CLUSTERING PUNE

IBM Data Science Professional Specialization - Coursera

-HARSH GUPTA

INTRODUCTION

Pune is the second largest city in the Indian state of Maharashtra, after Mumbai. It is the ninth most populous city in the country with an estimated population of 3.13 million. Along with its extended city limits Pimpri Chinchwad and the three cantonment towns of Pune, Khadki and Dehu Road, Pune forms the urban core of the eponymous Pune Metropolitan Region (PMR).According to the 2011 census, the urban area has a combined population of 5.05 million while the population of the metropolitan region is estimated at 7.27 million. The city is considered to be the cultural capital of Maharashtra. It is also known as the "Oxford of the East" due to the presence of several well-known educational institutions. The city has emerged as a major educational hub in recent decades, with nearly half of the total international students in the country studying in Pune. Research institutes of information technology, education, management and training attract students and professionals from India and overseas. Several colleges in Pune have student-exchange programmes with colleges in Europe. In the last decade, the city has also seen an IT boom and many companies have their offices set up here.For this reason there has been in increase in population and hence the housing property has also increased drastically. There are some particular Neighborhood which have their prices higher than other.

PROBLEM TO BE SOLVED

In this project I have tried to find the reason and draw conclusion as to why those neighborhoods have higher prices, what are the reason influencing the rates of the flats and houses. The results from this project will help the house buyers to see what price they can expect for a particular neighborhood. Also it will help the house buyer to look for a neighborhood with particular requirements.

Data acquisition and cleaning

The data that will be used for this project will be from following sources:

- Names of the Neighborhood https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune
- The venues in each neighborhood. (FourSquare API)
- The average price of each neighborhood. https://www.99acres.com/property-rates-and-price-trends-in-pune

Other supporting data:

- Coordinates (Geocoder Python)
- GeoJson (http://data.beta.nyc)

The price per sqt feet has be used as it can then be used to estimate the price of flat required by the user. Data for Cost of renting will not be used or mentioned.

Data Collection and using it

- First the neighborhood was scraped from wikipedia
- Then from 99acers data about the price for each area was scraped
- For each neighborhood, call Geocoder Python to get its coordinate.
- For each neighborhood's coordinate, call FourSquare API to get the surrounding venues.

The data obtained from this process will be converted into a 2D DataFrame and then will used to solve the problem specified using Machine Learning,Data Sciece Knowledge

The DataFrame will contain the Neighborhood, cost of land per sqaure foot, top 10 places around it.

After the first processing of data, a second refinement was required because some of the area were too far away and didn’t make any sense to include them. Some of the land price could not be found, so they had to be entered manually.

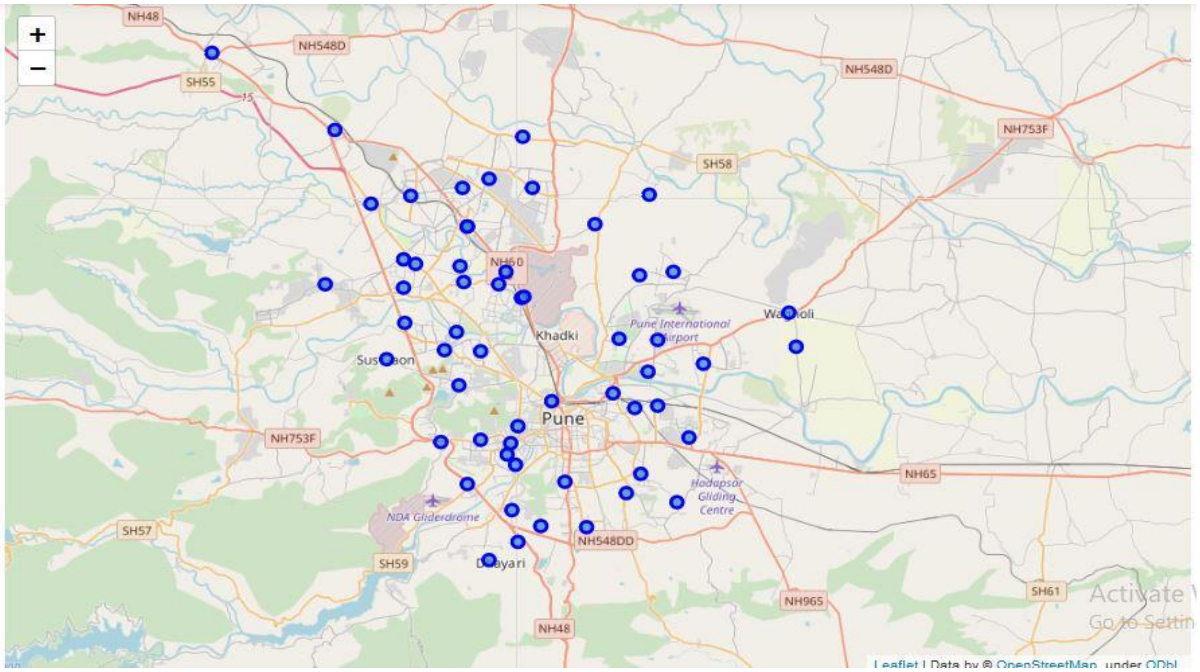
All the locality had a range of price, I used the higher end of the price, so that the user knows the maximum he has to pay for a locality.

CLUSTERING

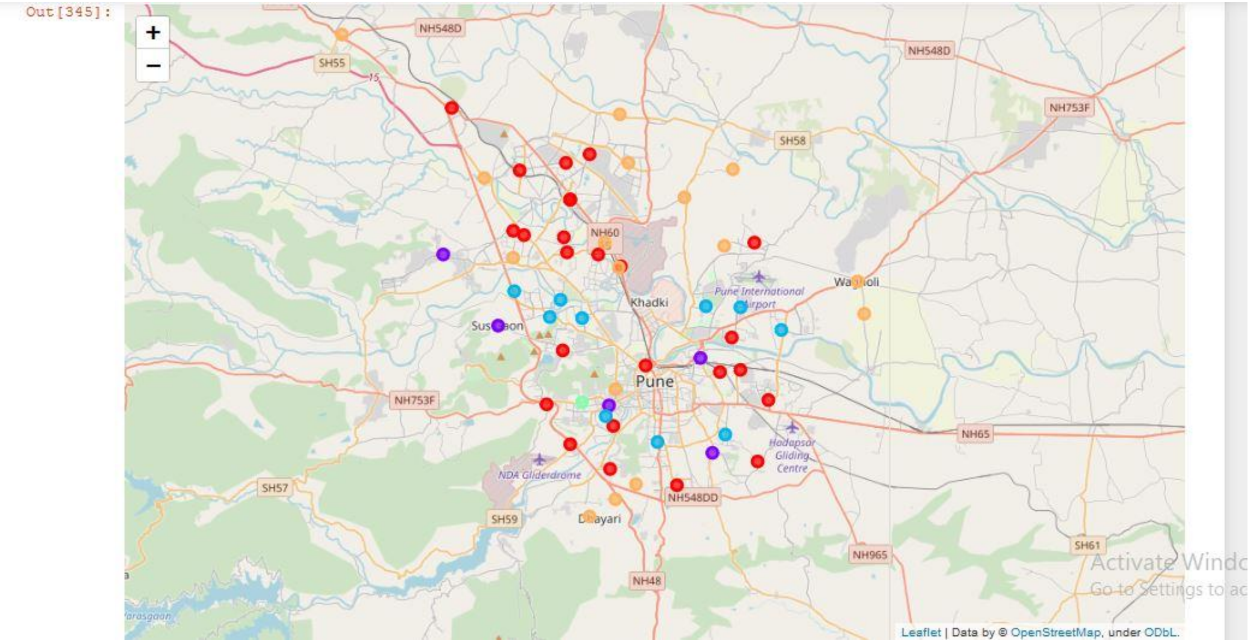
K-mean algorithm was used to cluster the region in 5 clusters, and then the following analysis was performed

- Cluster wise Top 3 places visited.
- Cluster wise prices of locality

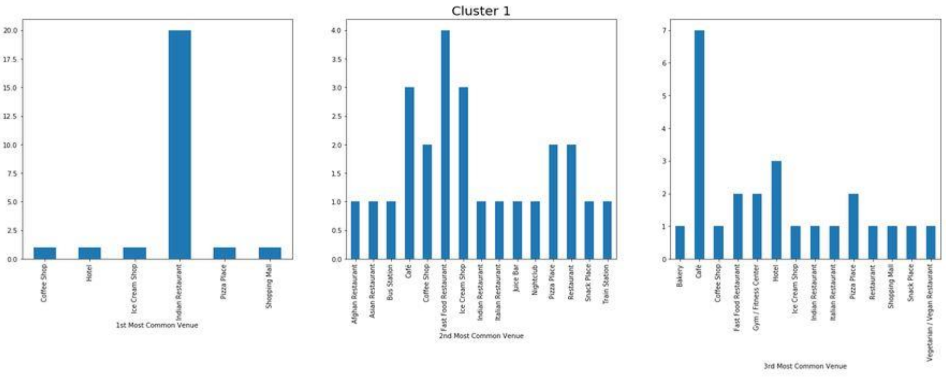
Here is a map of Pune before it was clustered.



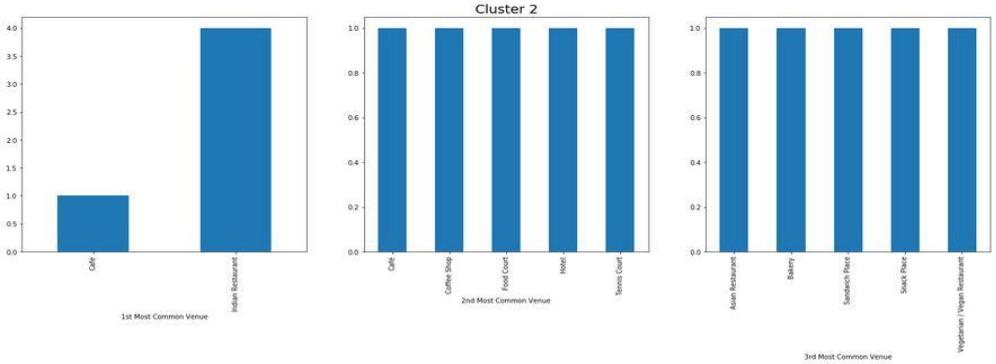
Now after clustering the map looked like this.



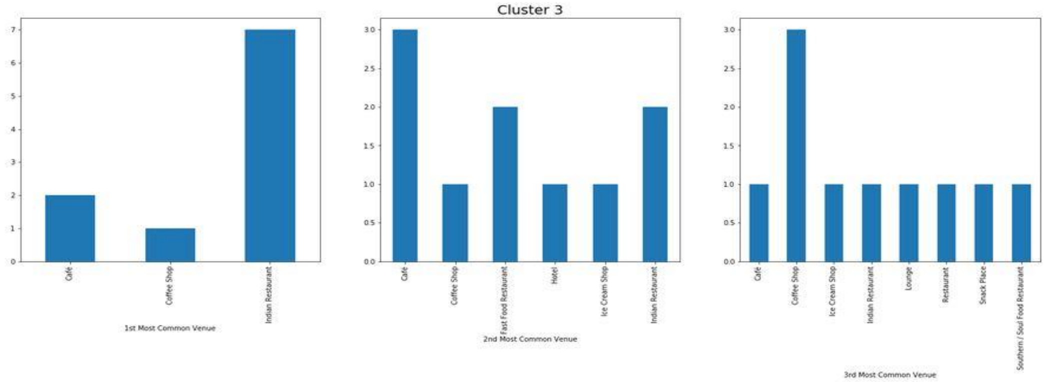
Cluster 1 Top 3 visited places



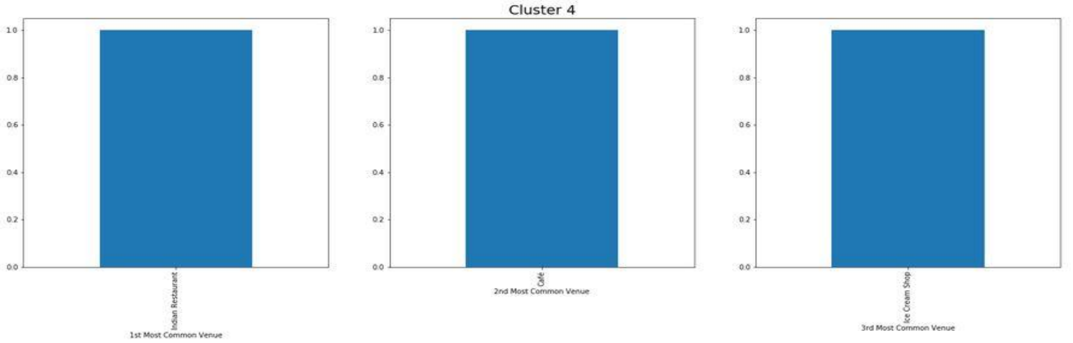
Cluster 2 Top 3 visited places



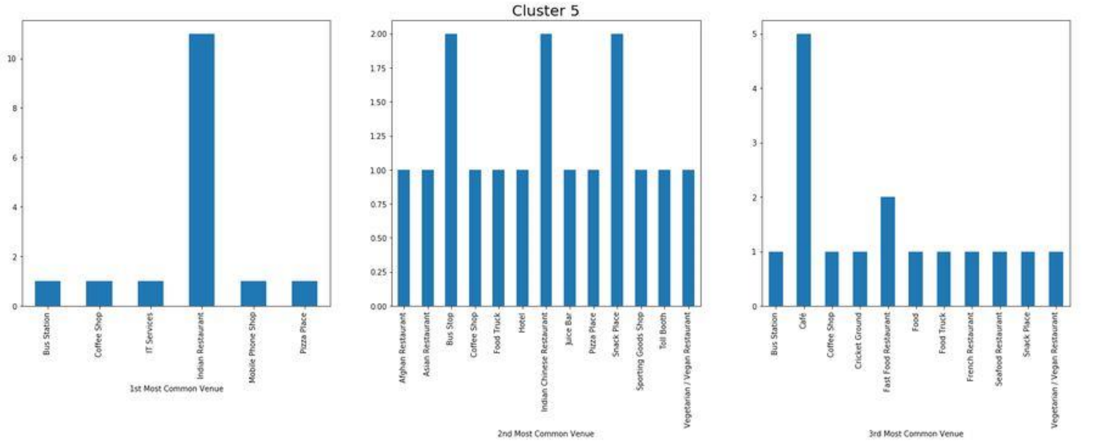
Cluster 3 Top 3 visited places



Cluster 4 Top 3 visited places

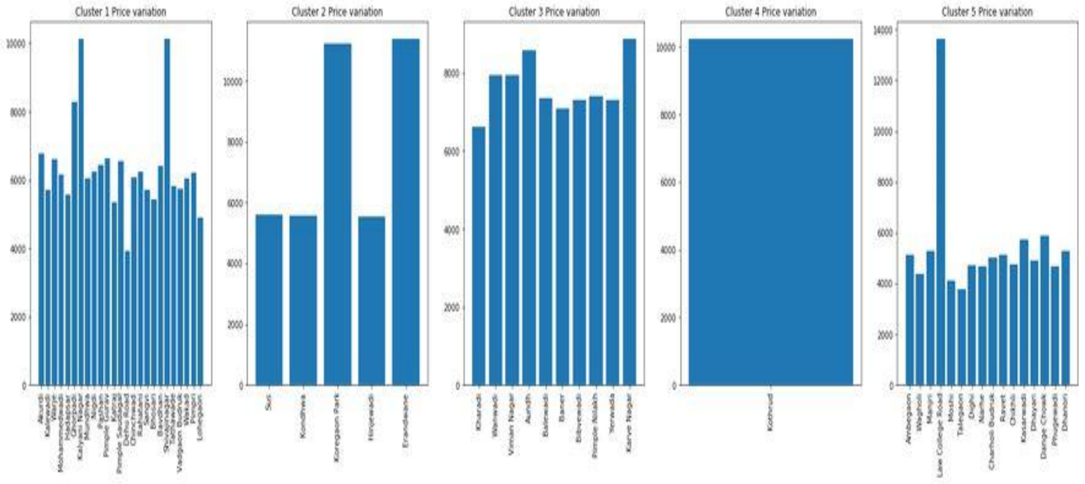


Cluster 5 Top 3 visited places



These graphs and maps allowed us to make some important observation and conclusion which will be mentioned in the ‘Observation and Conclusion’ Part.

Now we look into the cluster vs price graph to see the variation of price



OBSERVATION AND CONCLUSION

From the clustered maps, data set and visualisation, the following conclusions and observation can be drawn

- Cluster 1 and 5 are mostly located far off from the city center with the exception of 2-3 localities. These few localities however have their prices of lands. That is Cluster 1 and 5 if near to city center have the high prices of land.
- Cluster 3 has a mid-priced land rate and are generally somewhere in middle of city center and outskirts of the City. They have a lot restaurant, cafe around them
- Cluster 1 has Shopping malls or other public entertainment places. Also Closer they are to city Centre, the higher are the prices of land
- Pune in general has a lot of restaurants and people love to visit them

From the raw data we scraped out the useful data to create a model which clusters the neighborhood, and helps us to draw conclusions about the land / property rates. We also studied the data to draw other important conclusions. Though this model can be made better by considering the following changes which I could not do in the current project due to lack of time.

- All the restaurants and cafe be included in the same category
- Using more data about the city
- The distance from city center column being added to data frame.