Capstone Project - The Battle of Neighborhoods

Presented By – Ajay Tripathi

Jan 2020

Introduction

- **Background**
  - Moving to Hyderabad, India due to work and know nothing about the city; Have a family and looking for place to reside that is close to school, hospital, convenience store and office
- **Problem**
  - This project aims to find specific places of interest in and around various neighborhoods of Hyderabad to enable a family or individuals moving to Hyderabad either from within India or from any part of the world to view neighborhoods and close by venues of their interest and take a decision where they would like to either buy or rent a house
- **Interest**
  - People who are considering relocating to Hyderabad will be interested to identify & explore its neighborhoods and specific venues around each neighborhood in order to decide on a location based.

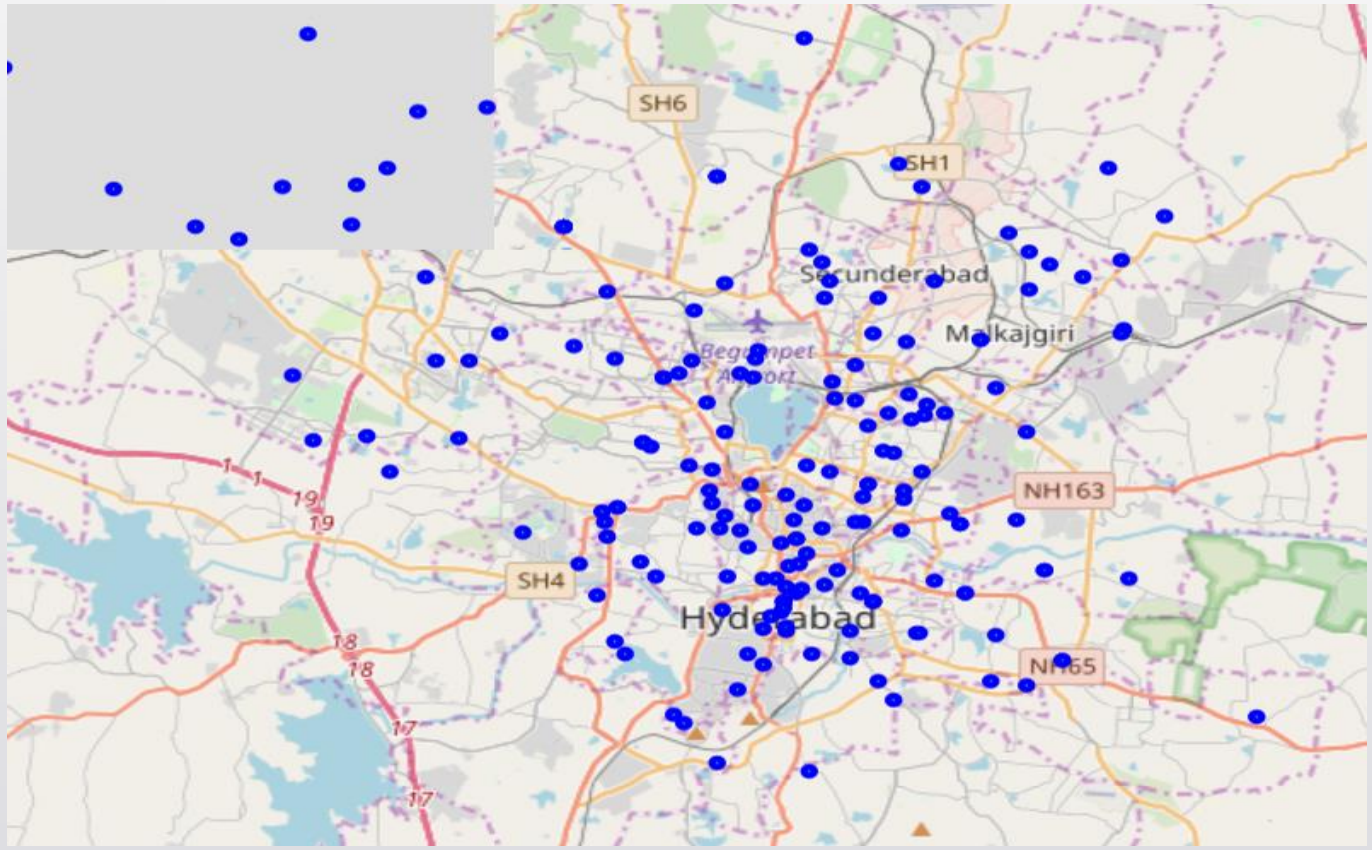
Data Acquisition & Cleaning

- **Data Acquisition -**
  - The data required for this project is a list of Hyderabad neighborhoods with their latitude and longitude information. The data will be web scrapped from Wikipedia links.
  - Hyderabad neighborhoods - https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India
  - Each neighborhood listed on this page points in the form of a link to neighborhood information page having more information about the neighborhood including its latitude and longitude information
- **Data Cleaning**
  - The data cleaning would be done in two steps using Beautiful Soup library in python to web scrap data
  - The first step would be to scrap the data from Hyderabad neighborhoods link and stored in a tabular format. After the web scrapping, string manipulation would be performed to get the neighborhoods names and their corresponding links information.
  - The second step would involve going through each neighborhood one by one and web scrapping the linked page(@link<X>) to extract latitude and longitude information. For each neighborhood information page, we'll extract the page content and perform string manipulation to extract Latitude and Longitude Information for the final dataset
  - Since the latitude and longitude information is given in degree, minutes and seconds on web pages, thet would be converted to decimals as required by the foursquare api while fetching from Wikipedia.

Methodology

Exploratory Data Analysis

We have quite several neighborhoods identified in the city and that probably will help in making a close observation of the places in and around. There 199 neighborhoods fetched from Wikipedia as shown below.



Modelling

- Using the final dataset containing the neighborhoods in Hyderabad along with the latitude and longitude, we can find specific venues that we selected within a 2KM radius of each neighborhood by connecting to the Foursquare API

	Neighborhood	Neighborhood_Latitude	Neighborhood_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	A.C. Guards	17.396944	78.456944	DineHill	17.405256	78.451674	Indian Restaurant
1	A.C. Guards	17.396944	78.456944	Paradise Restaurant	17.403602	78.452848	Indian Restaurant
2	A.C. Guards	17.396944	78.456944	Prince Hotel	17.394736	78.442410	Indian Restaurant
3	A.C. Guards	17.396944	78.456944	Sandarshini	17.405330	78.451656	Indian Restaurant
4	A.C. Guards	17.396944	78.456944	Al-humdulillah Hotel	17.390728	78.462992	Indian Restaurant
5	A.C. Guards	17.396944	78.456944	Sarvi	17.412698	78.449704	Indian Restaurant
6	A.C. Guards	17.396944	78.456944	Hyderabad House	17.402775	78.456992	Indian Restaurant
7	A.C. Guards	17.396944	78.456944	Kamat Hotel	17.404330	78.467602	Indian Restaurant
8	A.C. Guards	17.396944	78.456944	Paradise Food Court	17.412113	78.465443	Indian Restaurant

- Many venues returned by the API with restaurants being the highest in numbers
- The focus was moved to specific categories excluding restaurants ensuring API covered more of hospitals, schools, pharmacy and convenience stores.

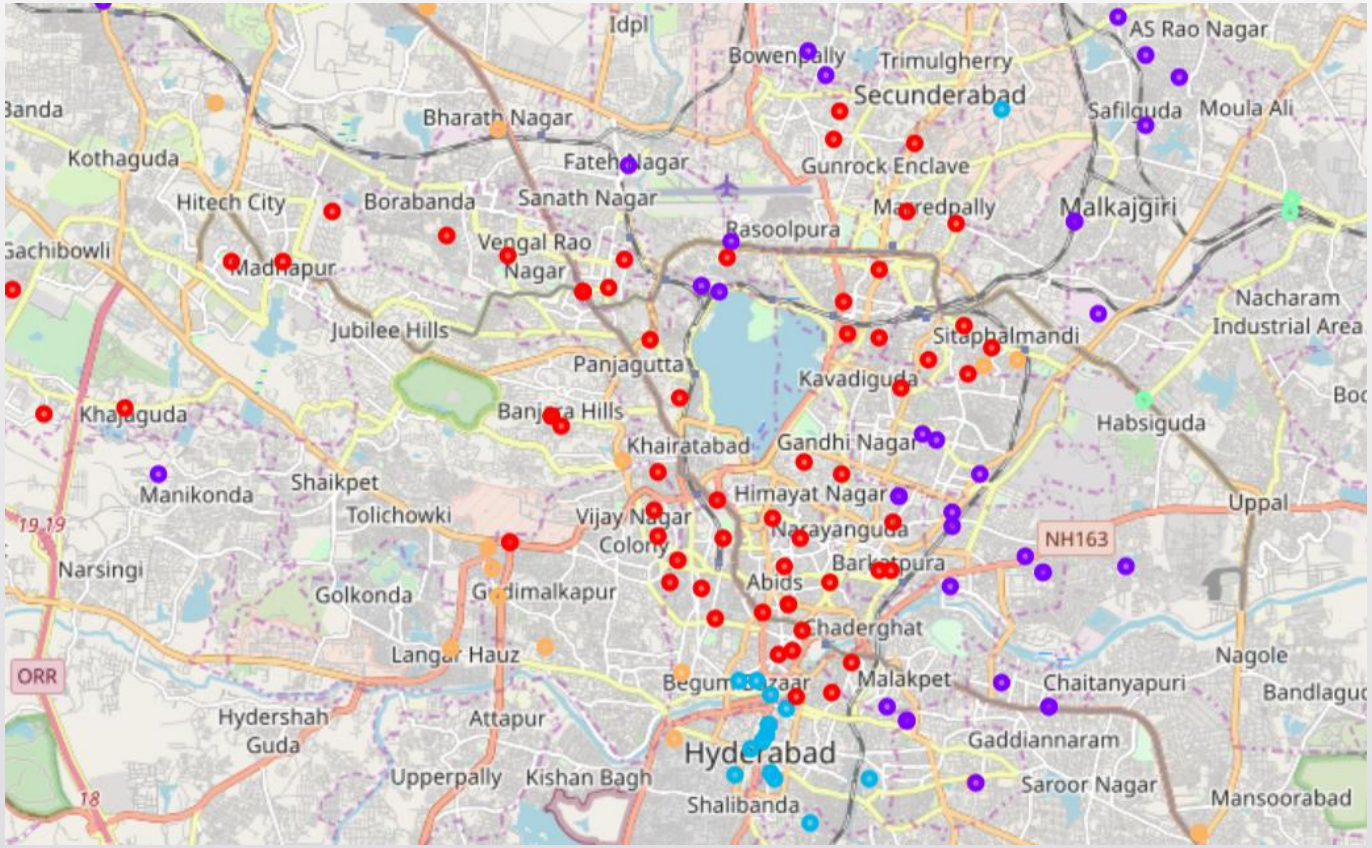
	Neighborhood	Latitude	Longitude	Cluster Labels	1st_Most_Common_Venue	2nd_Most_Common_Venue	3rd_Most_Common_Venue	4th_Most_Common_Venue
0	A.C. Guards	17.396944	78.456944	0	Hospital	Convenience Store	Medical Center	Grocery Store
1	Abhyudaya Nagar	18.991378	72.844164	0	Hospital	School	Pharmacy	Medical School
2	Abids	17.366111	78.476111	2	Hospital	School	Trade School	College Administrative Building
3	Adikmet	17.406944	78.512778	1	Convenience Store	Hospital	Pharmacy	Trade School
4	Afzal Gunj	17.373333	78.470833	2	Hospital	Chaat Place	School	College Administrative Building
5	Aghapura	17.388333	78.464444	0	Hospital	Hotel	Chaat Place	School
6	Alijah Kotla	17.357222	78.476944	2	Hospital	School	Trade School	College Administrative Building
7	Allwyn Colony	17.483333	78.416667	4	Hospital	Trade School	College Administrative Building	College Classroom
8	Amberpet	17.390833	78.523611	1	Convenience Store	Hospital	School	Trade School
9	Ameerpet	17.437500	78.448200	0	Hospital	Pharmacy	Convenience Store	Trade School

Modelling

- One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods
- K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster neighborhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood

Results

After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters. Each cluster is color coded for ease of presentation



Cluster 0

	Neighborhood	Latitude	Longitude	Cluster Labels	1st_Most_Common_Venue	2nd_Most_Common_Venue	3rd_Most_Common_Venue	4th
0	A.C. Guards	17.396944	78.456944	0	Hospital	Convenience Store	Medical Center	
1	Abhyudaya Nagar	18.991378	72.844164	0	Hospital	School	Pharmacy	
5	Aghapura	17.388333	78.464444	0	Hospital	Hotel	Chaat Place	
9	Ameerpet	17.437500	78.448200	0	Hospital	Pharmacy	Convenience Store	
10	Ashok Nagar, Hyderabad	17.406944	78.488611	0	Hospital	High School	Convenience Store	
11	Attapur	17.395833	78.431111	0	Hospital	Medical Center	Grocery Store	
13	Azampura	17.376111	78.490278	0	Hospital	Chaat Place	School	

- Visualizing the cluster0, we see that it is the largest with 144 neighborhoods and dense in the center also. This cluster covers most of the important venues that would probably be of anyone's interest and covers hospitals, schools, pharmacy and convenience stores

Cluster 1

- The second cluster 1 has good number of mix of venues and ideal for families to consider

3rd_Most_Common_Venue	3rd_Most_Common_Venue
Convenience Store	5
Emergency Room	1
High School	3
Hospital	7
Pharmacy	8
School	2
Supermarket	2
Trade School	11

2nd_Most_Common_Venue	2nd_Most_Common_Venue
Convenience Store	17
Hospital	7
Pharmacy	2
School	6
Trade School	7

1st_Most_Common_Venue	1st_Most_Common_Venue
College Administrative Building	1
College Classroom	1
Convenience Store	11
Department Store	2
Hospital	18
School	4
Student Center	1
Supermarket	1

Cluster 2

	Neighborhood	Latitude	Longitude	Cluster Labels	1st_Most_Common_Venue	2nd_Most_Common_Venue	3rd_Most_Common_Venue	4th_Most_Common_Venue
2	Abids	17.366111	78.476111	2	Hospital	School	Trade School	College Administrative Building
4	Afzal Gunj	17.373333	78.470833	2	Hospital	Chaat Place	School	College Administrative Building
6	Alijah Kotla	17.357222	78.476944	2	Hospital	School	Trade School	College Administrative Building
22	Begum Bazaar	17.373333	78.473889	2	Hospital	Chaat Place	School	College Administrative Building
33	Chatta Bazaar	17.368800	78.479200	2	Hospital	Chaat Place	School	College Administrative Building
37	Dabirpura	17.366111	78.476111	2	Hospital	School	Trade School	College Administrative Building

- Cluster 2 covers has most common venues as hospitals, school and college building and would be of interest to someone wanting to live close by to hospital.

Cluster 3

	Neighborhood	Latitude	Longitude	Cluster Labels	1st_Most_Common_Venue	2nd_Most_Common_Venue	3rd_Most_Common_Venue	4th_Most_Common_Venue
42	ECIL 'X' Roads	17.451749	78.567116	3	Gas Station	Trade School	Hospital	College Administrative Building
52	Habsiguda	17.419167	78.541389	3	Gas Station	High School	Trade School	Hospital
88	Mallapur	17.450000	78.566667	3	Gas Station	Trade School	Hospital	College Administrative Building

- Small cluster contains mostly gas stations, trade schools & college buildings

Cluster 4

	Neighborhood	Latitude	Longitude	Cluster Labels	1st_Most_Common_Venue	2nd_Most_Common_Venue	3rd_Most_Common_Venue	4th_Most_Common_Venue
7	Allwyn Colony	17.483333	78.416667	4	Hospital	Trade School	College Administrative Building	College Classroom
24	Bharat Nagar	17.463333	78.428889	4	Hospital	Trade School	College Administrative Building	College Classroom
31	Champapet	17.348333	78.550833	4	Hospital	Trade School	College Administrative Building	College Classroom
39	Dhoolpet	17.374444	78.460833	4	Hospital	Trade School	College Administrative Building	College Classroom

- Cluster 4 mostly contains hospitals, trade school and college classrooms and building which may not of high interest to individuals or families except students who may move to the city for education purpose

Discussion

- The aim of this project is to help people who want to relocate to the Hyderabad in India, they can choose the neighborhoods to which they want to relocate based on the common venues any family would want to have close by.
- For a family I feel that the neighborhoods in Cluster 0 are more suitable due to the venues in that cluster, these neighborhoods have common venues as categorized which is ideal for a family. The choices of neighborhoods may vary from person to person.

Conclusion

This project helps a person get a better understanding of the neighborhoods with respect to the required venues in that neighborhood. It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a neighborhood. We have just taken family into consideration for people moving from outside Hyderabad to have a better sense of the neighborhood they would want to move considering venues of their interest proximity. This data set can also be looked from the perspective of neighborhoods crime statistics and Airbnb price predictor based on location….something to consider in future analysis