered with few restaurants and Coffee Shops. Property cost in these areas is not as high as that of Cluster 1 which has Jubilee Hills, Nallagandla etc which are very costly. Owing to the limited facilities in this cluster, nominal costs in these areas can be understood.

	Locality	Price_Min_Rs_per_SqYd	Price_Max_Rs_per_SqYd	Avg_Price_Rs_per_SqYd	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Mo Comm Ven
9	Pati	16830	18360	17595.0	17.493985	78.238534	1	ATM	Deli / Bodega	Clothing Store	Cocktail Bar	Coffee Shop	Comf Fo Restaur
16	Janardana Hills	12622	12622	12622.0	17.435982	78.368646	1	Indian Restaurant	Fast Food Restaurant	Department Store	Asian Restaurant	Coffee Shop	Keb Restaur
21	Thimmapur	6120	7268	6694.0	17.167167	78.295430	1	Train Station	Indian Restaurant	Fried Chicken Joint	French Restaurant	Clothing Store	Cock

Cluster 2 is filled with areas around the Airport where facilities are limited. Also with respect to cost, properties in this area are less costly compared to others.

	Locality	Price_Min_Rs_per_SqYd	Price_Max_Rs_per_SqYd	Avg_Price_Rs_per_SqYd	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Comm Ve
20	Rajiv Gandhi International Airport	9562	10328	9945.0	17.240601	78.429535	2	Café	Vegetarian / Vegan Restaurant	Airport Lounge	Fast Food Restaurant	Bookstore	In Restau
11	TurkaYamjal	13770	17978	15874.0	17.270651	78.580953	2	ATM	Deli / Bodega	Clothing Store	Cocktail Bar	Coffee Shop	Com F Restau
17	P Janardhan Reddy Nagar	12622	12622	12622.0	17.441167	78.366652	2	Indian Restaurant	Fast Food Restaurant	Ice Cream Shop	Sandwich Place	Department Store	H

Cluster 3 and Cluster 4 have limited amenities for living in these areas. Cluster 4 which contains Appa junction is comparatively costly than other areas in Cluster 3 due to its proximity to IT corridor and Water Bodies. Appa Junction can also be considered as a suitable place for investment owing to its nominal cost and facilities.

	Locality	Price_Min_Rs_per_SqYd	Price_Max_Rs_per_SqYd	Avg_Price_Rs_per_SqYd	Latitude	Longitude	Cluster Labels	1st Most Common Venue	[illegible] Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
8	Patighanpur	16830	18742	17786.0	17.50746	78.243907	3	Movie Theater	Furniture / Home Store	Chinese Restaurant	Clothing Store	Cocktail Bar	Coffee Shop

	Locality	Price_Min_Rs_per_SqYd	Price_Max_Rs_per_SqYd	Avg_Price_Rs_per_SqYd	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
4	Appa Junction	26392	35190	30791.0	17.34971	78.360158	4	Athletics & Sports	Vegetarian / Vegan Restaurant	Deli / Bodega	Clothing Store	Cocktail Bar	Coffee Shop

5. Conclusion and Future Work

This project helps anyone interested in investing in residential properties of Hyderabad by giving necessary overview of the best localities based on individual needs. In combination with the prices and clusters of similar venues this analysis best helps in decision making for investment. This work can be further improved by including features related to population densities, industrial locations, roadways and water body proximity etc for better and holistic decision making.