ENTERTAINMENT CENTER CLASSIFICATION IN INDIAN CITIES

Avilash

COURSERA CAPSTONE PROJECT

IBM DATA SCIENCE SPECIALISATION

CONTENTS

India
States and Union Territories
○ State Capital
● Union Territory Capital
Ladakh
Kargil
Leh
Jammu and Kashmir
Srinagar
Jammu
Himachal Pradesh
Shimla
Chandigarh
Punjab
Chandigarh
Dehradun
Uttarakhand
Haryana
Delhi
New Delhi
Arunachal Pradesh
Itanagar
Sikkim
Gangtok
Rajasthan
Jaipur
Lucknow
Uttar Pradesh
Bihar
Patna
Dispur
Assam
Nagaland
Kohima
Shillong
Meghalaya
Imphal
Manipur
Agartala
Tripura
Aizawl
Mizoram
Jharkhand
Ranchi
West Bengal
Kolkata
Gandhinagar
Gujarat
Bhopal
Madhya Pradesh
Chhattisgarh
Raipur
Odisha
Bhubaneshwar
Daman and Diu
Diu (DD)
Daman
Dadra and Nagar Haveli
Silvassa
Mumbai
Maharashtra
Telangana
Hyderabad
Yanam (PY)
Amaravati
Panaji
Goa
Andhra Pradesh
Arabian Sea
Karnataka
Bay of Bengal
Chennai
Bengaluru
Port Blair
Pondicherry
Puducherry
Mahé (PY)
Karaikal (PY)
Tamil Nadu
Kavaratti
Lakshadweep
Kerala
Andaman and Nicobar Islands
Thiruvananthapuram
Indian Ocean

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

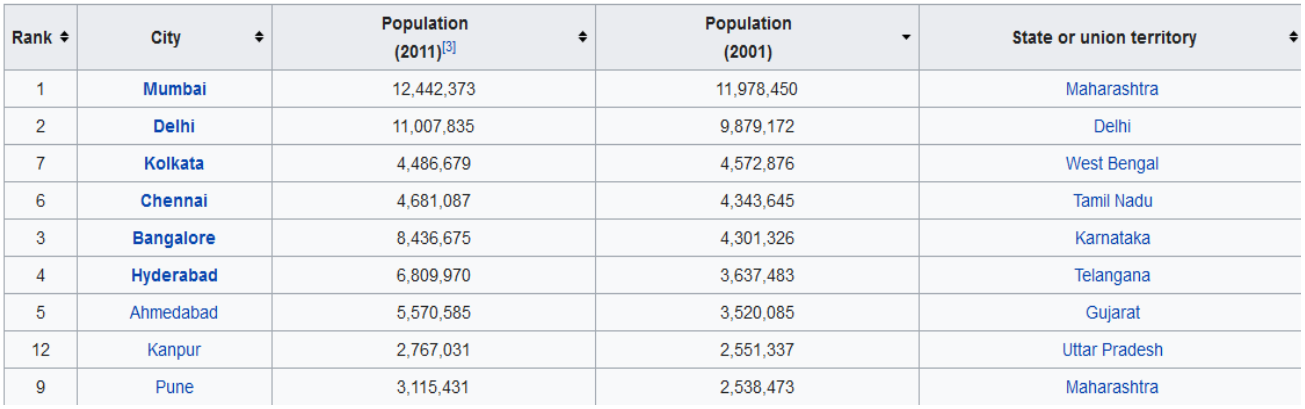
Now lets discuss the dataset for a bit. Please understand that the dataset description here will be completely initial (with prerequisites and assumptions) and it will change in the due course of this project.

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.