**Cities of India**

Rahul thaker

4th March 2020

**1. Introduction**

1.1 Background

India is a country located in Asia. It is the 2nd most populated country in the world. India has 29 states and 7 union territories. Having a lot of population creates densely populated cities. Major cities such as Mumbai, Delhi, Hyderabad have a lot of people working in different sectors. A person having a job or a business may have to move to different city for other opportunities. It would be helpful to know which city would be comfortable.

1.2 Problem

As there is a higher probability for businesses to grow and open a branch in another major city or for employers to have more workforce In another location. Employees not being able to habituate themselves in a certain location or the growth of a business could come to a stop because of a location which may not be suitable. Knowing similarities between locations and analysis of major cities becomes important at this stage

**2.Data acquisition and cleaning**

2.1 Data source

The latitudes and longitudes for the major Indian cities

Could be found on <https://simplemaps.com/data/in-cities> which could be downloaded in an excel file and then be processed using pandas library.

The information regarding individual cities could be then be found using the FOURSQAUE API <https://developer.foursquare.com/docs/api> by providing the latitudes and longitudes.

2.2 Data cleaning

Much Data cleaning is not required in this case. The data should be read from the simplemaps html website which provides a clean data set of latitudes and longitudes of major Indian cities. Depending on the number of Cities that are to be taken into consideration slicing of the data source obtained must be done.

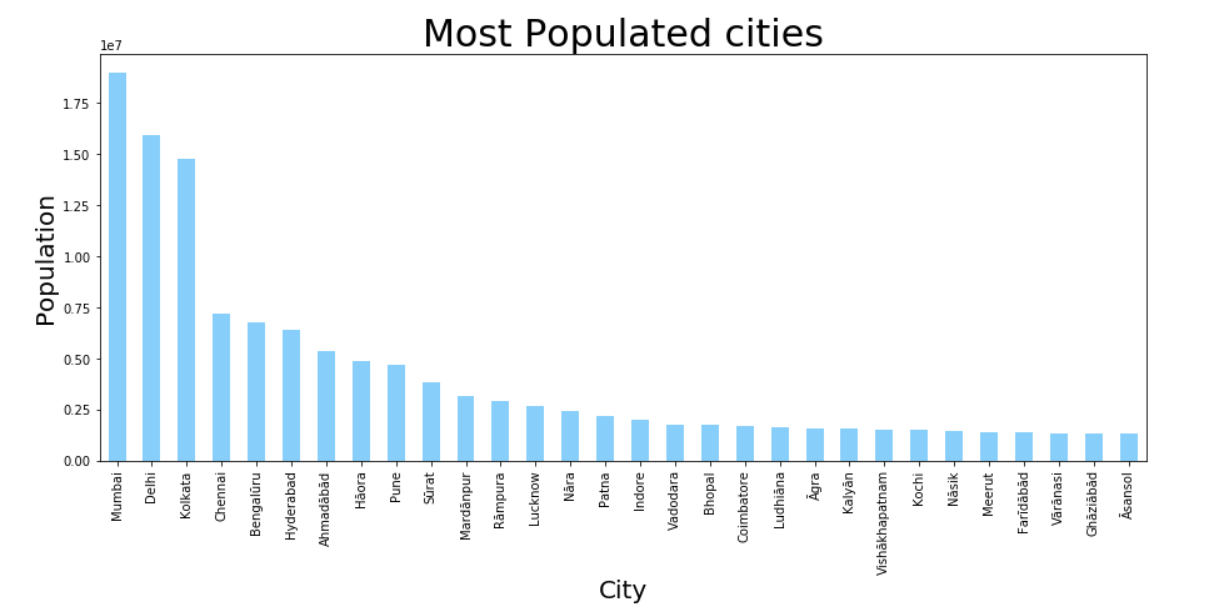
The data set comes with 8 attributes for each city out of which we require only 3 attributes namely latitude ,longitude and the name of the city. I have taken the top 30 most populated cities in India.



**3.Data Analysis**

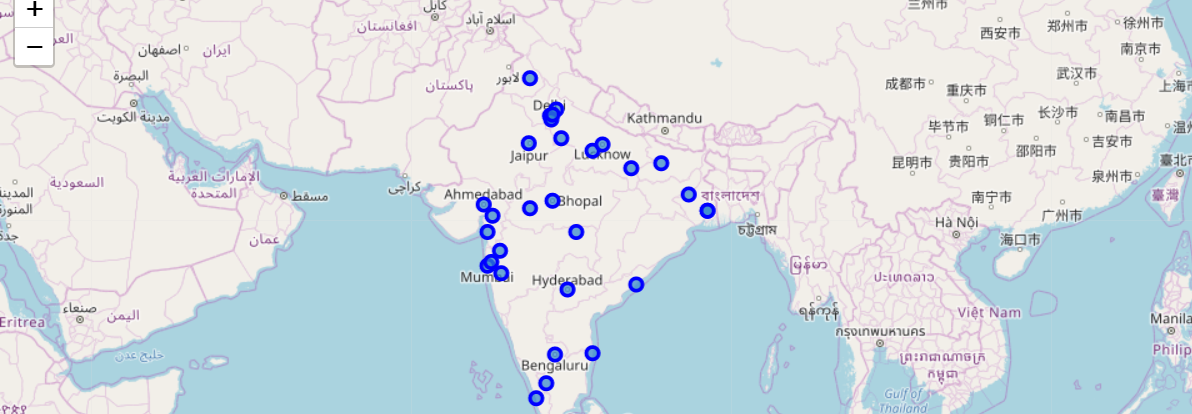
**3.1 Initial visuals**

To get an idea about the population of different cities an initial bar graph can be visualized as follows



As the final similarity between cities is going to be represented using markers on the map of India an initial look of the cities on the map of India is helpful.

Using the folium library we can generate interactive maps for different locations based on the provided latitude and longitude. India’s latitude and longitude are 20.5937° N, 78.9629° E. The cities could similarly be represented. This is the visual

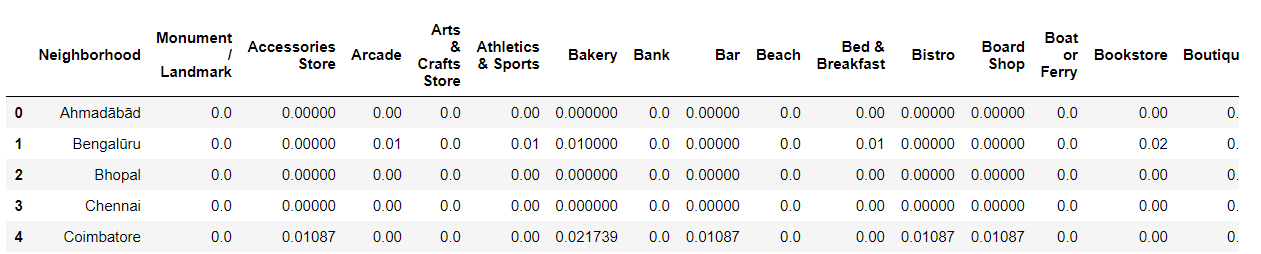


**3.2 Data processing:**

Foursqaure api provides different types of data based on the api request made and the parameters given. In our case we find all the different venues and places available in the city for example: theatres, parks, gyms, different types of restaurants etc.

So using this the different venues for all the cities are found. By repeatedly making api calls for their respective latitude and longitudes therby creating a dataframe with all the venues for a given radius for all the cities.

Since for applying a clustering model we have to use numeric data. Hence using one hot encoding the different categories are converted into different columns.

Finally after cleaning and processing a dataframe with points is created I.e each city is given points based on the different types of venues it has.

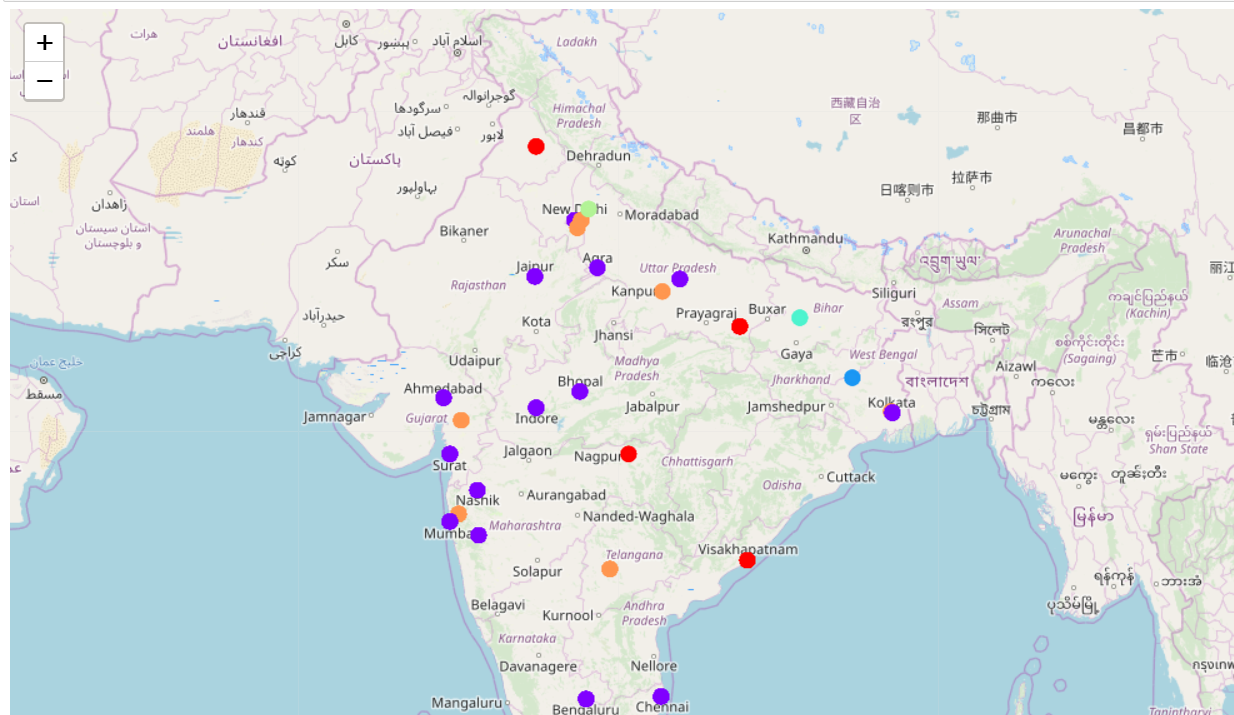
Now we have all the cities with their different attributes.

Hence we apply a KMeans clustering model. Which gives makes clusters of different data points In this case different cities based on their similarity since the attributes are different venues.

Finally each city is labelled and it belongs to a cluster based on its similarity with other cities.

**4.Result and conclusion.**

The clustering done by the Kmeans model could be represented using a folium map as done initially. The visual looks somewhat like this.



The cities based on their attributes are divided into groups represented by a specific color which is what we required.

Further based on the data derived while creating attributes analysis could be done for different attributes

For example: one of the attributes I have taken is western food available in each city which included a combination of different restaurants and food types from donuts to burgers to pizzas. This attribute could be visualized for each city. I thought it could be beautifully represented by using a pie chart as follows.

