IBM APPLIED DATA SCIENCE CAPSTONE PROJECT
FINAL REPORT
HYDERABAD CITY: LOCALITY RECOMMENDATION SYSTEM

Source: The New Indian Express

COURSERA CAPSTONE PROJECT
ARJUN GIDWANI
JUNE 2020

INTRODUCTION:

Hyderabad

(Source: https://hyderabad.telangana.gov.in/about-district/**):**

Hyderabad is the capital of India's most tech-savvy state, Telangana. This city is also known as the "City of Pearls" due to its history of being a global centre of trade in rare diamonds, emeralds and most importantly, fresh water pearls. Hyderabad was founded in 1951 and was planned as a grid with the Charminar at the centre. It has grown beyond the confines of the original walled city, to include a new town, north of the Musi river, the Military Cantonment at Secunderabad and the famous high-tech estate nicknamed, "Cyberabad".

Hyderabad district is a city-district in the state of Telangana which includes a part of the metropolitan capital city area of Hyderabad. It is the smallest in terms of area among all the districts in the state (217sq.km) but has the highest human density. According to the 2011 census, the population of this district was 3,943,323.

Greater Hyderabad Municipal Corporation (GHMC) and the Demography of Hyderabad:

(Source: https://worldpopulationreview.com/world-cities/hyderabad-population/)

The Greater Hyderabad Municipal Corporation (GHMC) is the civic body that oversees Hyderabad. It was created in 2007 to oversee the "18 circles" of the city. This increased the area of Hyderabad to approx. 650 sq.km and the population grew by 87%. The GHMC has a population of approx. 10 million which makes it the 6th most populous urban agglomeration in India. It is the local government for the cities of Hyderabad and Secunderabad. At present the GHMC is spread across 4 districs – Hyderabad District, Medchal District, Ranga Reddy District and Sanga Reddy District.

Most of the Hyderabad population comprises Telugu and Urdu speaking people. The minority communities include Tamil, Marathi, Kannada, Marwari, Gujarati, Keralites, Punjabis and communities from Uttar Pradesh. The foreign population in Hyderabad comprises Hadhrami Arabs (of Yemeni descent), African Arabs, Armenians, Iranians, Pathans and Turkish people.

As per the 2011 census, 24% of Hyderabad residents, were migrants from other parts of India.

Understanding the population growth:

As Hyderabad is one of the fastest growing metropolitan areas in India, it faces issues in terms of employment, housing and essential services. There has been a 264% increase in Hyderabad's slum population bringing it up to 30% in 2014. This has been attributed to two factors:

- Inefficient Urban Planning
- Greater rural-to-urban migration

Problem Description:

The objective of this project is to analyse and select the best possible localities in the city of Hyderabad for a contract working executive to live in/relocate to.

A challenge faced by Sindhi working executives relocating to Hyderabad for 1-2 year projects from Mumbai, Ahmedabad and Pune living with family on rent in Hyderabad, is finding the right locality to live in based on the population density, the rental rates and the availability of specific conveniences in the locality. The committee members of the Sindhi Community Centre of Hyderabad approached me to find a solution for this problem.

Using Data Science methodology and ML techniques like clustering, the project aims to provide the solutions to the question:

If a Sindhi working executive is planning to relocate to Hyderabad, based on his specific requirements , in which locality would you recommend he shift to?

Target Audience:

While the project is designed to answer the question put forward by the Sindhi Community Centre, the target audience of this project could be any working executive in particular, either looking to shift to a rented house in a different locality in Hyderabad with family or looking to relocate to Hyderabad with family for a short period of up to 2 years.
In the case of this project, the additional parameter (of specific convenience) that will be factored in is that of the presence of vegetarian restaurants in the locality. This is being undertaken as most of the Sindhi working executives with families relocating to Hyderabad and approaching the Sindhi Community Centre for guidance, are vegetarian and are particular about relocating to areas where there is a reasonable presence of vegetarian restaurants.

DATA:

Data Required:

1. A list of the Zones, Circles and Wards which come under the Greater Hyderabad Municipal Corporation.
2. The Coordinates (The Latitudes and Longitudes) of each ward (equivalent of a neighbourhood). This is required particularly to plot the map and to get the venue data using the Foursquare API.
3. Venue Data
4. Ward-wise average rent data for 2BHK
5. Population Data for each Ward

These are important for clustering and setting the parameters of the recommender model.

Sources of Data:

The GHMC Zones, Circles and Wards List:
https://www.ghmc.gov.in/Documents/GHMC_Circles.pdf
https://www.ghmc.gov.in/Documents/Wards.pdf
The lists from the GHMC website have been manually combined due to the lack of reliable open source list.

The Coordinates of each ward:
Due to the unavailability of open source data on geographical coordinates on Wikipedia of government websites, this data was manually collected for each using google maps to produce the dataframe for the same.

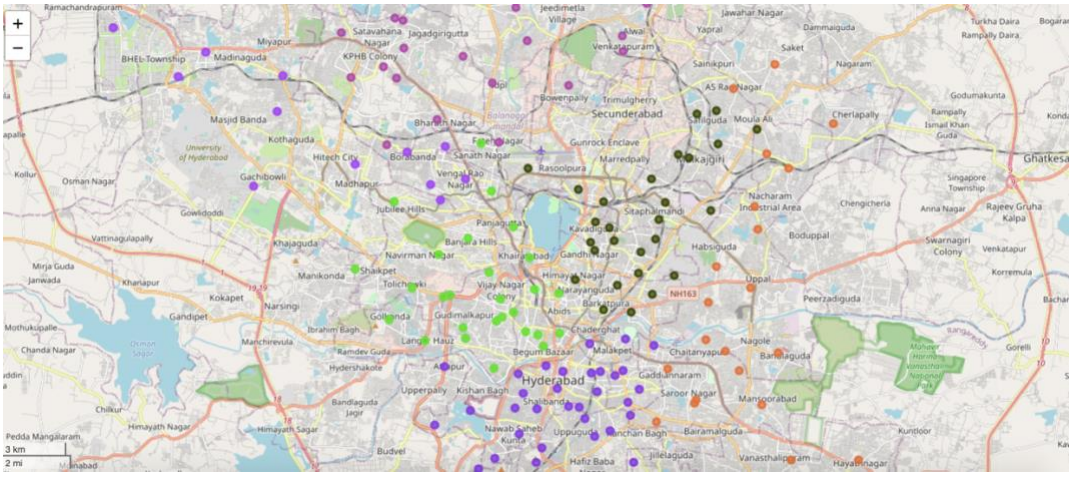
	Zone	Ward	Latitude	Longitude
0	LB Nagar (East Zone)	Kapra	17.488731	78.568384
1	LB Nagar (East Zone)	Dr. AS Rao Nagar	17.479077	78.550862
2	LB Nagar (East Zone)	Cherlapally	17.464962	78.593184
3	LB Nagar (East Zone)	Meerpet HB Colony	17.452821	78.564846
4	LB Nagar (East Zone)	Mallapur	17.446885	78.574279

Venue Data:

Foursquare API to get the Venue Data:

	Ward	Zone	Ward Latitude	Ward Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kapra	LB Nagar (East Zone)	17.488731	78.568384	MedPlus	17.488267	78.569265	Pharmacy
1	Dr. AS Rao Nagar	LB Nagar (East Zone)	17.479077	78.550862	The Coffee Cup	17.483180	78.552104	Café
2	Dr. AS Rao Nagar	LB Nagar (East Zone)	17.479077	78.550862	Swagath Grand	17.482022	78.553261	Indian Restaurant
3	Dr. AS Rao Nagar	LB Nagar (East Zone)	17.479077	78.550862	Ushodaya Supermarket	17.482001	78.553056	Department Store
4	Dr. AS Rao Nagar	LB Nagar (East Zone)	17.479077	78.550862	Vodafone Store	17.482264	78.552923	Electronics Store

Map marking the Wards/Neighborhoods:



Ward-wise Avg. Rent Data:
Main Data: https://www.makaan.com/price-trends/property-rates-for-rent-in-hyderabad
The data which could not be found in this list were manually added to the list be searching on the following sites:

- housing.com
- magicbricks.com

Ward-Wise Population Data Source:
https://indikosh.com/city/708723/greater-hyderabad

METHODOLOGY:

1. Creating the main data sheet:

Step 1: We need to get the list of neighbourhoods / wards in the city of Hyderabad. The GHMC lists below are the only credible, updated and reliable sources online for the list of zones, wards and circles in the city of Hyderabad.
https://www.ghmc.gov.in/Documents/GHMC_Circles.pdf
https://www.ghmc.gov.in/Documents/Wards.pdf
The list has been manually compiled from the GHMC sources as they could not be scraped owing to the formats and lack of reliability of other open sources.

Step 2: We need to get the geographical coordinates of the wards in the list. To do this, we have the option of using the Google API. However, owing to the limitations of usage, this list has been compiled manually using Google Maps. This has helped in ensuring accuracy of data as coordinates of Wards, spelt differently from their official names (official names are on the GHMC list) have also been correctly captured.

Step 3: We need the population data and the average rent data for each ward. The population data has been retrieved manually using the data in the following link:
https://indikosh.com/city/708723/greater-hyderabad

The population data has been given according to ward numbers (which can be found on the GHMC list). The final data has been compiled referencing both lists.

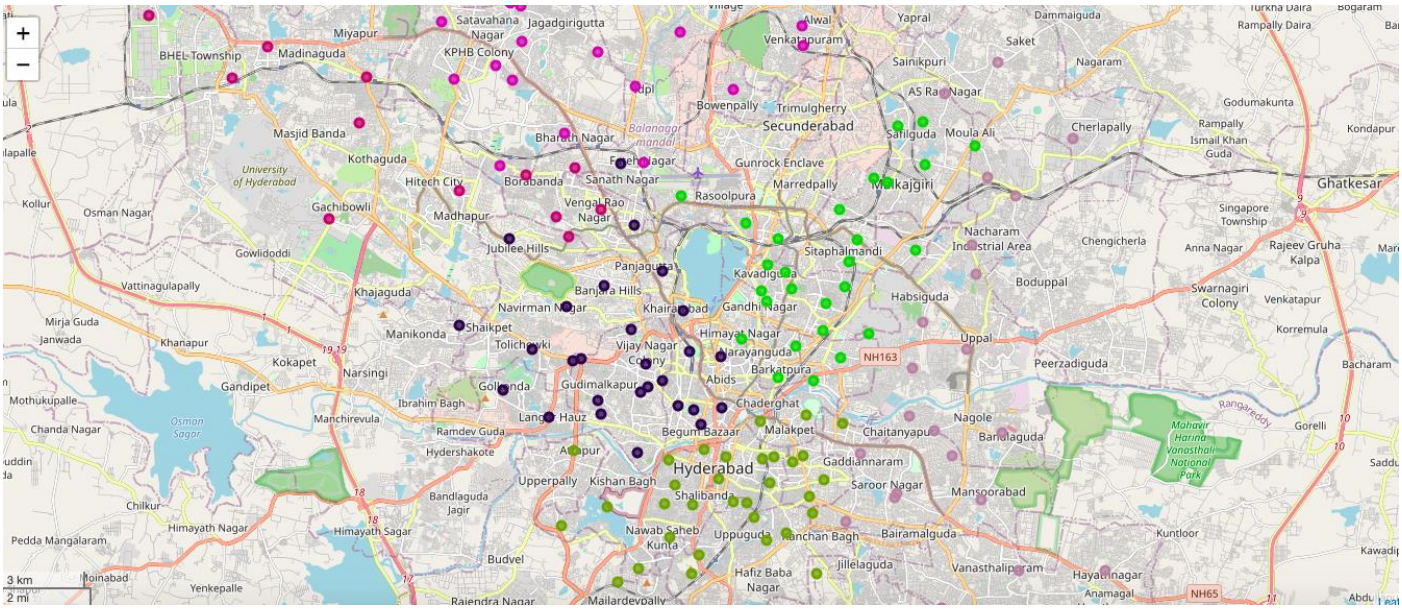
The average rent data list has been manually created using 2bhk rent data in the following link:
https://www.makaan.com/price-trends/property-rates-for-rent-in-hyderabad

As the data was unavailable for all the wards, the data for the missing wards was compiled from namely two sites:

- housing.com
- magicbricks.com

We "read" the final csv file into the program in order to do an exploratory analysis of the data and create the recommendation system.

We also create a folium map using the coordinate data to visualize the Zones in Hyderabad with their respective Wards.

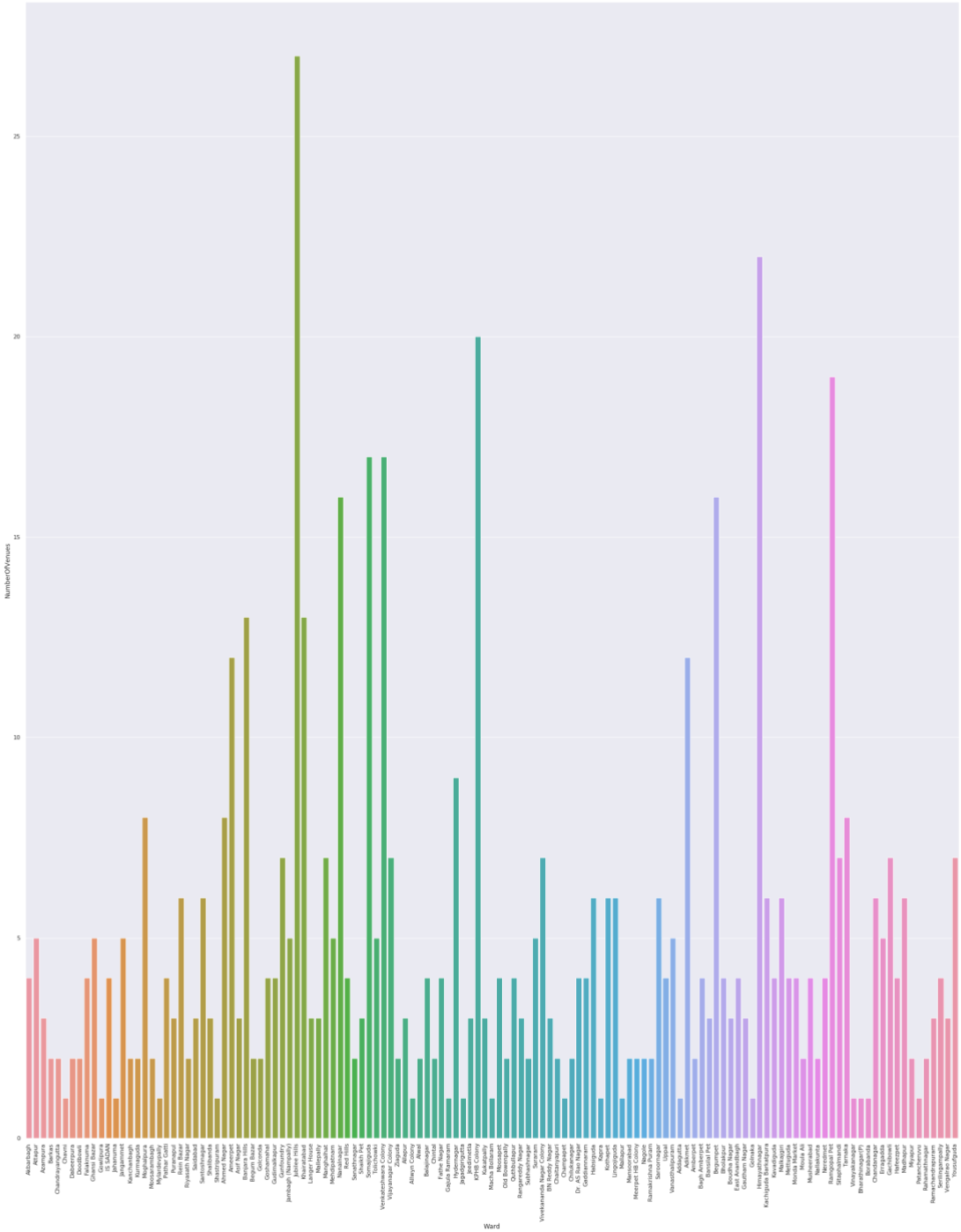


2. Identify and Explore the Venues using Foursquare API:

We use Foursquare API to retrieve all the venues within a 500 metre radius of each Ward. In order to achieve this, we need to first register for a Foursquare Developer Account. Upon registration we obtain the Foursquare ID and the Foursquare secret key. We use these ids along with the geographical coordinates of each ward in a python loop to make API calls to Foursquare.Foursquare returns this venue data in a JSON format. We extract the venue name, category and the venue coordinates (latitude & longitude). With this data we can check how many venues are returned for each ward and how many unique categories can be

compiled from the returned venues. We then analyse the ward by grouping each row by ward and the mean frequency of occurrence of each venue category. This way we prepare the data for use in clustering.

In the graph below, we can see that the wards with the highest venue count are: Jambagh (Nampally) and Himayathnagar.

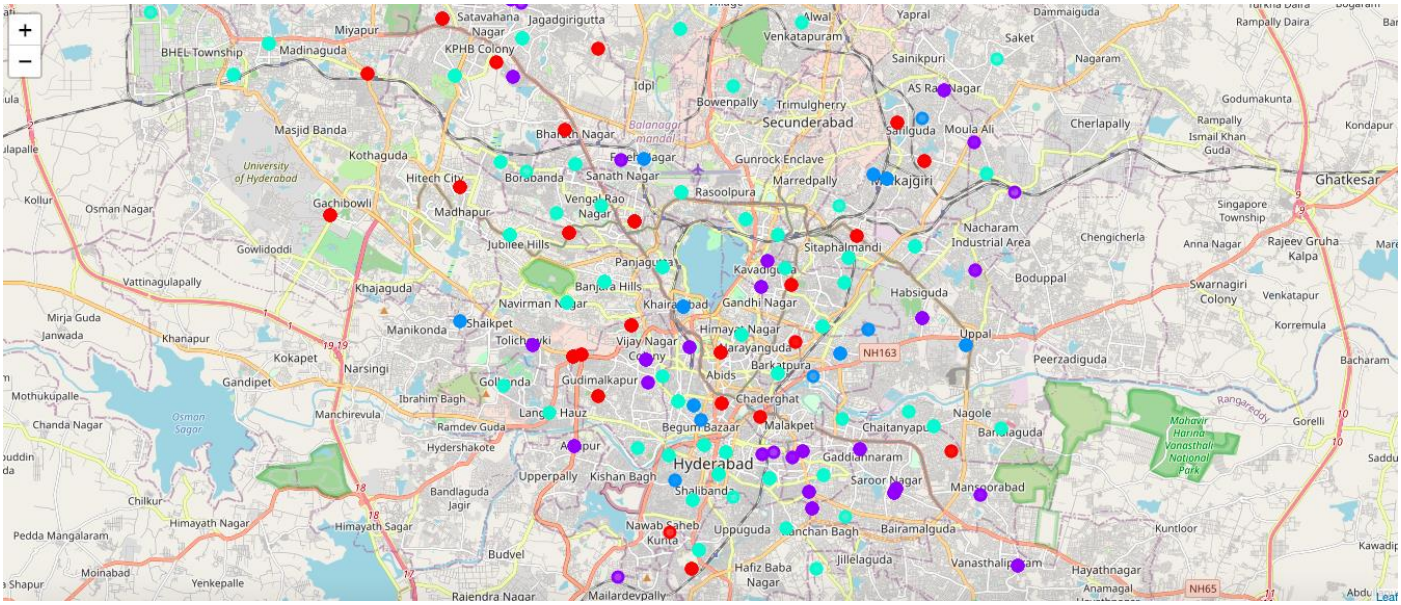


3. Perform Clustering using K-Means Clustering:

We perform clustering on the data using K-Means clustering. The K-Means clustering algorithm identifies k number of centroids and then allocates every data point to the nearest cluster, keeping the number of centroids as small as possible. It is a simple and popular unsupervised machine learning algorithm and is best suited to obtain the desired results of this project. We identify the optimum k number using the elbow method.

In this program based on the wards, venue data retrieved and their categories, the optimum k value we begin with is 4. The wards are then grouped into 4 clusters. We display these clusters on the Hyderabad City Map.

Colour of Circles	Cluster Number
Red	0
Purple	1
Blue	2
Light Green	3



4. Creating the Recommender System:

The Locality Recommendation system is created based on the Population of the Ward, the Average Monthly Rent of a 2BHK apartment in the Ward and a special parameter, which in this case is the venue category: Vegetarian Restaurant. We assess the number of Vegetarian Restaurants in Each Ward.
We also retrieve the top 10 venues in each ward in order to be able to understand the locality better and to build a better recommendation system.

Ameerpet is taken as a target or reference venue based on user input, for this program, in order to eventually recommend localities within the target or reference cluster.

RESULT:

[84]:

	Ward	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Ahmed Nagar	Venue Category_Indian Restaurant	Venue Category_Pizza Place	Venue Category_Hotel	[0.23475159972514792]
1	Ameerpet	Venue Category_Indian Restaurant	Venue Category_Fast Food Restaurant	Venue Category_Women's Store	[0.5303522750453613]
2	Azampura	Venue Category_Hotel	Venue Category_Tourist Information Center	Venue Category_Indian Restaurant	[0.41677435957312037]

The Result List produced gives us the names of ideal localities with similarity to Ameerpet (in house rental rates, population and in the number of vegetarian restaurants) and the top 3 venues in those localities. The ranking is also given in order to assist the user in making a better informed choice.

DISCUSSION:

The most important observation which could alter the results is:

- An alternate cluster number (k number) could significantly alter the results. Based on a high or low k number the results could be either overfitted or underfitted. It, therefore, is very important to conduct an analysis of the number of clusters prior to clustering. In this program we have used the elbow method to arrive at k (the number of clusters).

To be able to factor in vegetarian restaurant options along with giving greater priority to the rental rates and the population density, we need to create a Non-vegetarian restaurant category which is equal to 1- the total number of Vegetarian restaurants for each ward. In the final algorithmic equation, we give this parameter the least weightage so as to give a higher ranking to wards with Vegetarian Restaurants.

CONCLUSION:

The Hyderabad Locality Recommendation system relies on the important factors of Avg.Rent, Population ,the special parameter (which in this case is the Vegetarian Restaurant requirement) and the Foursquare API to analyse the venues in each Ward/Locality.