**Introduction**

Comparing neighborhood of cities in two states of India namely Maharashtra and Karnataka to determine a locality to set up a business. This can later be extended for Pan India.

Target audience would be anyone who is interested to do a business in these two states (or in India at a later stage)

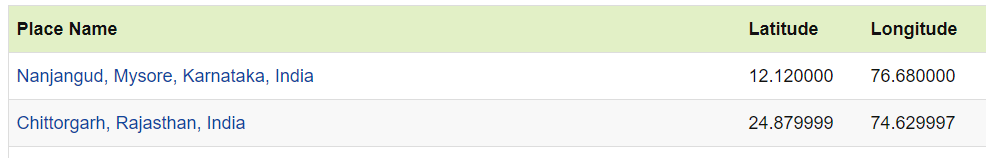
This would help them identify specific locations based on the locality of the targeted cities.

**Dataset**

Data of Indian cities is scraped from the following URL:

<https://www.latlong.net/category/cities-102-15.html>

This contains 794 cities spread out across 8 HTML pages in the following format:



The field ‘Place Name’ represents the Name of cities, ‘Latitude’ field represents the Latitude and the ‘Longitude’ field represents the Longitude of the city.

This table is stored in a dataframe df. From df, cities in Maharashtra and Karnataka are filtered out (which is the requirement for our project) and stored in another dataframe df\_mah.

Along with the above described table, locality based data is derived from Foursquare using the developer version.

**Methodology**

Once we obtain the dataframe df\_mah, the data points are plotted on map using Folium library.

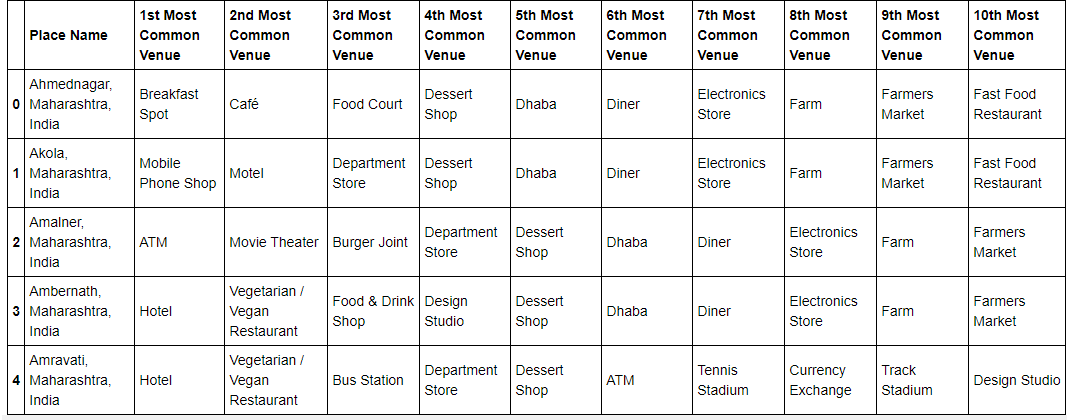
Next we provide Client Id, Client Secret of Foursquare to get nearby 100 (LIMIT =100) venues to these cities in df\_mah within a radius of 500m (radius=500). This information is converted into another dataframe India\_venues:



Then we use One hot encoding to convert these Categorical features of ‘Venue Category’ to columns of 1s and 0s and store in dataframe name mah\_onehot.

Since, there will be repetitions in the column Place Name, next we use groupby and club these using mean.

Next, top 10 most common venues are sorted in dataframe mah\_venues\_sorted



Now, it’s the time to perform clustering. We use K Means and divide the cities into 5 clusters. Place Name having NaN as cluster labels are dropped down.

Finally, these clusters are plotted on map using 5 different colors.

**Results**

We obtain 5 clusters for the cities in Karnataka and Maharashtra. These clusters can be seen in the output as Cluster 1, Cluster 2, Cluster 3, Cluster 4 and Cluster 5.

**Discussion**

The clusters are formed using an unsupervised Machine Learning algorithm – K Means. Here, only 2 States are considered due to limit of API calls in free developer account of Foursquare. Later, for commercial purposes, this project can be extended for Pan India considering all the states and the dataframe df consisting of 794 rows.

**Conclusion**

The cities in the column ‘Place name’ were successfully clustered. Hence, the business owner can now easily identify the locations of his interest and also find similar locations in the 2 states of Maharashtra and Karnataka.