Similarity and difference between major Indian cities

Utkarsh Mishra

1. Introduction

1.1 Background

This project is focused on India. India is a country in South Asia. It is the second-most populous country, the seventh-largest country by land area, and the most populous democracy in the world. India is a federal union comprising 28 states and 8 union territories. With this much diversity, Indian cuisine consists of a wide variety of regional and traditional cuisines. Given the range of diversity in soil type, climate, culture, ethnic groups, and occupations, these cuisines vary substantially from each other, using locally available spices, herbs, vegetables, and fruit. Indian foodways have been influenced by religion, in particular Hindu cultural choices and traditions. In this project I looked for differences and similarities between all major cities in india.

1.2 Problem

Using the data available is it possible to determine whether different cities have something in common? If someone wants to move/ expand his business to another city what should he look for?

1.3 Interest

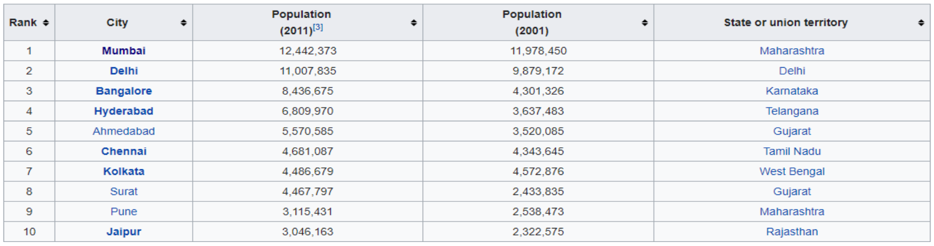
This will be useful for people expanding their business to other cities. Anyone moving to another city can see how similar another city is to their own city.

2. Data acquisition and cleaning

2.1 Data sources

For this project we need three datasets.

1. This project is focused on India so firstly we need data of all the major cities. To collect the name and state of major cities in India this data from Wikipedia is used

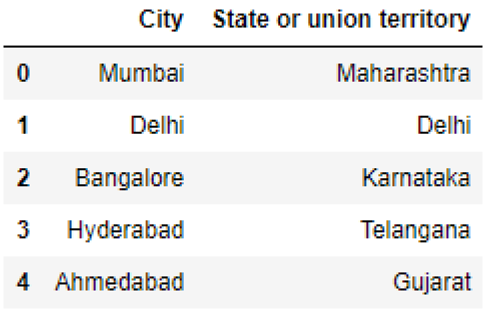


Rank ⬍	City ⬍	Population (2011)[3] ⬍	Population (2001) ⬍	State or union territory ⬍
1	**Mumbai**	12,442,373	11,978,450	Maharashtra
2	**Delhi**	11,007,835	9,879,172	Delhi
3	**Bangalore**	8,436,675	4,301,326	Karnataka
4	**Hyderabad**	6,809,970	3,637,483	Telangana
5	Ahmedabad	5,570,585	3,520,085	Gujarat
6	**Chennai**	4,681,087	4,343,645	Tamil Nadu
7	**Kolkata**	4,486,679	4,572,876	West Bengal
8	Surat	4,467,797	2,433,835	Gujarat
9	Pune	3,115,431	2,538,473	Maharashtra
10	**Jaipur**	3,046,163	2,322,575	Rajasthan

2. Now that we have names and state of the major cities we can use geocoder to get latitude and longitude of all the cities
3. Finally we use Foursquare to collect relevant data of all the major venues

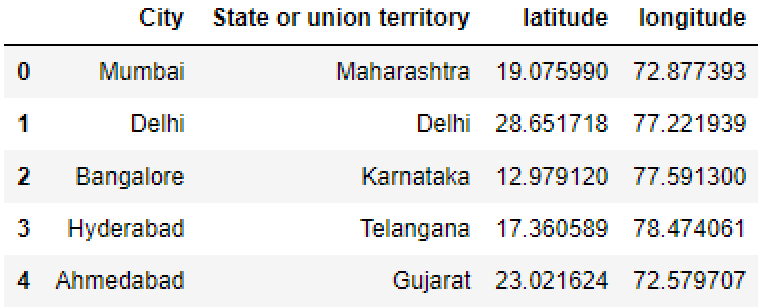
2.2 Data cleaning and Feature selection

After importing city data to dataframe using pandas library, firstly unnecessary columns need to be removed so in the first step of data cleaning we remove population and rank data from our data to get city name and state. This dataframe contains 319 cities so we need to limit the no. of cities to 100 to get only the major cities of India.



	City	State or union territory
0	Mumbai	Maharashtra
1	Delhi	Delhi
2	Bangalore	Karnataka
3	Hyderabad	Telangana
4	Ahmedabad	Gujarat

Now to use Foursquare API to get relevant data, location of all the cities needs to be added in the dataframe. We use geocoder to get the location of all cities and add this data as latitude and longitude to our dataframe. When we use geocoder we did not get the location of some of the cities and this will result in error when we proceed further so this also needs to be removed. After this process we end up with 90 cities.



	City	State or union territory	latitude	longitude
0	Mumbai	Maharashtra	19.075990	72.877393
1	Delhi	Delhi	28.651718	77.221939
2	Bangalore	Karnataka	12.979120	77.591300
3	Hyderabad	Telangana	17.360589	78.474061
4	Ahmedabad	Gujarat	23.021624	72.579707

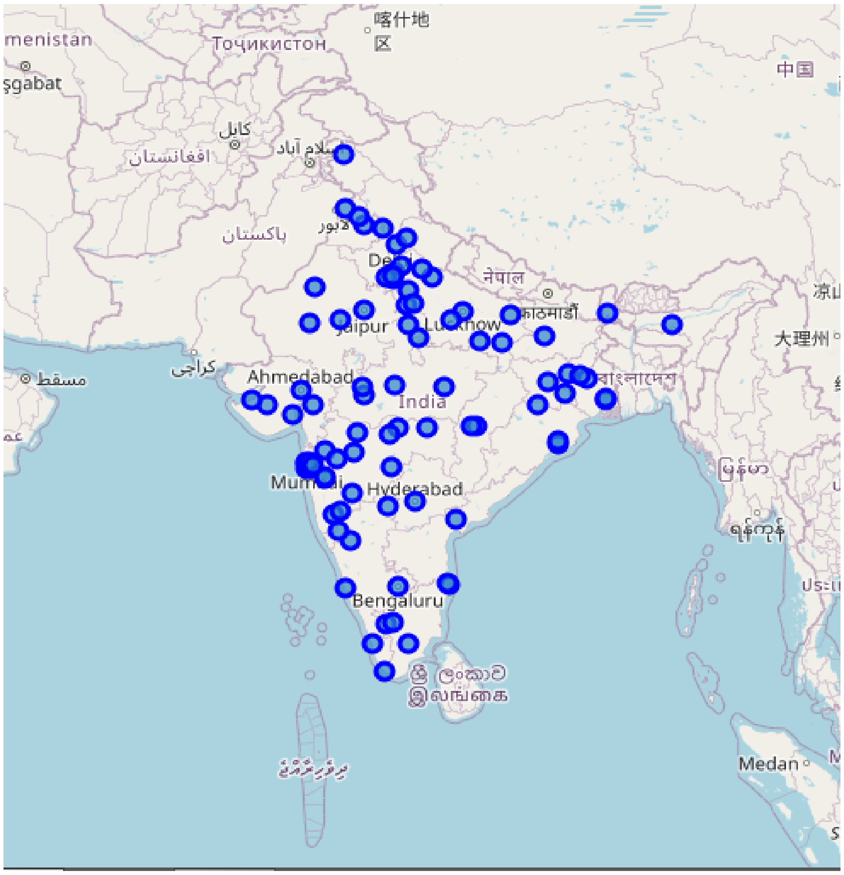
After cleaning location data our dataframe will have 4 columns name, state, latitude and longitude of the city. We use the Foursquare API to get the top 100 major venues of all cities. This includes restaurants, bars, cafes, historic sites, parks and many other popular venues.

We use the city name, venues category, state their location as our feature to further evaluate the similarity and other parameters between cities.

3. Exploratory Data Analysis

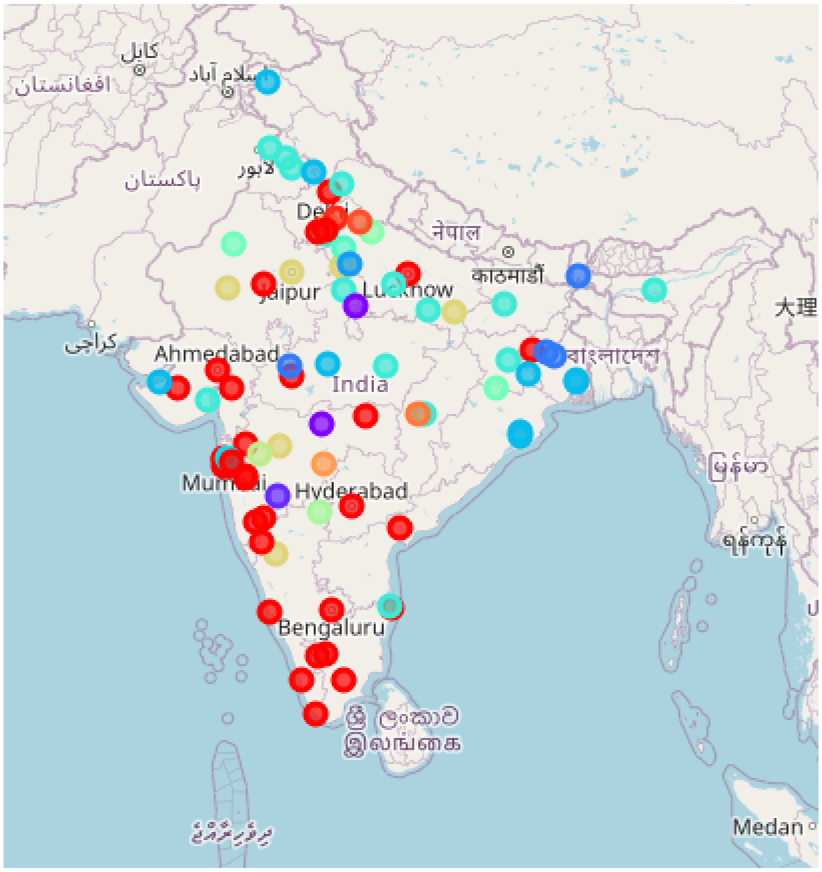
3.1 Visualization of cleaned data

After cleaning the data, the first step towards data analysis is to visualize available data and then determine what further data is needed to complete the analysis. It is very easy to visualize data on a map using the folium library. So I created a map using the folium library with India on focus.



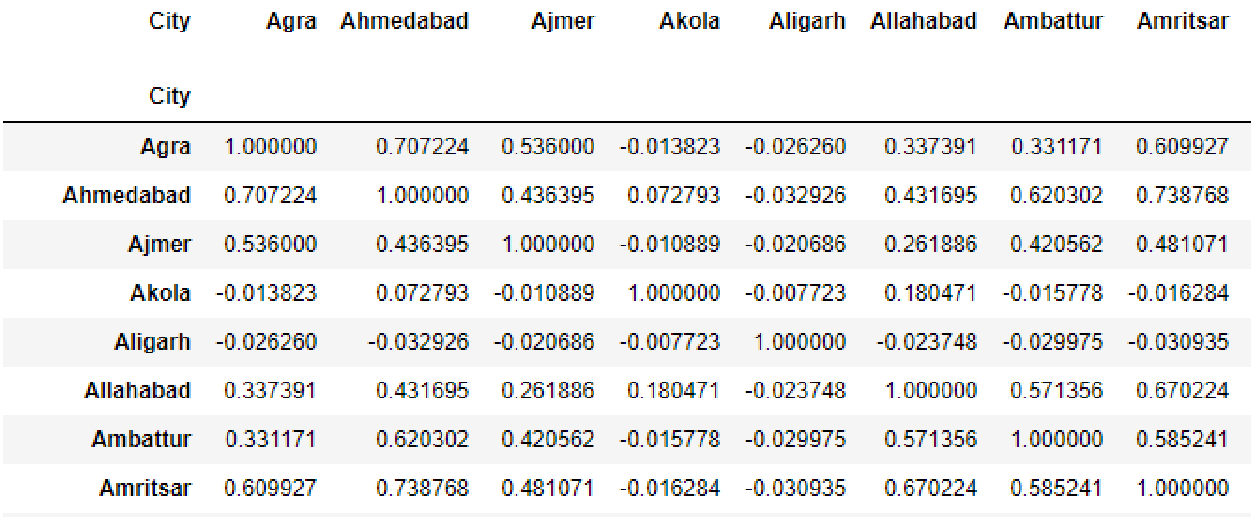
3.2 Getting data for cluster analysis and cluster analysis

After visualizing the location data, for cluster analysis we need a little more data such as popular venues in cities, their location and their category. For cluster analysis we transform data and get top venues for each city. Finally we perform cluster analysis for different values of clusters and retrieve the result as a map.



3.3 Correlation between different cities

To compare cities with each other, we need a dataset which contains the venue category for every city and that's why india_grouped is the best choice for it, but before we can use it for our purpose we need to transform it. Transposing dataframe to get each column as a city for correlation analysis. After performing correlation analysis we get the correlation matrix. Now we can see which city is similar to others by comparing two cities.



City	Agra	Ahmedabad	Ajmer	Akola	Aligarh	Allahabad	Ambattur	Amritsar
City								
Agra	1.000000	0.707224	0.536000	-0.013823	-0.026260	0.337391	0.331171	0.609927
Ahmedabad	0.707224	1.000000	0.436395	0.072793	-0.032926	0.431695	0.620302	0.738768
Ajmer	0.536000	0.436395	1.000000	-0.010889	-0.020686	0.261886	0.420562	0.481071
Akola	-0.013823	0.072793	-0.010889	1.000000	-0.007723	0.180471	-0.015778	-0.016284
Aligarh	-0.026260	-0.032926	-0.020686	-0.007723	1.000000	-0.023748	-0.029975	-0.030935
Allahabad	0.337391	0.431695	0.261886	0.180471	-0.023748	1.000000	0.571356	0.670224
Ambattur	0.331171	0.620302	0.420562	-0.015778	-0.029975	0.571356	1.000000	0.585241
Amritsar	0.609927	0.738768	0.481071	-0.016284	-0.030935	0.670224	0.585241	1.000000

4. Results and Observations

Venue Analysis

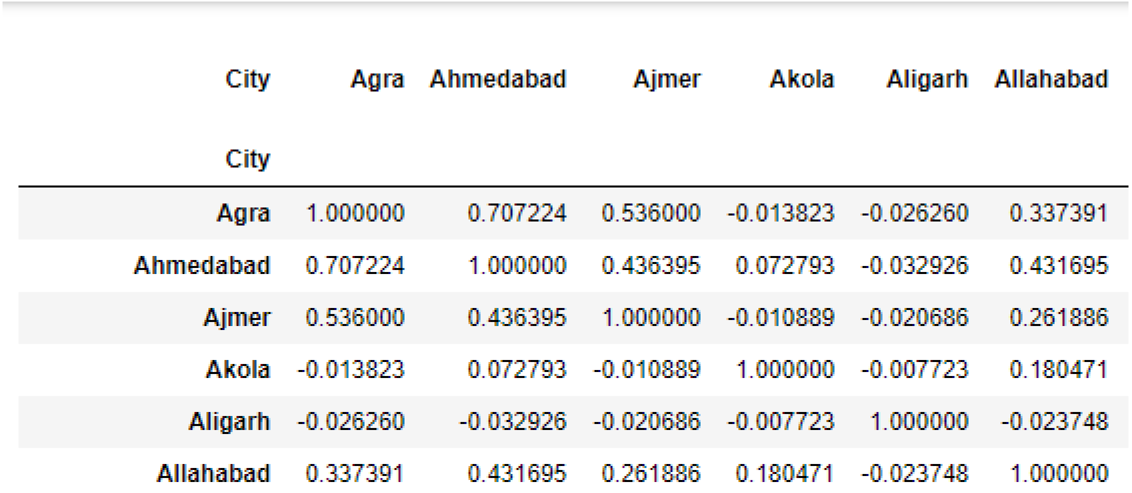
After performing data analysis we can clearly see that most of the cities are similar to one another as we get most of the cities in only three clusters. We see the top ten venues that are most common are.

Venues

1. Indian Restaurant
2. Hotel
3. Café
4. Pizza Place
5. Train Station
6. Fast Food Restaurant
7. Shopping Mall
8. Multiplex
9. Restaurant
10. Coffee Shop

Correlation Analysis

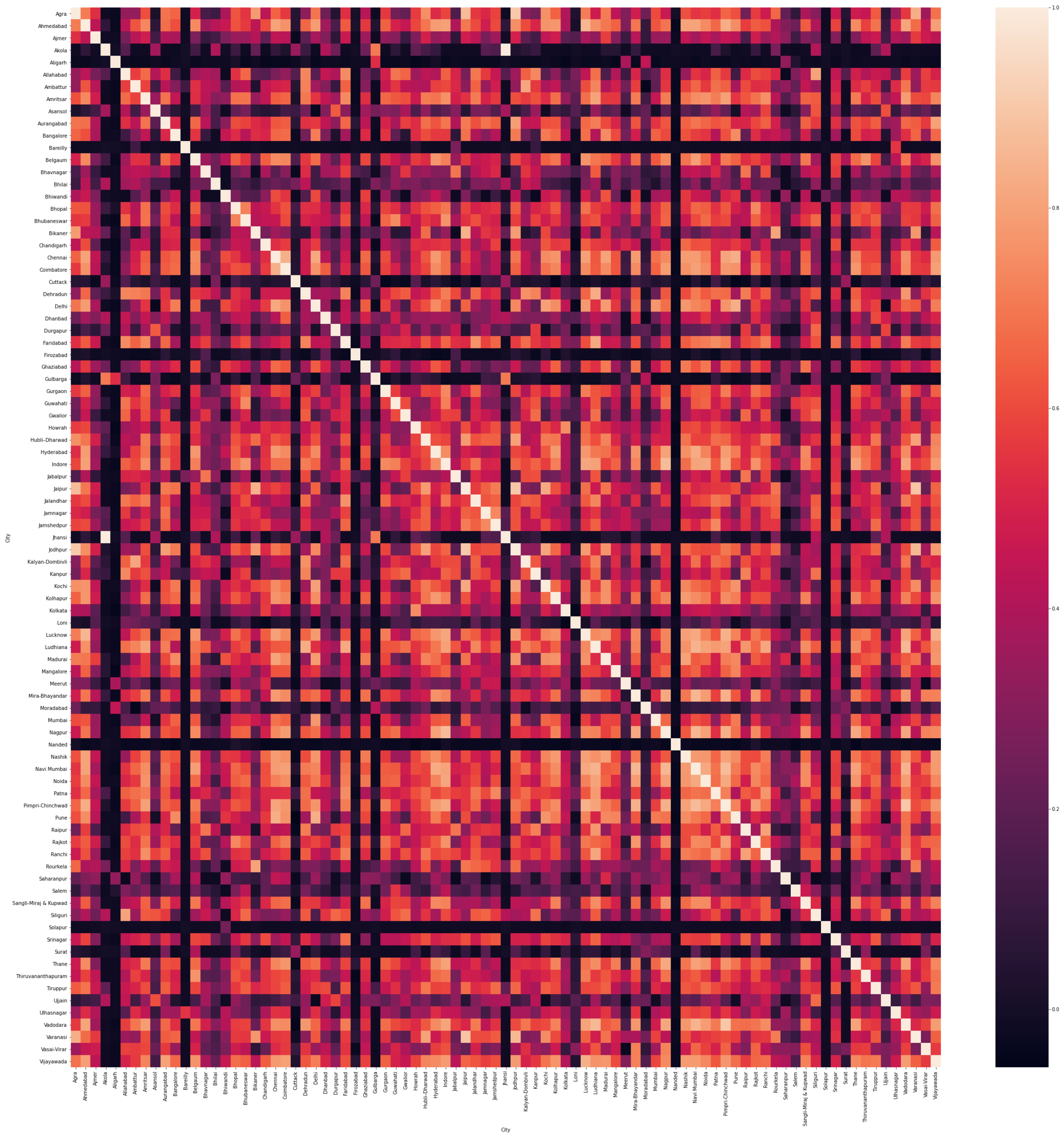
Even though India has very high diversity in terms of people, Indian cities are more or less very similar to each other nowadays. All the basic amenities are available in almost every city except for a few outliers.



City	Agra	Ahmedabad	Ajmer	Akola	Aligarh	Allahabad
City						
Agra	1.000000	0.707224	0.536000	-0.013823	-0.026260	0.337391
Ahmedabad	0.707224	1.000000	0.436395	0.072793	-0.032926	0.431695
Ajmer	0.536000	0.436395	1.000000	-0.010889	-0.020686	0.261886
Akola	-0.013823	0.072793	-0.010889	1.000000	-0.007723	0.180471
Aligarh	-0.026260	-0.032926	-0.020686	-0.007723	1.000000	-0.023748
Allahabad	0.337391	0.431695	0.261886	0.180471	-0.023748	1.000000

As from the analysis it is very clear that almost all the cities have very high correlation with each other. There are very few cities who have negative correlation with other cities.

When we Plot correlation heatmap using seaborn library we can visualize it easily



From the above chart it is clearly visible that few cities which are negatively correlated with others and shown in black strips.

If someone wants to see if the other city is similar or different then they can use this heat map or matrix above to determine correlation.

It can also be used for comparable analysis between two or more cities.

5. Recommendation and Conclusion

In this project I analysed that most of the cities are similar in context to the popular category of venues they have. I also found out there are 13 cities which are different from the rest and if someone wants to expand their business then they can look in which group the city belongs to and they can make the decision based on this information.

Now we can answer the question asked in the introduction section that it is possible to determine whether different cities have something in common. If someone wants to move/ expand his business to another city they can look at correlation heat map and decide based on that.