Data Science Capstone Project: Optimal Location to startup Tuition Classes or franchise of existing Classes in Mumbai

Published on December 4, 2020

Table of contents

Introduction: Business Problem

In this project we will try to find an optimal location to startup Tuition Classes or franchise of existing Classes. This analysis will be targeted to stakeholders interested in opening a private Tuition Classes or management of existing Classes to expand thier bussiness & revenue by opening new branch/franchise in Mumbai, India.

Since there are lots of Classes in Mumbai we will try to detect locations that are not already crowded with Classes and having good number of existing Schools & Colleges to ensure sufficient student admistions and less bussiness competition. We will also consider Commercial Property Rates of the locations.

We will use our data science powers to generate a few most promissing Area based on these criteria so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decission are:

- Number of existing Schools & Colleges in the area.
- Number of existing Tutorials & Classes in the area.
- Commercial Property Rates of the area.

Following data sources will be needed to extract/generate the required information:

- List of all Areas in Mumbai along with their geographical coordinates (latitude & longitude) & location is web-scraped from **Wikipedia** website.
- Commercial property rates of each area is collected from major Indian **online real estate platforms** (like MagicBricks & 99Acres).
- Neighbourhood Schools, Colleges and Classes in each Area will be retrieved using **Foursquare API** (a location data provider).

Methodology

In this project we will direct our efforts on detecting areas of Mumbai having high number of Schools & Colleges, low number of existing private tuition Classes and affordable property rates.

Steps:

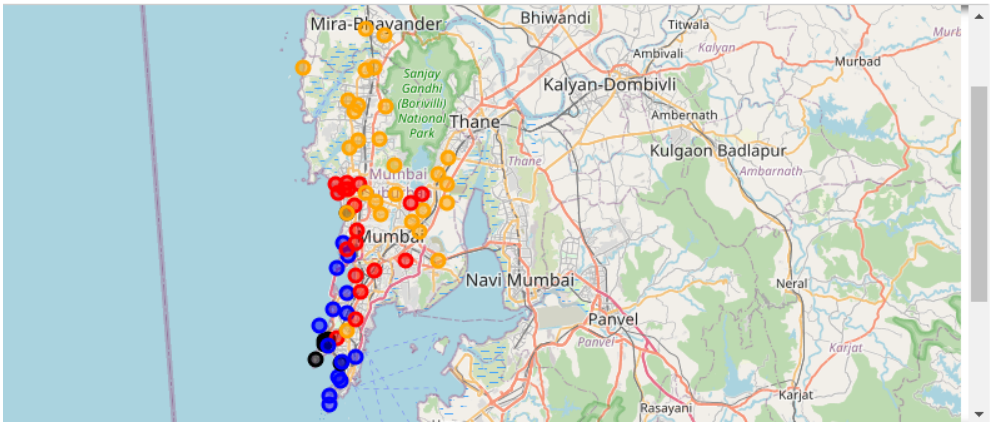
Make dataframe containing area name, location, latitude & longitude data (all 67 Areas) web-scrapped from Wikipedia page (**https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai#Other**) and commercial property rate (per square feet) from major Indian online real estate platforms (like MagicBricks & 99Acres).

	Area	Location	Latitude	Longitude	Rate
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400	18000.0
1	Chakala, Andheri	Western Suburbs	19.111388	72.860833	14750.0
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373	18000.0
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.827210	18000.0
4	Lokhandwala	Andheri,Western Suburbs	19.130815	72.829270	18000.0

Plot all Areas on Map showing respective Property Rates using python **folium** library.

Color code:

Property Rate	Color
High Property Rate	Black
Medium High Property Rate	Blue
Medium Low Property Rate	Red
Low Property Rate	Orange



Collect JSON from **Foursquare API** containing data of all nearby Schools, Colleges & Classes in each of the Areas in radius of 2 kilometres and extract required data from JSON & put it in the dataframes. Then groupby dataframes using Area name such that we will get

(1) Total number of schools/colleges in each Area.

	Area	Venue
0	Aarey Milk Colony	10
1	Altamount Road	63
2	Amboli	59
3	Amrut Nagar	38
4	Asalfa	44

(2) Total number of Classes in each Area.

	Area	Venue
0	Altamount Road	9
1	Amboli	21
2	Amrut Nagar	8
3	Asalfa	18
4	Ballard Estate	4

Calculate **Business Opportunity Index** for each Area. A good Business Opportunity Index means there are more Schools/Colleges and less Classes in a Area.

- *Business Opportunity Index = (Total no. of Schools & Colleges) / [(Total no. of Classes & Tutorials) + 1]*

Merge all data into a single dataframe such as:

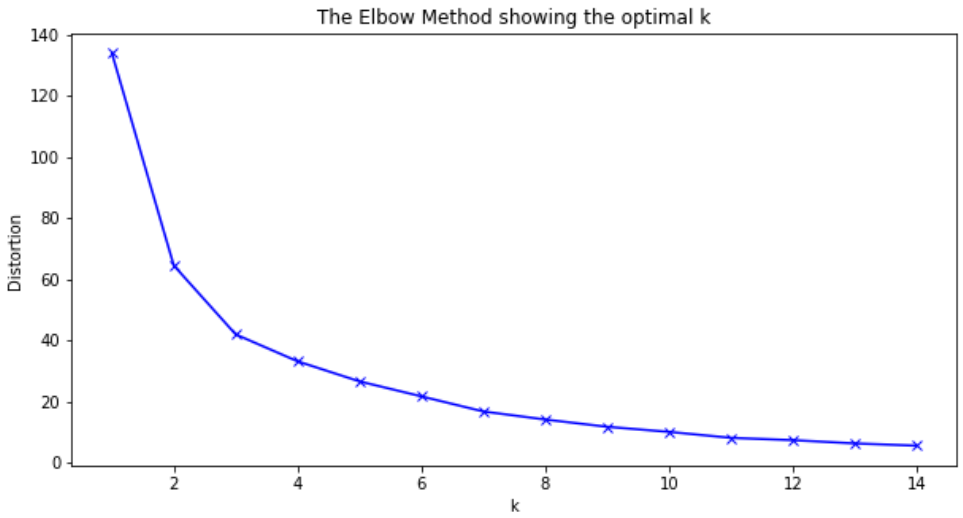
	Area	Latitude	Longitude	Rate	Schools_Colleges	Tutorials_Classes	Startup Opportunity Index
0	Parel	18.99	72.840	18000.0	40.0	2.0	13.333
1	Kanjurmarg	19.13	72.940	10550.0	13.0	0.0	13.000
2	Worli	19.00	72.815	27500.0	25.0	1.0	12.500
3	Churchgate	18.93	72.820	28500.0	61.0	4.0	12.200
4	Ballard Estate	18.95	72.840	29000.0	60.0	4.0	12.000

Cluster Areas using **k-Means clustering** algorithm. Parameters use for clustering each Area will be:

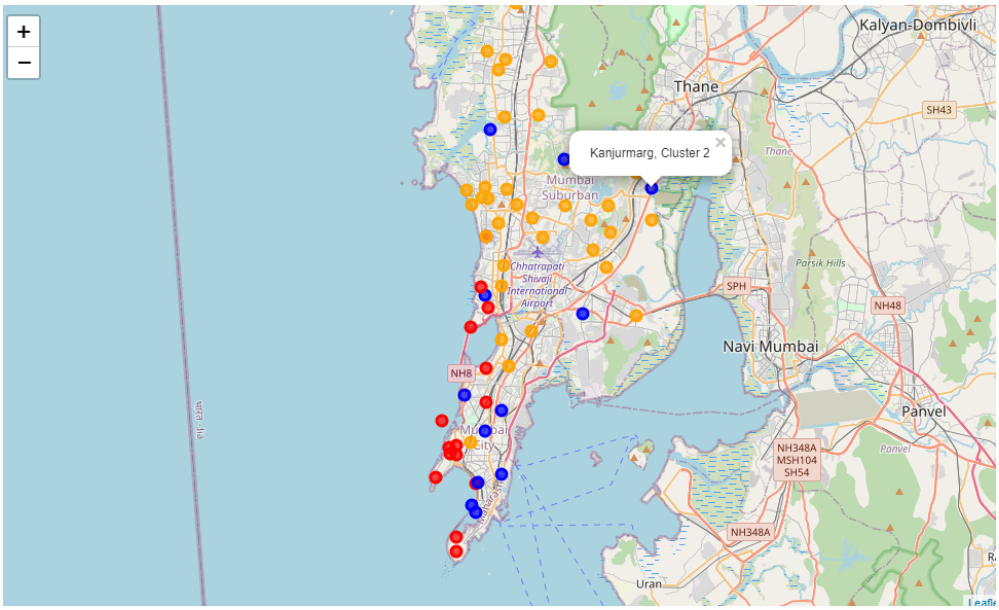
(1)Property rates.

(2) Business Opportunity Index. (A good Business Opportunity Index means there are more Schools/Colleges and less Classes in a Area)

Use **Elbow method** for deciding number of clusters.



The Elbow Method suggests 2 as the optimum number of clusters, as the distortion in the data drops suddenly at 2 clusters and very slowly decreases from there. But considering insights from the exploratory data analysis, we will proceed with grouping the data into 3 clusters, as it will be easier to gather insights that way.



Cluster insights:

Cluster	No. of Areas	Average Property Rate	Average No. of Schools & Colleges	Average No. of Tutorials & Classes	Startup Opportunity Index
0	40	15616.250000	42.625000	13.575000	3.267425
1	15	33983.333333	44.333333	7.600000	6.093867
2	12	21687.500000	43.000000	3.166667	10.754750

All Clusters have similar 'Average No. of Schools & Colleges'

- Cluster 0: Low 'Average Property Rate', More 'Average No. of Tutorials & Classes', Low 'Startup Opportunity Index'.
- Cluster 1: High 'Average Property Rate', Moderate 'Average No. of Tutorials & Classes', Medium 'Startup Opportunity Index'.
- Cluster 2: Medium 'Average Property Rate', Less 'Average No. of Tutorials & Classes', High 'Startup Opportunity Index'.

Cluster	Property Rate	Startup Opportunity Index	No. of Bussiness Competitors
0	Cheap	Low	High
1	Expensive	Medium	Medium
2	Mid-Range	High	Low

Selecting "Cluster 2" as it has less business competitors & moderate/affordable property rates.

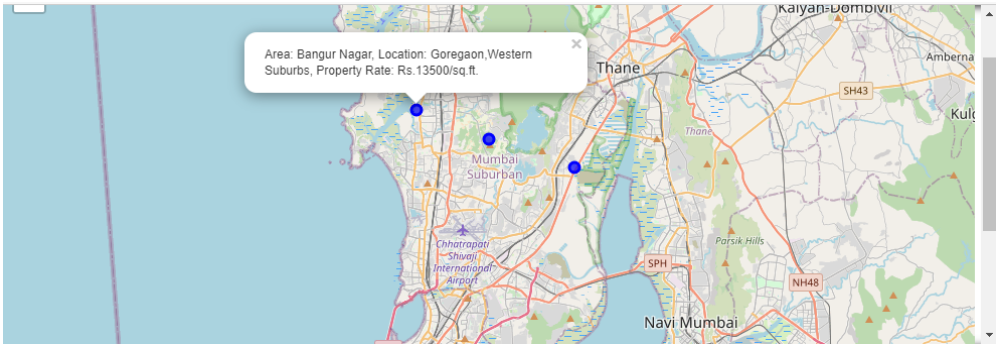
```
Cluster_2
```

	Area	Latitude	Longitude	Rate	Schools_Colleges	Tutorials_Classes	Startup Opportunity Index	Cluster
0	Parel	18.990000	72.840000	18000.0	40.0	2.0	13.333	2
1	Kanjurmarg	19.130000	72.940000	10550.0	13.0	0.0	13.000	2
2	Worli	19.000000	72.815000	27500.0	25.0	1.0	12.500	2
3	Churchgate	18.930000	72.820000	28500.0	61.0	4.0	12.200	2
4	Ballard Estate	18.950000	72.840000	29000.0	60.0	4.0	12.000	2
5	Nariman Point	18.926000	72.823000	33500.0	56.0	4.0	11.200	2
6	Bangur Nagar	19.167362	72.832252	13500.0	50.0	4.0	10.000	2
7	Aarey Milk Colony	19.148493	72.881756	12950.0	10.0	0.0	10.000	2
10	Marine Lines	18.944700	72.824400	29000.0	76.0	7.0	9.500	2
11	Dagdi Chawl	18.977129	72.829131	17000.0	44.0	4.0	8.800	2
12	Chembur	19.051000	72.894000	19250.0	26.0	2.0	8.667	2
14	Pali Naka	19.062742	72.829396	21500.0	55.0	6.0	7.857	2

Now, for more short listing of Areas apply a limit on Property prices depending on the investing capacity of the stakeholder. Here we are taking Rupees 15,000/- per square feet.

	Area	Location	Latitude	Longitude	Rate	Schools_Colleges	Tutorials_Classes	Startup Opportunity Index	Cluster
0	Kanjurmarg	Eastern Suburbs	19.130000	72.940000	10550.0	13.0	0.0	13.0	2
1	Bangur Nagar	Goregaon,Western Suburbs	19.167362	72.832252	13500.0	50.0	4.0	10.0	2
2	Aarey Milk Colony	Goregaon,Western Suburbs	19.148493	72.881756	12950.0	10.0	0.0	10.0	2

Top 3 optimal Areas (1) Kanjurmarg. (2) Bangur Nagar. (3) Aarey Milk Colony. on Map



Results

Optimal Locations to startup Tuition Classes or franchise of existing Classes in Mumbai having high number of Schools & college and less existing Classes and affordable property rates are

- Kanjurmarg
- Bangur Nagar
- Aarey Milk Colony

Discussion

Our analysis shows that:

- Property rates are increasing as we move from North to South region of Mumbai. 'Cluster 2' is in central region of Mumbai, therefore having mid-range Properties Rates.
- All 3 Clusters are having similar number of Schols & Colleges. But 'Cluster 2' is having lowest number of existing Classes & Tutorials, resulting in less Bussiness competition, therefore having good Bussiness Opportunity.

Suggestions for Future research:

- Population Density of each Areas can be web-scrapped from Wikipedia and can be a parameter/feature in clustering locations.
- Areas with high Population Density should be given preferences.

Conclusion

In this project, we have gone through the process of identifying the business problem , specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters(no. of cluster decided by Elbow Method) based on their similarities, and lastly providing recommendations to the stakeholders interested in opening a private Tuition Classes or management of existing Classes to expand thier bussiness & revenue by opening new branch/franchise in Mumbai. To answer the business question that was raised in the introduction session , the answer proposed by this project is: The Areas in cluster 2 are the most optimal locations to open a new Classes and with a properties rates limit depending on investment capacity of stakeholders(Rupees 15,000/- per square feet), top 3 optimal Areas are (1) Kanjurmarg. (2) Bangur Nagar. (3) Aarey Milk Colony.

Published by

Sahil Ansari

#Data_Science_Capstone_Project: Optimal Location to startup Tuition Classes or franchise of existing Classes in Mumbai. This project is a part of "IBM Data Science Professional Certificate" on Coursera.

#datascience #dataanalytics #python #machinelearning #datavisualization #coursera

Sahil Ansari