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)

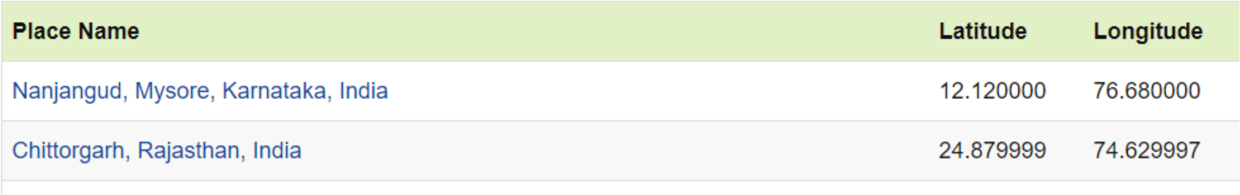
This would help them identify specific locations based on the locality of the targeted cities.

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:



Place Name	Latitude	Longitude
Nanjangud, Mysore, Karnataka, India	12.120000	76.680000
Chittorgarh, Rajasthan, India	24.879999	74.629997

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

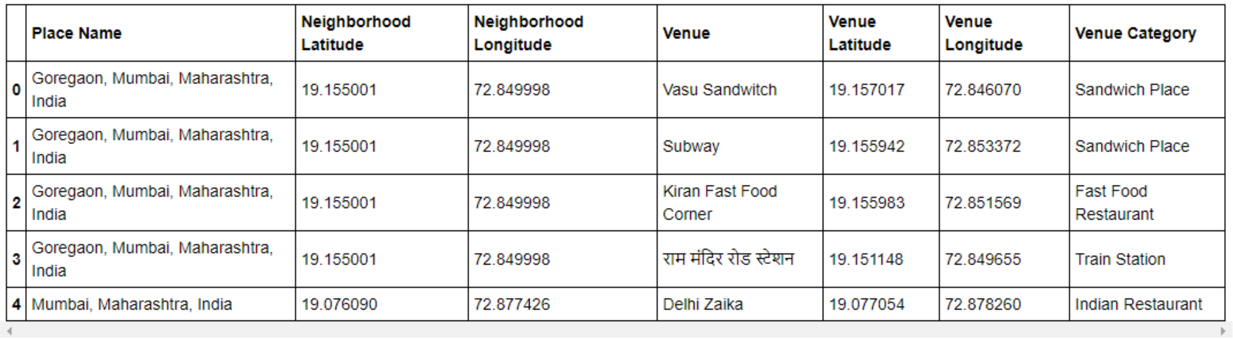
This table is stored in a dataframe df. From df, cities in Maharashtra and Karnataka are filtered out (which is the requirement for our project) and stored in another dataframe df_mah.

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:

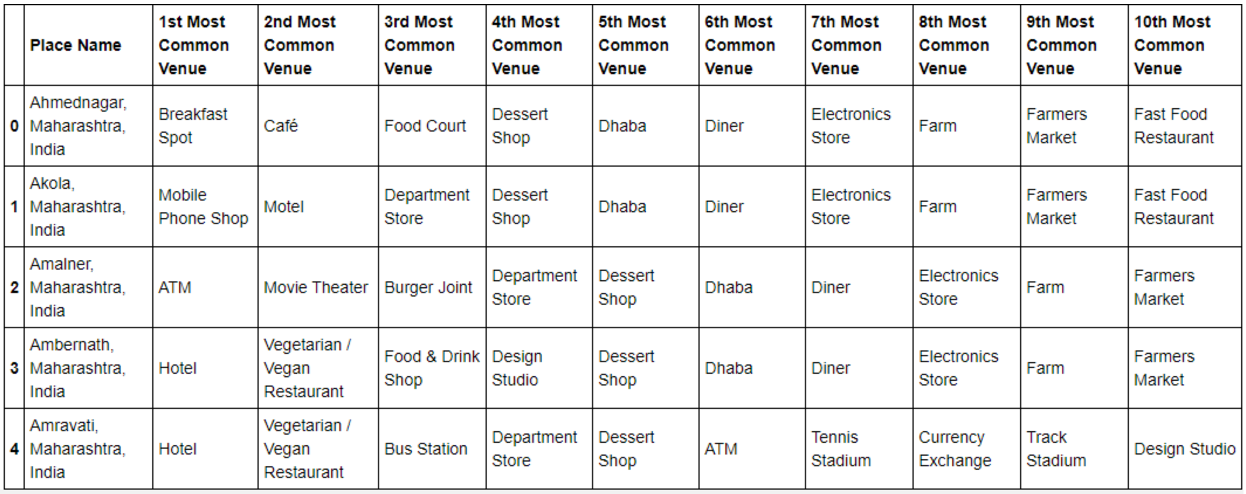


	Place Name	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Goregaon, Mumbai, Maharashtra, India	19.155001	72.849998	Vasu Sandwitch	19.157017	72.846070	Sandwich Place
1	Goregaon, Mumbai, Maharashtra, India	19.155001	72.849998	Subway	19.155942	72.853372	Sandwich Place
2	Goregaon, Mumbai, Maharashtra, India	19.155001	72.849998	Kiran Fast Food Corner	19.155983	72.851569	Fast Food Restaurant
3	Goregaon, Mumbai, Maharashtra, India	19.155001	72.849998	राम मंदिर रोड स्टेशन	19.151148	72.849655	Train Station
4	Mumbai, Maharashtra, India	19.076090	72.877426	Delhi Zaika	19.077054	72.878260	Indian Restaurant

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



	Place Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Ahmednagar, Maharashtra, India	Breakfast Spot	Café	Food Court	Dessert Shop	Dhaba	Diner	Electronics Store	Farm	Farmers Market	Fast Food Restaurant
1	Akola, Maharashtra, India	Mobile Phone Shop	Motel	Department Store	Dessert Shop	Dhaba	Diner	Electronics Store	Farm	Farmers Market	Fast Food Restaurant
2	Amalner, Maharashtra, India	ATM	Movie Theater	Burger Joint	Department Store	Dessert Shop	Dhaba	Diner	Electronics Store	Farm	Farmers Market
3	Ambernath, Maharashtra, India	Hotel	Vegetarian / Vegan Restaurant	Food & Drink Shop	Design Studio	Dessert Shop	Dhaba	Diner	Electronics Store	Farm	Farmers Market
4	Amravati, Maharashtra, India	Hotel	Vegetarian / Vegan Restaurant	Bus Station	Department Store	Dessert Shop	ATM	Tennis Stadium	Currency Exchange	Track Stadium	Design Studio

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.