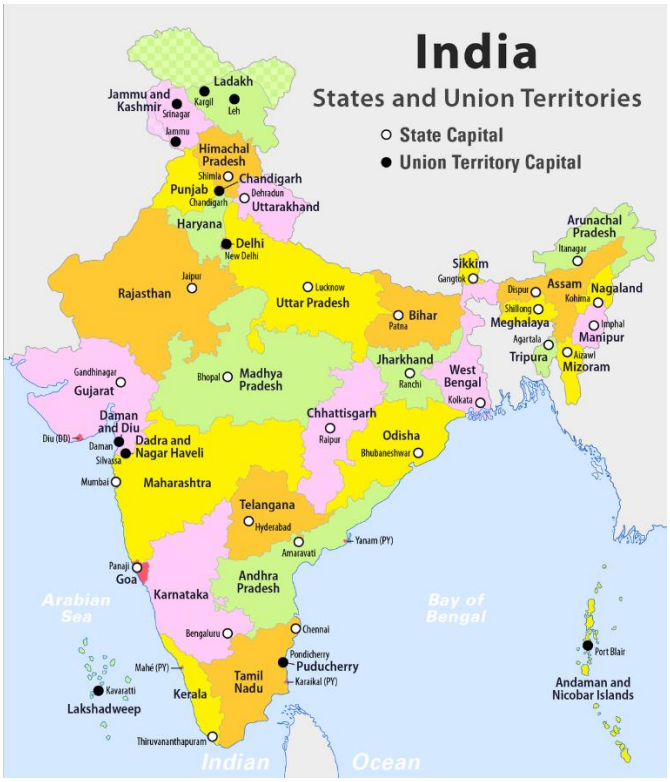
ENTERTAINMENT CENTER CLASSIFICATION IN INDIAN CITIES

Avilash Banerjee (Developer Associate @ SAP Labs)

COURSERA CAPSTONE PROJECT

IBM DATA SCIENCE SPECIALISATION

CONTENTS



Introduction

Overview

India is the second most populated country in the world and its diversity is without equal. India's ever growing cities are *ever-changing* and *ever-expanding.* No two cities are alike in India, neither in climate or culture or language etc. Also various factors determine the livability in these cities with Mumbai being the financial capital and city of dreams while Bangalore is considered the utopia for IT engineers. In such a vast and diverse (and with ever increasing population) it becomes very difficult to classify Indian cities based on a fixed criteria. How many things need to be considered -> population, climate, economic sustainability etc ? This becomes a nightmare for investors who want a proper classification on a fixed criteria to invest in some real estate in the city and are worried about their returns on investments. Sure there will be articles to overcome such a demand but as already explained above, the ever changing nature of the cities can become a bit of a roadblock and can lead of costly miscalculations and a loss of millions. For example Bangalore was a sleepy, peaceful city a few decades ago. Suddenly it has changed into an IT powerhouse with a bustling city life with pubs and nightclubs adorning its Silicon-Valley texture. Thus there is a need for real time data analysis to find a point-of-time analytical model to determine the criteria provided. Hence we come across our problem statement and my solution provided for such.

Problem Statement

Entertainment center classification in Indian cities.

Business Need

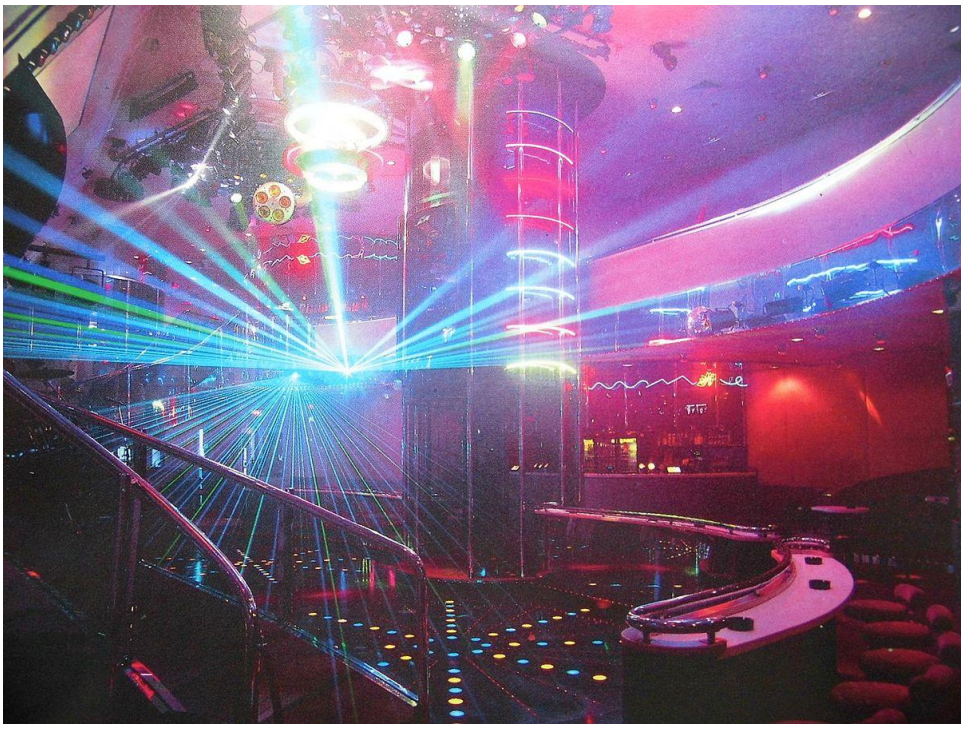
A point of time exploratory data analysis is needed to classify Indian cities to find the best viable cities for entertainment centers to be built/invested in. As explained above, due to the ever-changing nature of Indian cities a point of time data classification model will help us to understand when and which city is viable for investment by a business owner. For example every one knows that Goa and Mumbai are the well known cities for nightclubs and pubs, but how do we determine the rest ? Some of the cities in our list will certainly shock our business partners !! 😉

Business Assumption

We will be working on Indian cities classified according to their population. We will use authentic dataset provided by **Wikipedia** to do our data analysis due to its easy access and no attached monetary charge. Other factors of Indian cities will not be taken into account(economy, climate etc) as we have found that Kolkata and Bangalore are both hubs of entertainment centers despite been two very different cities in terms of climate, culture, finance etc. All other assumptions will be mentioned in code while performing the analysis. **Foursquare** will be our primary location data provider as we can easily access trending venues classified by categories (such as nightclubs, bars etc)

Problem Description

Lets discuss our problem before diving in. Lets assume an investor (lets call him Mr. Sharma) wants to open and invest in a chain of nightclubs in Indian cities. Mr. Sharma is a PIO(person of Indian origin) who is visiting India for the first time. His knowledge of India comes from social media and the internet in general. He believes that he can open from Goa/Mumbai and carry on from there. But when he visits India, he gets shocked. He cannot believe the diversity and the pure challenge in front of him from a business perspective. He consults a lot of people, online forums, news articles and the end result is that he is very confused. Some people are asking him to open in Chennai while others prefer Kochi/Hyderabad/Delhi. So he reaches out to data scientists for help in giving him real time classification of Indian cities (the parameter been entertainment venues such as nightclubs) so that he can choose 7-10 cities of his choice to invest in (safely). So we have decided to solve his problem using exploratory data analysis to provide him the classification he asks for. Also the entire model built for him will be real time (based on foursquare's ever changing dataset and Wikipedia's ever active community). I believe as an aspiring data scientist, the data can be as close to perfection as possible.



Dataset Description

Now lets discuss the dataset for a bit. Please understand that the dataset description here will be completely initial (with prerequisites and assumptions) and it will change in the due course of this project.

We will begin with our main resource page :
https://en.wikipedia.org/wiki/List_of_cities_in_India_by_population
This page accurately depicts our initial problem requirement of a list of Indian cities classified in order of their population. Out of 319 cities, we will only be considering the population centers with above 1,000,000 population (i.e. in 2011 it is 46 cities) as Mr Sharma wants populated cities to draw attraction from its native population rather than depending on tourist populations(think post Covid fear !!). The city names will be classified according to their latitude and longitude so that they can be displayed on the map of India. Some online service (like geocoder) will be used to get the city coordinates. This dataset will be changed throughout the course of the project via modification and update to create our final cluster dataset based on our data analysis and k-means clustering that we aim to complete. The final dataset can effectively predict the near-perfect scenarios containing the best 7-10 cities to open up the entertainment centers. Exciting isn't it !!

Initial analysis and planning

Once we have hold of our dataset containing City name, latitude and longitude we can dive deep into our data analysis. We will begin by querying foursquare for entertainment venues in these cities (preferably trendy) like nightclubs, bars etc. With out retrieved list we can choose top 5-10 entertainment venues for the cities. Then we can classify similar cities based on these venues and find the similarities between them. Finally using a clustering algorithm we can cluster similar cities together to get the idea of opening entertainment centers in these cities. The clusters can also be referenced along with its visual representation to get an idea of how the cities were classified and on what basis is the classification. This will provide a flexibility to the business owner in the sense that he may choose to open his chain of venues at a city more likely to have restaurants and breweries than to open one in nightclubs and bars.

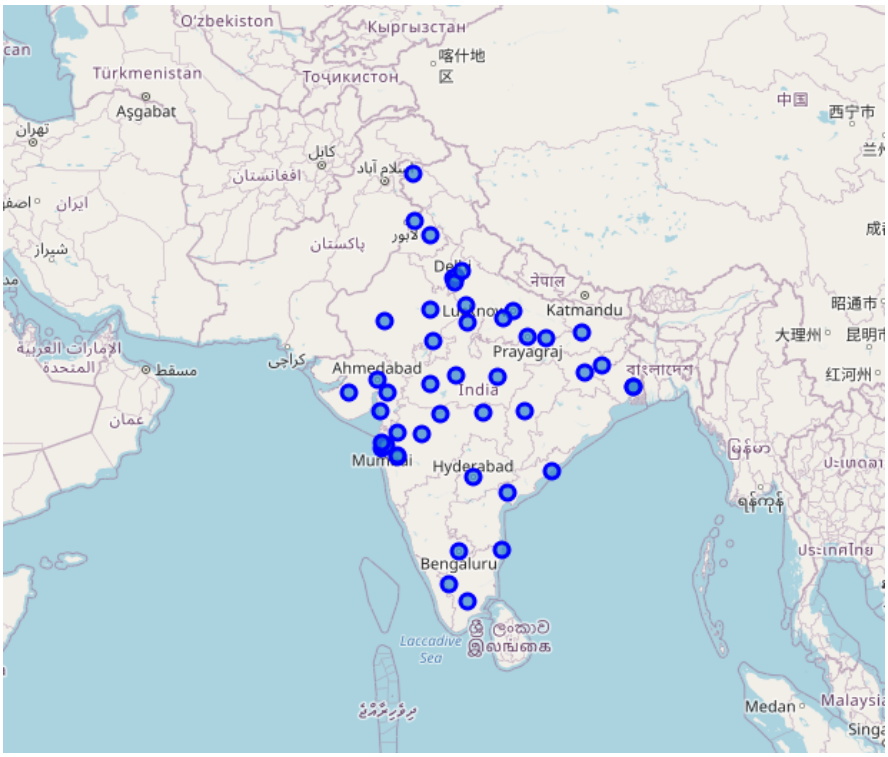
Please understand that these initial assumptions will be used as a base for further assumptions while performing the analysis due to the ever-changing nature of data. (Unfortunately cities cannot be static hence we will challenge ourselves with dynamic data )

Rank	City	Population (2011)[3]	Population (2001)	State or union territory
1	**Mumbai**	12,442,373	11,978,450	Maharashtra
2	**Delhi**	11,007,835	9,879,172	Delhi
7	**Kolkata**	4,486,679	4,572,876	West Bengal
6	**Chennai**	4,681,087	4,343,645	Tamil Nadu
3	**Bangalore**	8,436,675	4,301,326	Karnataka
4	**Hyderabad**	6,809,970	3,637,483	Telangana
5	Ahmedabad	5,570,585	3,520,085	Gujarat
12	Kanpur	2,767,031	2,551,337	Uttar Pradesh
9	Pune	3,115,431	2,538,473	Maharashtra

* A small clipping of the dataset to be used.

Methodology

Now lets dive into our work. Firstly we shall webscrap the source https://en.wikipedia.org/wiki/List_of_cities_in_India_by_population and get the bare minimum dataframe containing only City and Population(2011) data and populated with cities having a pop. above 1,000,000(our only prerequisite criteria). So far its easy. But now we need to plot it on our map. For this we shall have to get the latitudes and longitudes of each city. As Google Maps have made their API usage chargeable, we shall use python's Nominatim geolocator library to get us our coordinates. It may not be as sophisticated and trustworthy as Google Maps but being an open source project, it is by far the best option available for a Capstone Project. Now with our coordinates data we can easily plot the map of India using code.



Zoomed out map of India generated from code

Now we have to cluster the cities based on our inherent criteria (entertainment centers like nightclubs). Lets use Foursquare's documentation to help us. As per this resource https://developer.foursquare.com/docs/build-with-foursquare/categories/ , the master venue category Nightlife Spot and its child categories like Bars, Nightclubs etc fulfills our requirement. Using the API reference doc we create a dataset of the categories we need from the Foursquare data source.

Now begins the fun part. For each city we have to use Foursquare's explore API endpoint to get all the venues in each city (restricting radius to 20km from city center as Mr Sharma would want his investments to be near to the city center rather than on the outskirts) which have the same venue category id as in our category dataset created from Nightlife Spot category. A bit of a complex coding

algorithm certainly, but very powerful as to create a dataset containing the cities along with their venues. After one hot encoding and cleaning our dataset, our new dataframe is ready.

Using best practices, we run a loop to get top 10 venues from each city and arrange them in a dataframe alongside their city name, population and coordinates. Now our final dataframe is close to being ready.

Now comes the tough part. Due to this being a capstone project and due to the nature of the complexities involved, we choose the number of kclusters at a basic 5 (after various trial and error methods used) and a new dataframe is created that includes the cluster as well as the top 10 venues for each city. This dataframe is now plotted on the map to get the varying clusters.

Cluster 0

Cluster 1

Cluster 2

Cluster 3

Cluster 4

Brilliant isn't it ? This is almost perfect in the eyes of a data scientist (an aspiring one of course). Now let us dive into the results in the next section.

Results

As is apparent in previous section, our 5 cluster approach has provided dividends. Now let us come to the hard part – interpreting results !! In this section we will only decipher the results. The main discussion will take place in the next section

Let us take a look at each cluster:-

Cluster 0 :

	City	1st Most C	2nd Most Co	3rd Most C	4th Most C	5th Most Com	6th Most Com	7th Most Com	8th Most C	9th Most C	10th Most Common Venue
38	Gwalior	Pub	Gastropub	Bar	Lounge	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Nightclub	Night Market
39	Jabalpur	Pub	Nightclub	Bar	Beach Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Night Mar	Lounge

Consisting of only two cities (and well below the population expectations of Mr Sharma) the cluster will not be considered. The popular venues as per their rankings are pubs, gastropubs, bars etc. Nightclubs are sparse and is only popular in Jabalpur it seems.

Cluster 1 :

	City	1st Most C	2nd Most	3rd Most Cor	4th Most Con	5th Most Comm	6th Most Com	7th Most Com	8th Most C	9th Most Con	10th Most Common Venue
21	Ludhiana	Bar	Gastropub	Hookah Bar	Lounge	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Nightclub
27	Rajkot	Bar	Brewery	Sake Bar	Lounge	Whisky Bar	Sports Bar	Speakeasy	Pub	Nightclub	Night Market
30	Varanasi	Bar	Beer Bar	Brewery	Gastropub	Sake Bar	Hotel Bar	Pub	Night Mar	Whisky Bar	Sports Bar
31	Srinagar	Bar	Beach Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Nightclub	Night Market	Lounge
35	Navi	Bar	Beach Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Nightclub	Night Market	Lounge
43	Madurai	Bar	Brewery	Hotel Bar	Cocktail Bar	Wine Bar	Beer Garden	Dive Bar	Gastropub	Hookah Bar	Beer Bar

Consisting of 6 cities it is also not in the list of likely been chosen by Mr Sharma. These cities mostly to have a variety of bars. Nightclubs are not amongst the top 5 most popular venues in these cities.

Cluster 2 :

	A	B	C	D	E	F	G	H	I	J	K	L	M
		City	1st Most C	2nd Most	3rd Most C	4th Most Com	5th Most Com	6th Most Con	7th Most C	8th Most C	9th Most C	10th Most Common Venue	
24		Ranchi	Hotel Bar	Lounge	Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Nightclub	Night Market	
45		Kota	Hotel Bar	Bar	Lounge	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Nightclub	Night Market	

Once again, a cluster consisting of only 2 cities where nightclubs seem to be the 9th most popular venues. Not worth considering.

Cluster 3 :

	City	1st Most C	2nd Most Con	3rd Most Comm	4th Most Comm	5th Most Comm	6th Most Comm	7th Most Comm	8th Most Com	9th Most Com	10th Most Common Venue
0	Mumbai	Bar	Lounge	Brewery	Pub	Beer Bar	Nightclub	Whisky Bar	Sports Bar	Speakeasy	Sake Bar
1	Delhi	Bar	Lounge	Pub	Beer Garden	Cocktail Bar	Nightclub	Whisky Bar	Sports Bar	Speakeasy	Sake Bar
11	Kanpur	Bar	Hookah Bar	Gastropub	Sports Bar	Pub	Lounge	Wine Bar	Dive Bar	Cocktail Bar	Hotel Bar
14	Thane	Bar	Pub	Lounge	Brewery	Nightclub	Wine Bar	Dive Bar	Gastropub	Hookah Bar	Hotel Bar
16	Visakhapa	Bar	Lounge	Beer Garden	Gastropub	Pub	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Nightclub
18	Patna	Bar	Whisky Bar	Pub	Hotel Bar	Lounge	Sports Bar	Speakeasy	Sake Bar	Nightclub	Night Market
22	Agra	Bar	Lounge	Hotel Bar	Sports Bar	Pub	Whisky Bar	Speakeasy	Sake Bar	Nightclub	Night Market
23	Nashik	Wine Bar	Lounge	Sports Bar	Bar	Brewery	Cocktail Bar	Dive Bar	Gastropub	Hookah Bar	Hotel Bar
26	Meerut	Bar	Lounge	Hookah Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Nightclub	Night Market
28	Kalyan-Dc	Bar	Lounge	Pub	Hookah Bar	Sports Bar	Gastropub	Nightclub	Wine Bar	Beer Garden	Dive Bar
29	Vasai-Vira	Bar	Lounge	Sports Bar	Dive Bar	Hookah Bar	Whisky Bar	Speakeasy	Sake Bar	Pub	Nightclub
32	Aurangab	Bar	Lounge	Sports Bar	Brewery	Dive Bar	Gastropub	Hotel Bar	Nightclub	Whisky Bar	Speakeasy
33	Dhanbad	Bar	Lounge	Beach Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Nightclub	Night Market
34	Amritsar	Bar	Beer Bar	Hotel Bar	Brewery	Lounge	Sake Bar	Hookah Bar	Nightclub	Whisky Bar	Sports Bar
36	Allahabad	Bar	Hotel Bar	Pub	Hookah Bar	Lounge	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Nightclub
41	Vijayawac	Bar	Brewery	Sake Bar	Pub	Hotel Bar	Lounge	Whisky Bar	Sports Bar	Speakeasy	Nightclub
44	Raipur	Pub	Bar	Sports Bar	Cocktail Bar	Dive Bar	Gastropub	Hookah Bar	Sake Bar	Nightclub	Night Market

A mega cluster consisting of 17 cities. Hmm..certainly one of great interest to Mr Sharma. It seems to have the cities with the major population centers also(Mumbai, Delhi etc). While bars and lounges seem to be the order of business here, there is also mentions of nightclubs amongst their popular venues (albeit infrequent). Mr Sharma would like to study this cluster in deep and would consider our analysis.

Cluster 4 :

	City	1st Most Cor	2nd Most Co	3rd Most Comr	4th Most C	5th Most Commc	6th Most Commc	7th Most Comr	8th Most Con	9th Most Com	10th Most Common Venue
2	Bangalore	Brewery	Lounge	Pub	Bar	Hookah Bar	Wine Bar	Cocktail Bar	Dive Bar	Gastropub	Hotel Bar
3	Hyderaba	Lounge	Hotel Bar	Nightclub	Brewery	Gastropub	Hookah Bar	Pub	Beer Garden	Cocktail Bar	Sports Bar
4	Ahmedab	Hotel Bar	Lounge	Speakeasy	Nightclub	Hookah Bar	Bar	Whisky Bar	Sports Bar	Sake Bar	Pub
5	Chennai	Lounge	Bar	Pub	Nightclub	Whisky Bar	Cocktail Bar	Hotel Bar	Sports Bar	Speakeasy	Sake Bar
6	Kolkata	Nightclub	Pub	Lounge	Hookah Ba	Bar	Gastropub	Karaoke Bar	Brewery	Dive Bar	Cocktail Bar
7	Surat	Sports Bar	Cocktail Bar	Lounge	Brewery	Pub	Night Market	Bar	Whisky Bar	Speakeasy	Sake Bar
8	Pune	Lounge	Pub	Bar	Nightclub	Brewery	Gastropub	Hookah Bar	Cocktail Bar	Beach Bar	Whisky Bar
9	Jaipur	Lounge	Bar	Hotel Bar	Hookah Ba	Nightclub	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub
10	Lucknow	Lounge	Hookah Bar	Nightclub	Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Night Market
12	Nagpur	Lounge	Bar	Pub	Nightclub	Beach Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Night Market
13	Indore	Lounge	Pub	Hookah Bar	Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Nightclub	Night Market
15	Bhopal	Cocktail Bar	Pub	Hookah Bar	Hotel Bar	Nightclub	Lounge	Bar	Night Market	Whisky Bar	Sports Bar
17	Pimpri-Ch	Lounge	Pub	Bar	Brewery	Nightclub	Cocktail Bar	Gastropub	Hookah Bar	Beach Bar	Whisky Bar
19	Vadodara	Lounge	Sports Bar	Bar	Speakeasy	Hotel Bar	Whisky Bar	Sake Bar	Pub	Nightclub	Night Market
20	Ghaziabac	Lounge	Sports Bar	Hookah Bar	Bar	Pub	Nightclub	Hotel Bar	Whisky Bar	Speakeasy	Sake Bar
25	Faridabad	Lounge	Nightclub	Bar	Pub	Gastropub	Karaoke Bar	Beer Garden	Dive Bar	Cocktail Bar	Hookah Bar
37	Howrah	Nightclub	Pub	Lounge	Hookah Ba	Bar	Brewery	Gastropub	Night Market	Whisky Bar	Sports Bar
40	Coimbato	Nightclub	Bar	Lounge	Beach Bar	Whisky Bar	Sports Bar	Speakeasy	Sake Bar	Pub	Night Market
42	Jodhpur	Hotel Bar	Lounge	Cocktail Bar	Hookah Ba	Bar	Night Market	Whisky Bar	Sports Bar	Speakeasy	Sake Bar

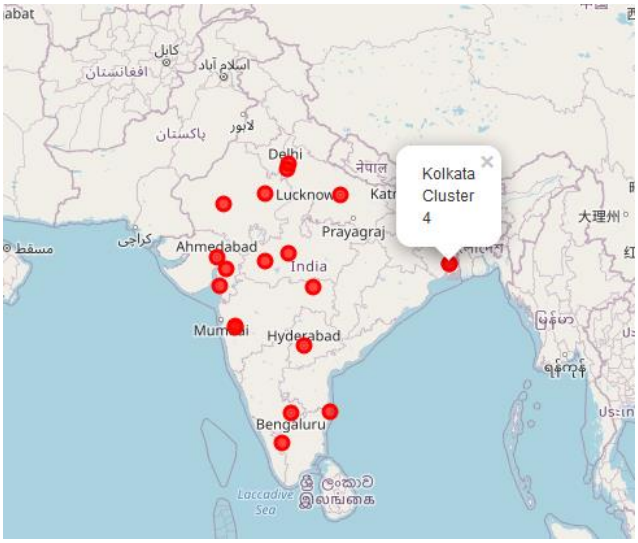
The final cluster and also a mega cluster by the looks of it. This cluster consists of 19 cities and immediately attracts Mr Sharma's interest. As we can see, nightclubs are not only amongst the most popular venues here, you can also pick out 5-7 cities just by looking at the dataset's 1st and 2nd common venues (Kolkata,Howrah,Coimbatore,Faridabad,Hyderabad etc.). This gives Mr Sharma a diverse option of cities across India (north, south east and west) and also with varying climates, population, economic viability etc. This will certainly be our recommended cluster.

Discussion

Now we have reached almost to the end of our report. Based on our result analysis above, Mr Sharma finds clusters 3 and 4 most viable (and we agree with him). Now based on our own analysis we provide our recommendations. As we can see cluster 3 has influential and economic powerhouse cities like Mumbai and Delhi. It also has major population centers. It will be economically viable if Mr Sharma does decide to invest in this cluster. Surely it seems bars and lounges seem more popular here then nightclubs, but such an assumption can be deceiving also. Being economical powerhouses these cities can be a great investment option for a nightclub chain. A drawback can be that the choice of cities may be a bit less. Our recommended cities based on the results shown will be Mumbai, Delhi, Thane, Kalyan-Dombivli and Amritsar. As already mentioned above the number of cities to be chosen can a bit low. But they are certainly viable economically, and seeing the map, quite well connected with each other too.

Now lets come to our recommended cluster. As already mentioned above, the choices are more and also nightclubs appear to be amongst the popular venues here. The cities are spread far across entire India taking in its diversity and culture. They are also placed in major economic zones with higher HDIs (compared to national average) and have major centers of population. Our recommended cities will be Kolkata, Howrah, Coimbatore, Faridabad, Hyderabad, Lucknow and Chennai. Based on our own analysis and recommendation I would opt out for Ahmedabad as it is situated in a dry state (alcohol is not permitted for consumption/sale in the state of Gujarat except by trusted government establishments and restaurants/hotels). The only drawback here that we can envision is the cities are a bit too spread out and Mr Sharma may not want an investment chain so far distributed from each other.

With this we come to the conclusion of our report and leave our findings to him to take his final decision.



Cluster 4 in all its glory

Conclusion

So we have come to the end of our report. This report and its preparation has been a great learning experience. The ability to easily cluster cities and provide a fast, point of time data analysis to meet business requirements is of great importance in the field of data science and analysis. I want to thank IBM and Coursera for collaborating on such a wonderful Data Science Specialization capped with a challenging Capstone Project to always keep us interested. I also want to extend my thanks to my learning peers who have not only fairly judged my assignment submissions, but have also helped out with helpful tips and tricks and suggestions through the discussion forums. Lastly I hope the fictional Mr Sharma (can be anyone really) really makes use of our recommendations and suggestions provided above and wish him the best in his investment venture of a string of nightclubs in India's cities.

Bibliography

- Wikipedia (https://en.wikipedia.org/wiki/Main_Page)
- Foursquare Developer Docs (https://developer.foursquare.com/docs/)
- Python documentations for Nominatim (https://nominatim.org/release-docs/develop/)