IBM Data Science Professional Certificate

Capstone Project: The Battle of the Neighbourhoods

Segmenting and Clustering selected Neighbourhoods of Hyderabad

Johny Ijaq

July 18, 2020

1. Introduction

1.1. Background

Hyderabad city is the capital city of the Indian state of Telangana. The city has an estimated population of around 8 million, making it 4th largest city in India, while the population of the metropolitan area was estimated above 9 million. Hyderabad city is known for its rich history, food and its multi-lingual culture, both geographically and culturally. As a resident of this city, I decided to explore Hyderabad for my project.

Hyderabad, since the time of the Nizams, has been the seat of art, culture, literature, music, science, and a variety of other academic activities. In modern times as well, the city has managed to hold on to its rich academic tradition, with some of the most premier educational institutions in the country making their home on the city's landscape

The city has three central universities, including the top-ranked University of Hyderabad, The Maulana Abdul Kalam Azad National Urdu University and English and Foreign Languages University, along with the famous state university such as Osmania University. The city and surrounding areas are home to top engineering institutes such as Indian Institute of Technology - Hyderabad, International Institute of Information Technology, Hyderabad, Birla Institute of Technology & Sciences. Hyderabad also hosts big names in business schools such as Indian School of Business (ISB) and ICFAI Business School. Nalsar University of Law, Tata Institute of Social Science Research and Tata Institute of Fundamental Research are the other prestigious institutes in the city.

Hyderabad is also a major centre for research in pharma, biotech and defence-related science and technology. Apart from many defence research labs such as Defence Research and Development Organisation (DRDO) and space institutes, the city has prestigious research laboratories such as the Centre for Cellular and Molecular Biology (CCMB), Indian Institute of Chemical Technology, Centre for DNA and Fingerprinting Diagnostics (CDFD), National Institute of Nutrition (NIN) etc.

Incorporating a large number of universities, management colleges, research centres, and technical institutes, the capital of Telangana, is undoubtedly shaping itself rapidly to be the country’s higher education hub. This has led to the development of many business centres nearby these educational and research institutes. Exploring the venues around these centres could provide valuable insights about the business opportunities present in these areas.

1.2 Problem

I selected few neighbourhoods from Hyderabad that are hosting famous educational and research institutes. From this project I would like to explore all the great amenities and other types of venues that exist in the neighbourhood, such as restaurants, cafeterias, fast food joints, bakeries, stationaries, pharmacies, parks, schools and so on.

Main objective of this project is to find out the most common businesses in these areas, to explore the best business idea which one can start in these neighbourhoods and to select the area where the type of business one want to install is less intense. Finally, a combination of location data and machine learning was used to group the neighbourhoods into clusters.

1.3. Interest

The results from this project can be useful to anyone who is planning to start a business in the studied neighbourhoods. Anyone who is looking to find the best neighbourhood to open a shop of interest also finds this study useful. One can choose to compare different neighbourhoods in terms of a service, search for potential explanation of why a neighbourhood is popular, the cause of complaints in another neighbourhood, or anything else related to neighbourhoods.

2. Data sources

1. Latitude and longitude data for each neighbourhood was obtained from https://www.gps-latitude-longitude.com/address-to-longitude-latitude-gps-coordinates.

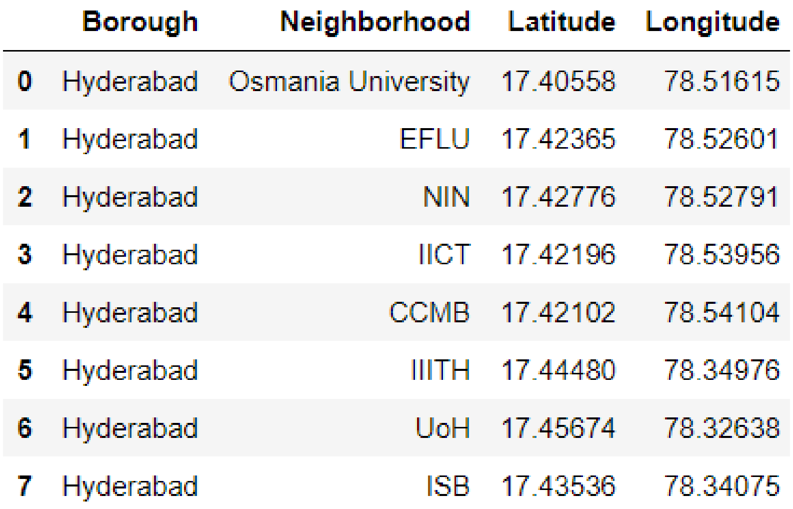
2. I used the Foursquare API to explore the nearby venues from the selected universities and research institutes in order to find the common venues or businesses. I searched for the venues up to 1 KM from the GPS coordinates (latitude and longitude) of the six neighbourhoods. Foursquare API helped me to search for a specific type of venues, explore a particular venue and to get trending venues around each neighbourhood. I retrieved the following information from the foursquare API for each neighbourhood.

   A. Venue Name
   B. Venue Category
   C. Venue Latitude
   D. Venue longitude

3. Methodology

3.1. Data Acquisition

Latitude and Longitude information of the selected neighbourhoods was obtained from Google. Data was tabulated in excel with *Borough*, *Neighbourhood*, *Latitude and Longitude* as components. File was saved in CSV format and was read using *pandas* library and used for further analysis.



	Borough	Neighborhood	Latitude	Longitude
0	Hyderabad	Osmania University	17.40558	78.51615
1	Hyderabad	EFLU	17.42365	78.52601
2	Hyderabad	NIN	17.42776	78.52791
3	Hyderabad	IICT	17.42196	78.53956
4	Hyderabad	CCMB	17.42102	78.54104
5	Hyderabad	IIITH	17.44480	78.34976
6	Hyderabad	UoH	17.45674	78.32638
7	Hyderabad	ISB	17.43536	78.34075

3.2. Exploring the selected Neighbourhoods in Hyderabad

3.2.1. Obtain the latitude and longitude values of Hyderabad city

First, GeoPy library was used to get the latitude and longitude values of Hyderabad City. GeoPy library is a Python client for several popular geocoding web services. It provides the coordinates (latitude and longitude values) for given addresses, cities, countries, and landmarks across the world using third-party geocoders and other data sources (1). In order to define an instance of the geocoder, we need to define a user_agent. Here the user agent was renamed as 'hyd_explorer'.

3.2.2. Create a map of Hyderabad showing the selected neighbourhoods

Python geospatial visualization library *folium* was used to visualize the selected neighbourhoods. Latitude and longitude values obtained from GeoPy was used to create a map of Hyderabad with neighbourhoods superimposed on top.

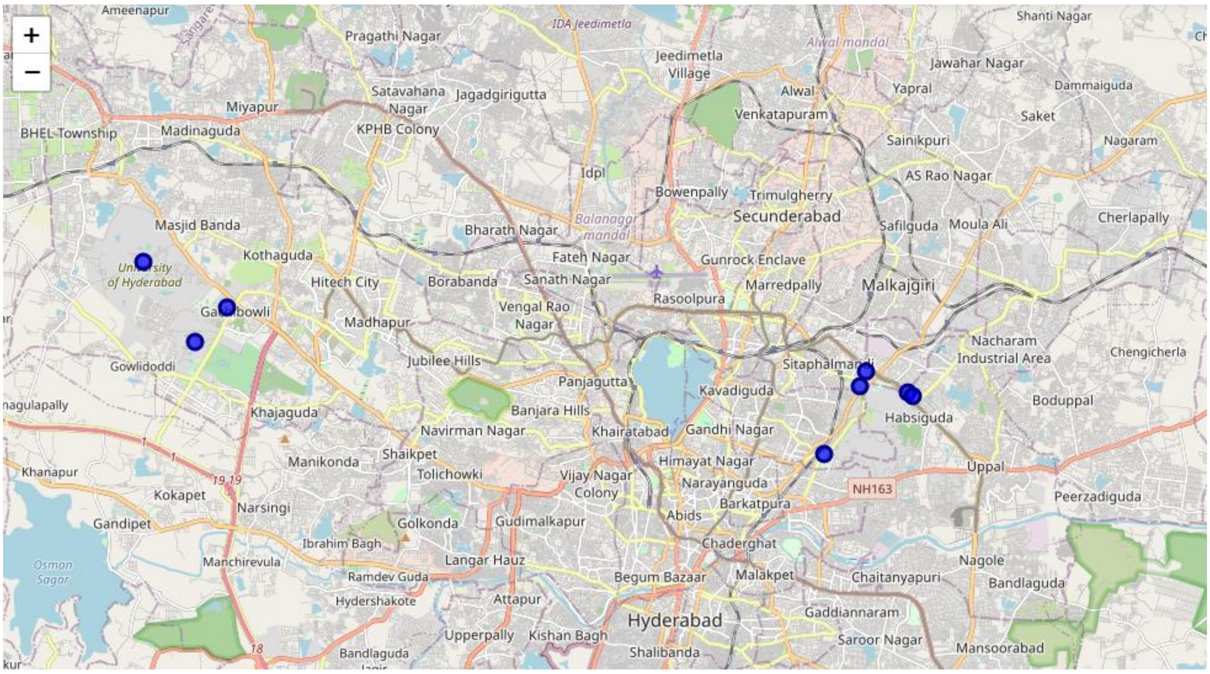


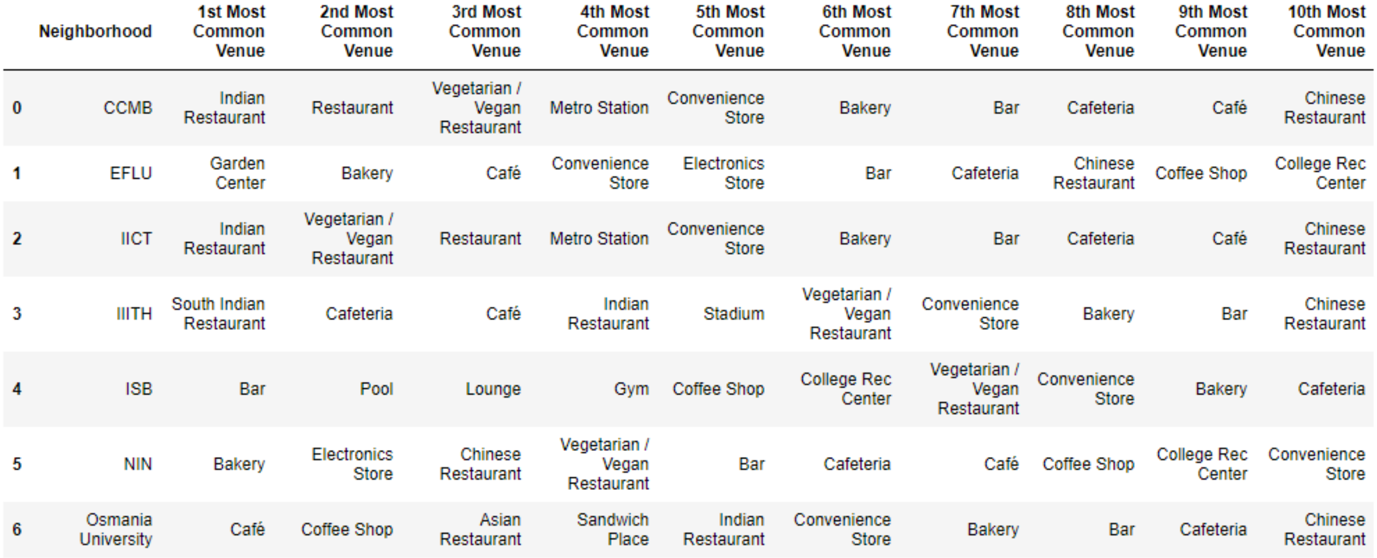
Figure 1. Map of Hyderabad with neighbourhoods superimposed on top.

3.2.3. Defining Foursquare Credentials and Version

Foursquare is a technology company that built a massive dataset of location data. For this project, Foursquare API was used to search for specific type of venues or stores and to learn more about particular venues around a given location. To make calls to the Foursquare API, uniform resource identifier (URI) has to be created (api.foursquare.com/v2). It should be appended with extra parameters depending on the data (viz. venues, users, tips etc.) that we seek from the database. Call request to the database was made by passing my developer account credentials, which include Client ID and Client Secret as well as the version of the API (date).

3.2.4. Obtaining venue information

Once the call was made to the Foursquare database, JSON file was returned containing all information the database has about the required venue. Information about the name of each venue, its unique ID, location, category, average rating, location and finally, tips posted about the venue. JSON structure is then cleaned into a pandas dataframe showing the name of the venue, its category, latitude and longitude. Similar analysis was performed for each neighbourhood and the list of venues nearby each neighbourhood were obtained. Each neighbourhood was analyzed by applying one-hot encoding to ‘venue category’ to know the occurrence of each category. Finally, top 10 venues for each neighbourhood were displayed.



	Neighborhood	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	CCMB	Indian Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Metro Station	Convenience Store	Bakery	Bar	Cafeteria	Café	Chinese Restaurant
1	EFLU	Garden Center	Bakery	Café	Convenience Store	Electronics Store	Bar	Cafeteria	Chinese Restaurant	Coffee Shop	College Rec Center
2	IICT	Indian Restaurant	Vegetarian / Vegan Restaurant	Restaurant	Metro Station	Convenience Store	Bakery	Bar	Cafeteria	Café	Chinese Restaurant
3	IIITH	South Indian Restaurant	Cafeteria	Café	Indian Restaurant	Stadium	Vegetarian / Vegan Restaurant	Convenience Store	Bakery	Bar	Chinese Restaurant
4	ISB	Bar	Pool	Lounge	Gym	Coffee Shop	College Rec Center	Vegetarian / Vegan Restaurant	Convenience Store	Bakery	Cafeteria
5	NIN	Bakery	Electronics Store	Chinese Restaurant	Vegetarian / Vegan Restaurant	Bar	Cafeteria	Café	Coffee Shop	College Rec Center	Convenience Store
6	Osmania University	Café	Coffee Shop	Asian Restaurant	Sandwich Place	Indian Restaurant	Convenience Store	Bakery	Bar	Cafeteria	Chinese Restaurant

Figure 2. List of top 10 venues for each neighbourhood.

3.2.5. Knowing the trending venues around each neighbourhood

To get trending venues or popular spots around a point of interest ‘explore’ endpoint was used instead of the ‘search’ endpoint, and the latitude and the longitude coordinates of the required venue were passed along with developer credentials. Then a call was made to the database, to get a list of trending spots around each neighbourhood.

3.4. Clustering Neighbourhoods

Next, to find similar neighbourhoods, they were grouped into clusters. k-means clustering algorithm was used to cluster the neighbourhoods based on their similarity. k-means is an unsupervised algorithm that group data based on the similarity of data points to each other. It is a type of partitioning clustering where it divides the data into k non-overlapping subsets or clusters. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar.

Clustering the neighbourhoods into distinct groups or clusters was done based on the number of clusters selected and the distance of each data point from the centre of cluster (centroid of cluster). To cluster the selected neighbourhoods the value of k was set to 4. This resulted in 4 clusters based on their similarity between the neighbourhoods.

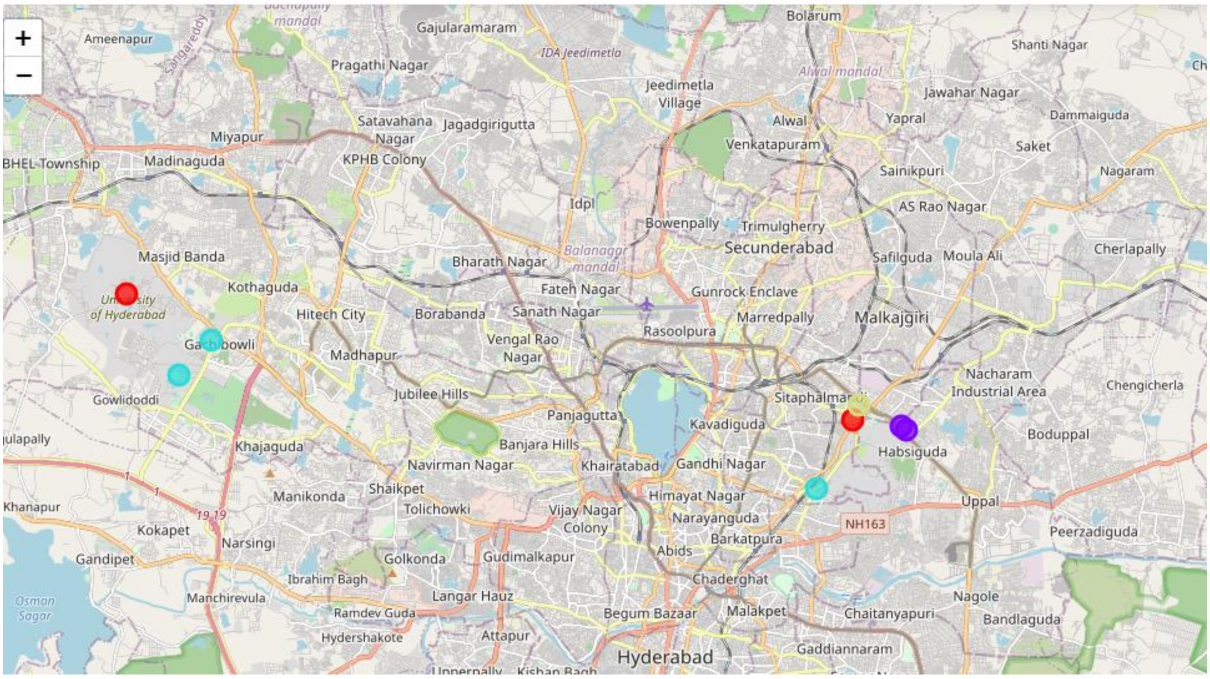


Figure 3. Figure showing four different clusters.

4. Results and Discussion

Restaurants, cafeterias and food courts dominated the venue list nearby all the 8 educational institutions. This shows the demand for food business in these areas. CCMB and IICT are located near to each other and shared most of the venues, as expected. Interestingly, Osmania University and IIITH which are 24 KMs apart from each other shared similar venues. No venues were shown nearby to University of Hyderabad.

The Hyderabad city has been attracting a large number of students not only from all over the country but also from abroad. This has created suitable environment to start a new business nearby many educational and research institutes. In this context, it would be helpful if we could gather the information regarding different business operating around these areas so that it would be easy to identify new opportunities. This analysis can be used to determine the business scenario near famous educational institutes and research centres in the city. Similar analysis can also be used to determine the neighbourhoods from different parts of the city that are similar in terms of ongoing businesses.

5. Conclusion

In this project, I have gone through the process of identifying the business problem, specifying the data required, preparing the data, performing the machine learning by using k-means clustering method and presenting the outcome pertaining to the problem.

References:

1. Hyderabad: A vibrant hub for higher education. (https://www.thehindubusinessline.com/news/education/hyderabad-a-vibrant-hub-for-higher-education/article30360645.ece)
2. 2. GeoPy’s documentation: https://geopy.readthedocs.io/en/stable/#