Clustering Apartment complexes in Bangalore

Aravind Ganesan

June 14th 2020

1. Introduction

1.1 Background

Bangalore is a city in the southern part of India and is commonly known as Silicon Valley of India because of being a technology and start-up capital of India. The city has also seen a real estate boom on the onset of millennium due to an increase in the affordability of the middle class. There is a value in understanding the real estate packages on offer based on the size and location. This will allow us to understand what are the various price segments in which apartments are offered and is it actually providing enough value for the price.

1.2 Problem Statement

This analysis will be focusing on a hypothetical prospective customer named Mr. Omar Little who is moving to Bangalore with family for a long time and is looking for an apartment.

He has the following requirements:

a) He is looking for a moderately sized 3 Bedroom apartment which is moderately (not premium) prized

b) He is looking for the apartments which have schools and hospitals in the vicinity (within 2 Kms of the apartment)

1.3 Who is this for?

This analysis is for potential homebuyers in any big city across India who are home-hunting and are looking for apartments which meet their affordability criteria. This helps us in understanding how apartment offerings are packaged according to size and price in the metro cities in India such as Bangalore.

2. Description of the Data

2.1 Source of Data

The data was sourced from Kaggle (here) where a user has collected the real estate website data from a website. The data had was available as 2 CSV files:

a) ***apartment* dataset*:*** *datasets_151373_358770_apartment_data.csv* – List of apartments in Bangalore with their longitude and latitude

b) ***apartment details* dataset*:*** *datasets_151373_358770_blore_apartment_data.csv* – The size (number of rooms), area, price of the apartments ()

2.2Data Cleaning

2.2.1 Data Cleansing for apartments.csv:

a) The apartment dataset has just 4 columns with the *apartment name*, *latitude* and *longitude* and *geometry*. The geometry column was dropped as it was a repeat of the coordinates again.

b) The names column had city appended to each name of the column. The city name was parsed out of the *names* column.

2.2.2 Data Cleansing for apartment details.csv

a) The apartment details csv had four columns namely *names, Price, Area, Unit type.*

b) The *Price* column were coded as alphanumeric (e.g. 100000 as 1 L, 1000000 as 1C) and the lower and upper prices were coded in the same column. The column was split based on '-' and the amounts were converted to a numeric value.

c) The *Unit Type* had BHK appended to it. To convert it to a numeric value, the BHK was split off.

d) Some of the apartments had types such as 1 RK Studio Apartment, 1 RK Apartment, 1RK, Studio. These were coded as 0.5 BHK apartments.

e) There were some properties which had an entry simply as a Plot. These entries were dropped.

f) The Price and Area fields were converted to Float data types.

g) The Mean price and Mean area per property was determined to standardize and cluster it later.

2.2.3 Merging

a) The data frames created above were merged into a single final data set which was used for further processing.

2.2.4 Summary of pre-processing

Input file name	Input Columns	Columns Dropped	Columns Created	Output Data frame Name
datasets_151373_358770_ apartment_data.csv	*names,lat,lon,geometry*	*geometry*	NA	bang_apt
datasets_151373_358770_blore_ apartment_data.csv	*names,price,area,Unit type*	*price,area,Unit type*	Size(BHK),Area-L,Area-H,Price-L-num,Price-H-num	bang_apt_details