**Segmentation of Hyderabad Localities**

Tina Shrivastava

March 12, 2020

1. **Introduction**
   1. **Background**

Hyderabad is the capital city of Telangana. Occupying 650 square kilometers (250 sq mi) along the banks of the Musi River, located on the Deccan Plateau in the northern part of South India. It consists of over 1000 neighborhoods with over 9.7 million population making in Fourth most populous city in India. Certain localities are highly developed areas while remaining fall under un-developed areas. So, when a person wants to buy a new house he/she is left with a question “Which areas offer good facilities for it’s residents?”

* 1. **Problem**

The main objective of this project is to determine which area stands in average pricing and huge no.of facilities to it’s residents. This project helps a individual to find out in which area he/she has to buy a new house.

* 1. **Interests**

An individual who wants to change to new house (or) a person who wants to rent or buy a new house.

1. **Data acquisition and cleaning**
   1. **Data sources**

I have collected data regarding to neighborhoods and plot average prices from manamm.com website using scrapping. I used geocoder and FourSquare API to get the location coordinates of individual areas and used these coordinates to find all nearby schools, colleges, hospitals, restaurants etc.

* 1. **Data cleaning**

The data which is scrapped consists of html tags which is unnecessary to the project. First, I have deleted all the unnecessary tags and I was left which 26 tables which consists of neighborhood details.

There are lot of missing values in the tables. So, I decided to remove the inconsistent rows and only keep the consistent rows and columns. The data in the <td> tag consists of special characters which cannot be processed by algorithm, so I split the value to replace special characters using python split and replace methods. After performing these steps, I was left with plain float value of price details.

Using geocoder, I was able to get the location coordinates of small set of localities using their names, but a large no of areas are left out. So I decided to use club nearby areas. After this, I was able to get latitude and longitude and Hyderabad neighborhoods.

**2.3 Feature Selection**

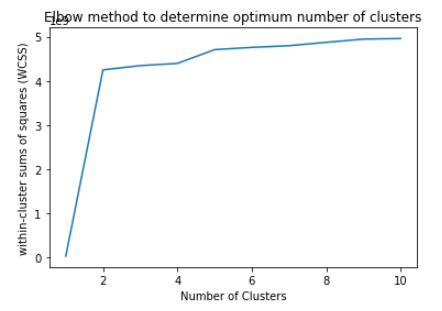
After data cleaning, there were 1200 rows with 10 features in the data. Upon examining , these features I was able to conclude that min and max plot prices can be clubbed as average pricing. Individual items such as schools, colleges, restaurants can be named under single data entry of Places\_To\_Visit. The areas which have null values are dropped out or clubbed into nearby bigger areas. At last, I was left with 620 rows with 6 features.

1. **Methodology**
   1. **Model Selection**

There are vast under of models to determine the solution to my problem. But, as I’m working with unknown dataset with no specific labels; I decided to go with unsupervised clustering methodology. In which I picked K means mechanism as it is a simple one and best suited for mid-range feature clustering.

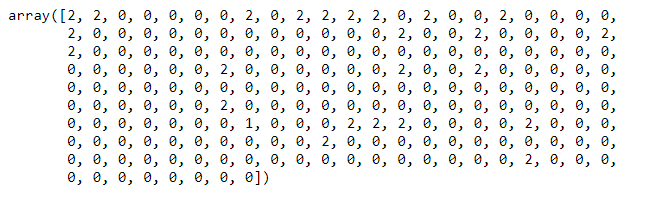
* 1. **Determining best K in K means**

To determine the best K, I used Elbow method to determine optimum number of clusters in the dataset. I have predicted the sample data with a value of clusters from 1 to 10. By observing the results, I was able to conclude that keeping k value to 2 is optimum as it doesn’t give a notable change in the metrics.



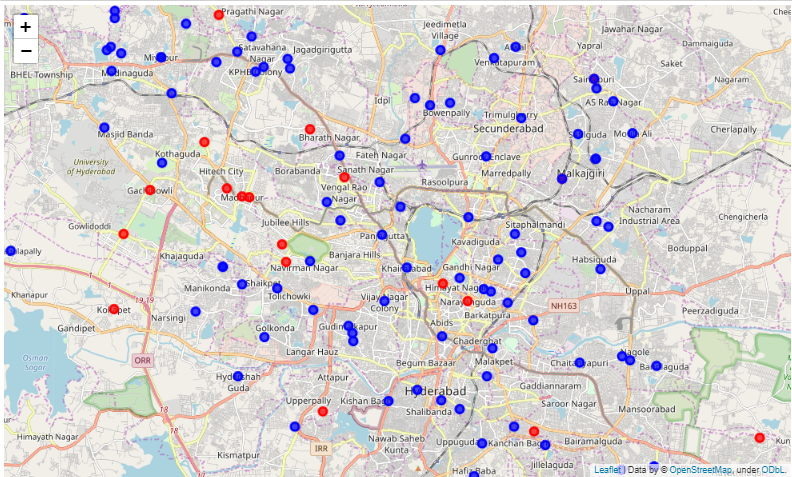
* 1. **Clustering**

Keeping no of clusters to 2, I have fitted the k means algorithm with train data chosen randomly. Later using the original data I have clusters all neighborhoods in the dataset as best areas to buy & not good areas to buy a new house.



1. **Results**

Using Folium I have created markers for every neighborhood in Hyderabad using there location coordinates. I have labelled all best areas with red color marker and not good areas with blue color marker and added these markers to folium map. The result looks as follows :



1. **Future Reference**

I was unable to add survey data about land i.e., drought/ flood areas during rainy seasons, highway connectivity areas, soil fertility/minerals, water minerals etc. These datasets would be very helpful in getting a better understanding of areas.

1. **Conclusions**

In this project I was able to determine certain areas such as Gachobowli, Upperpally, Jubilee Hills etc have a better life style and offer their residents a wide range of facilities and easy access to their offices, schools, colleges.