**MUMBAI REAL ESTATE PRICING AND CORRELATION WITH NEARBY FACILITIES**

Capstone Project- Final Report

*By*

**ARJUN C.**

[iamarjunchandra@gmail.com](mailto:iamarjunchandra@gmail.com)

**IBM – COURSERA**

**DATA SCIENCE PROFESSIONAL CERTIFICATE SPECIALIZATION**

**JANUARY 2019**

**TABLE OF CONTENT:**

[**I.** **Introduction:** 2](#_Toc529925814)

[**II.** **Data description:** 3](#_Toc529925815)

[**III.** **Methodology:** 5](#_Toc529925816)

[1. Web Scraping: 5](#_Toc529925817)

[2. Geocoding: 6](#_Toc529925818)

[3. Clustering: 8](#_Toc529925819)

[**IV.** **Results:** 9](#_Toc529925820)

[**V.** **Discussion:** 9](#_Toc529925821)

[**VI.** **Conclusion:** 10](#_Toc529925822)

[**References:** 11](#_Toc529925823)

[**Table of Figures:** 12](#_Toc529925824)

1. **INTRODUCTION**

Mumbai is one of the largest metropolitan cities of India and also the economical capital of India. The price of properties in Mumbai is sky rocket high and the price of some portion of Mumbai is par with that of the rate of properties in many U.S cities. In this capstone project for the ‘IBM Data Science Professional Certificate’ specialization course, I am building a clustering and segmentation model that will sort the various places in Mumbai based on its real estate price and amenities and venues available nearby.

Do the nearby venues and amenities available near a place affect the pricing of that place? What are the posh places in Mumbai? What are the facilities and venues nearby to each place? These are the main questions that will be addressed in this project. By the end of this project, answers for all these questions will be derived.

This project will be useful for a number of people for a number of purposes. Some of the main beneficiaries of this project are:

* People who are interested to buy a property in Mumbai.

In this project, the various places in Mumbai are classified in to various categories depending on their average price per sq.ft. So people can target respective places depending on their budget.

* Companies looking for a perfect place to open their outlet/office.

This project will be useful for those entrepreneurs who are searching for an ideal place in Mumbai as in this project all nearby attractions of different places in Mumbai will be determined. So companies can choose places which have nearby attractions that can suit their operations.

1. **DATA** **DESCRIPTION**

For the project, Data regarding various places in Mumbai is necessary. A real estate website called ‘Makaan’ has a tabular webpage containing details regarding all major places in Mumbai and their real estate price.

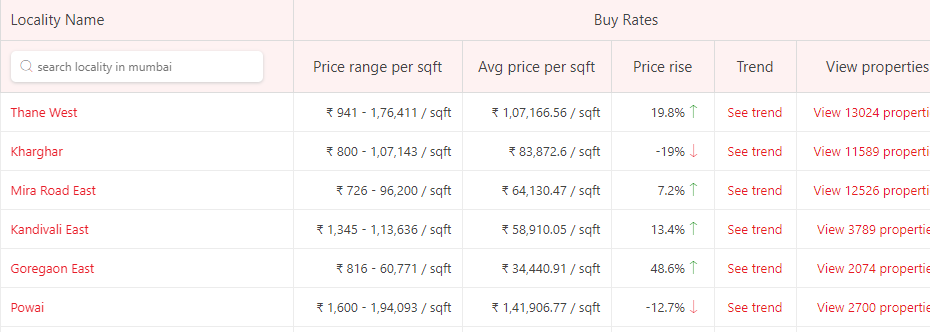


Fig.1: Web page containing necessary data

From the web page shown above, details regarding various places in Mumbai and their average price per sqft. are extracted using HTML parsing techniques. The extracted data frame’s sample is provided below.

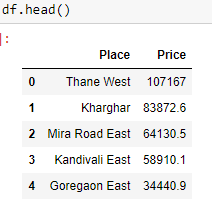


Fig 2. Data Frame’s first 5 rows

The latitude and longitude of each place will be extracted using geopy API, then the nearby attractions of each place will be determined with the help of foursquare API. Segmentation and clustering algorithm will be modeled to group places according to their real estate value.

1. **METHODOLOGY**

In the Project, the first step is collecting necessary information regarding the various places in Mumbai. For this, web scraping is performed and necessary data is extracted from the leading online real estate company in Mumbai, ‘Makaan’. Using the python library ‘BeautifulSoup’, the web page of ‘Makaan’ is parsed and data regarding the various places in Mumbai and the average price per sq.ft for these places are obtained. The price details are parsed into string format using ‘BeautifulSoup’ and it contained symbols and characters. The data is then preprocessed and using ‘regex’ converted to numerical value. The extracted data is then converted into a Data Frame.

Now that the name of various places in Mumbai and their real estate prices are collected, the next step is to find the various attractions/ famous land marks near these places. Various land marks and attractions near a place can be extracted from a geographical API like Google Maps API which gives precise results. However since the project is executed in a testing basis to categorize various places in Mumbai, throughout the project free API’s and services are used. So in the project, ‘FourSquare API’ is used for extracting information regarding nearby attractions of each place. For retrieving nearby venue details from ‘FourSquare API’, a GET request with the coordinates of each place needs to be performed. So before using ‘FourSquare API’, coordinates of each place was required. Geopy API was used for this task. Geopy is a free API that is easy to use as it doesn’t require credentials. However, Geopy failed to get the coordinates of many places which were then removed from the dataset.

Using Folium library, map of Mumbai is rendered and all the places for which the data was available is then marked in the map. Foursuare API returned details regarding all the major attractions near each place, their name and type of venue. These are then attached to the dataset and is analyzed. One hot encoding is performed on the newly added ‘venue category’ column and then the venue categories that are most common in each place is determined. KMeans algorithm is performed to classify the places into three categories.

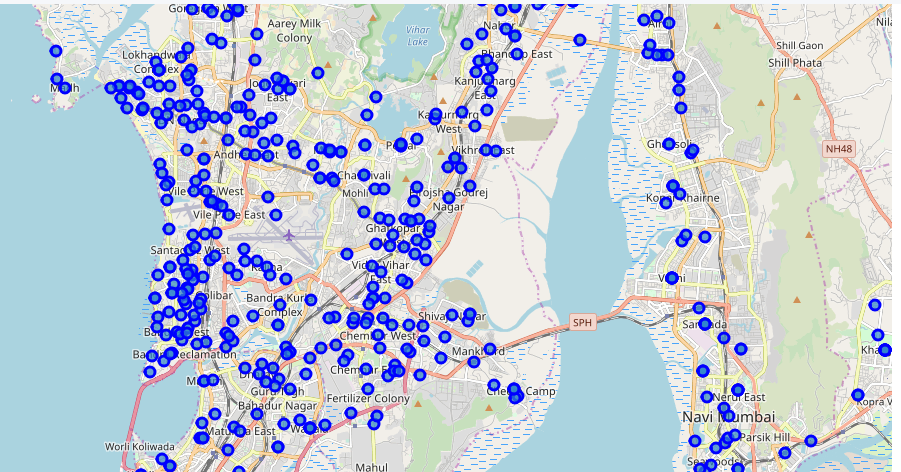


Fig3. Map of Mumbai with main places marked using folium

The kMeans algorithm is imported from the scikit library. Number of clusters was given 3 as the task was to categorize each place into three categories. The newly clustered places are rendered in map using folium, each cluster is represented in each colour for easy identification of clusters.

1. **RESULT**

* All major places in Mumbai were extracted along with their average price per sq.ft.
* For the data available, ‘Charkop Sector 1’ is the place in Mumbai that has the highest real estate value with an average price of Rs.177778 per sq.feet.
* ‘Hendre Pada’ is the region in Mumbai that has the least real estate value.
* From the data collected using Foursquare API, It has been determined that ‘Charkop Sector 1’ is in a commercial area with a lot of eateries and public hangouts
* ‘Hendre Pada’ is a place that’s at the outskirts or Mumbai, far away from the chaos of the city.
* The different land marks of each place is discovered.
* The data is made into three clusters based on their real estate value.

1. **DISCUSSIONS**

* The major challenge was procuring the necessary data.
* The data extracted from web scrapping had many missing columns which were removed from the data set as it wasn’t able to collect the missing data from any source.
* The Geopy geocoder API isn’t the best available. However, the fact that it’s free compensate for the instability of the API considering the scale of the project.
* The project just aims to put forward a insight into the different ways in which data science and machine learning can be incorporated into the field of real estate for improving the business.
* Slight tweaks and tuning in this project can develop this into a useful business model.
* The data frame size will not be constant each time due to the unreliability of geopy API.

1. **CONCLUSION**

It’s unfortunately that the analysis couldn’t produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Doing this project helps practicing every topic in the specialization, andthus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.