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 locality, transportation, water facilities, environment, parks, hospitals, schools 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 factors for investment in the city 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 co-ordinates(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 wrangled/scraped 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 is *Banjara Hills* is a highly developed area with prices of the place ranging from Rs. 5,312 - 7,098 per square foot. Secondly, pricing data extracted from previous step is used to visualize which areas of the city are more costlier 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 venues(restaurants, cafes, parks, studios etc), 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.