**Introduction**

Mumbai and Pune are two major cities in India. Both cities become a center of attention for residential, job employment, tourism, education, shopping and sports activity. Both cities are well known in India and become the top choice for local and foreign communities.

Brief information about both cities:

* **Mumbai:** (/mʊmˈbaɪ/; also known as Bombay) is the capital city of the Indian state of Maharashtra. It is the most populous city in India with an estimated city proper population of 12.4 million as of 2011. Along with the neighbouring regions of the Mumbai Metropolitan Region, it is the second most populous metropolitan area in India, with a population of 21.3 million as of 2016. Mumbai lies on the Konkan coast on the west coast of India and has a deep natural harbour. In 2008, Mumbai was named an alpha world city. It is also the wealthiest city in India and has the highest number of millionaires and billionaires among all cities in India.

(Source: <https://en.wikipedia.org/wiki/Mumbai>)

* **Pune:** (Marathi pronunciation: [puɳe]; English: /ˈpuːnə/;), formerly spelled Poona (1857–1978), is the second largest city in the Indian state of Maharashtra, after Mumbai. It is the ninth most populous city in the country with an estimated population of 3.13 million. Along with its Industrial twin Pimpri Chinchwad and the three cantonment towns of Pune, Khadki and Dehu Road, Pune forms the urban core of the eponymous Pune Metropolitan Region (PMR). According to the 2011 census, the urban area has a combined population of 5.05 million while the population of the metropolitan region is estimated at 7.27 million. Situated 560 metres (1,837 feet) above sea level on the Deccan plateau on the right bank of the Mutha river, Pune is also the administrative headquarters of its namesake district. In the 18th century, the city was the seat of the Peshwas, the prime ministers of the Maratha Empire and so was one of the most important political centers on the Indian subcontinent. Pune is ranked the No. 1 city in India in 'ease of living' ranking index. Pune is also a most important center for civil services training.

(Source: <https://en.wikipedia.org/wiki/Pune>)

**Objective**

In this project, we will study in detail the area classification using Foursquare data and machine learning segmentation and clustering. The aim of this project is to segment areas of Mumbai and Pune based on the most common places captured from Foursquare.

Using segmentation and clustering, we hope we can determine:

* The similarity or dissimilarity of both cities
* Classification of area located inside the city whether it is residential, tourism places, or others

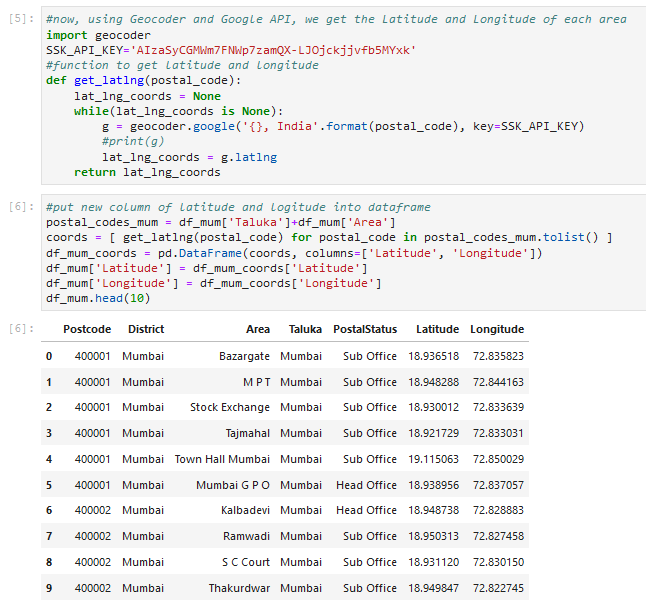
**Data**

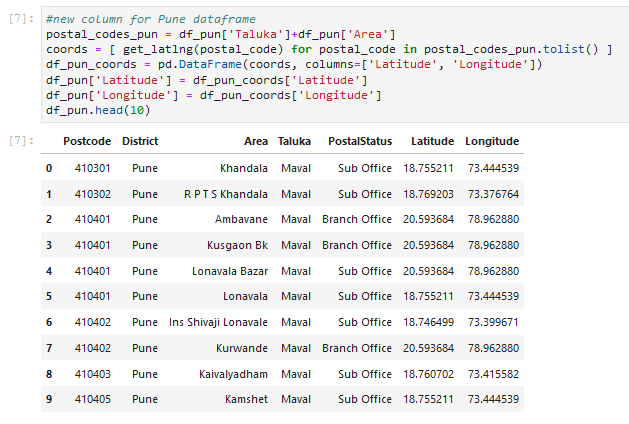
The data acquired from Wikipedia pages and restructure to csv file for easier manipulation and reading. Both files uploaded to my GitHub for references. Link to the files are:

* [Mumbai\_district.csv](https://github.com/skakkara-in-ibm-com/Coursera_Capstone_Assignment1/blob/master/Mumbai_district.csv)
* [Pune\_district.csv](https://github.com/skakkara-in-ibm-com/Coursera_Capstone_Assignment1/blob/master/Pune_district.csv)

Another aspect to consider for this project is the Foursquare data. I believe that the data as good as provided, meaning although we are using Foursquare data for segmentation and clustering, the amount and accuracy of data captured can't 100% determine correct classification in real world.

To start, let's get and look at the data. I've already downloaded it, so let's read it (from local drive) and load it to dataframe: Using geocoder, we able to get Latitude and longitude for each area.





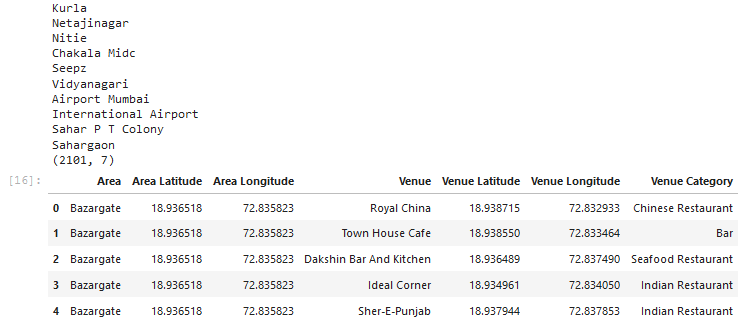
**Methodology**

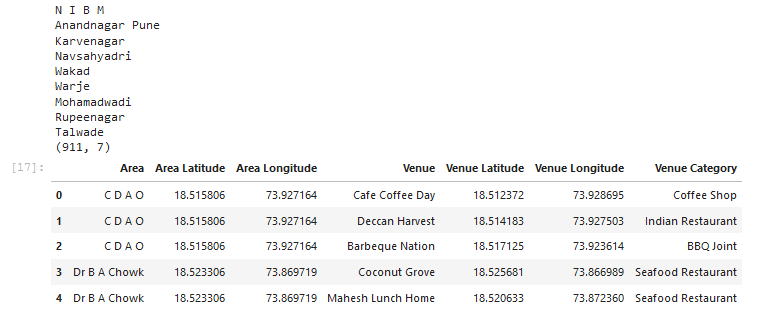
In this project, I will use the basic methodology as taught in Week 3 lab.

* Above, we have done convert addresses into their equivalent latitude and longitude values.
* Then we will use the Foursquare API to explore neighbourhoods in both cities, Mumbai and Pune
* After that, explore function to get the most common venue categories in each neighbourhood,
* and then use this feature to group the neighbourhoods into clusters

K-means clustering algorithm will be using to complete this task. And the Folium library to visualize the neighbourhoods in Mumbai and Pune and their emerging clusters.

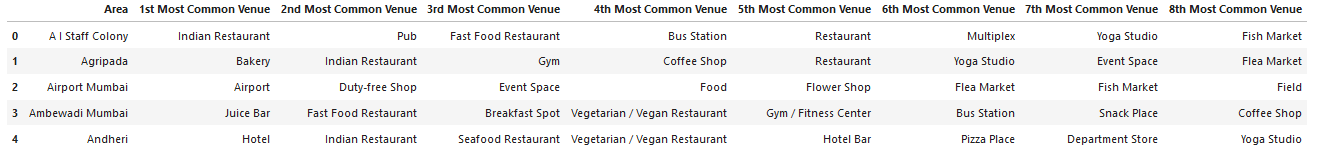
Based on dataframe analysis above, we found out that **Mumbai Taluka** in **Mumbai** and **Pune City Taluka** in **Pune** are both have the highest number of area within it those districts.





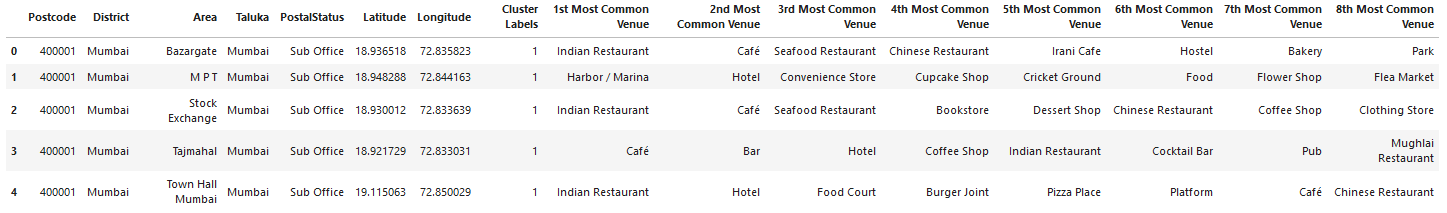
**Analyse Mumbai**

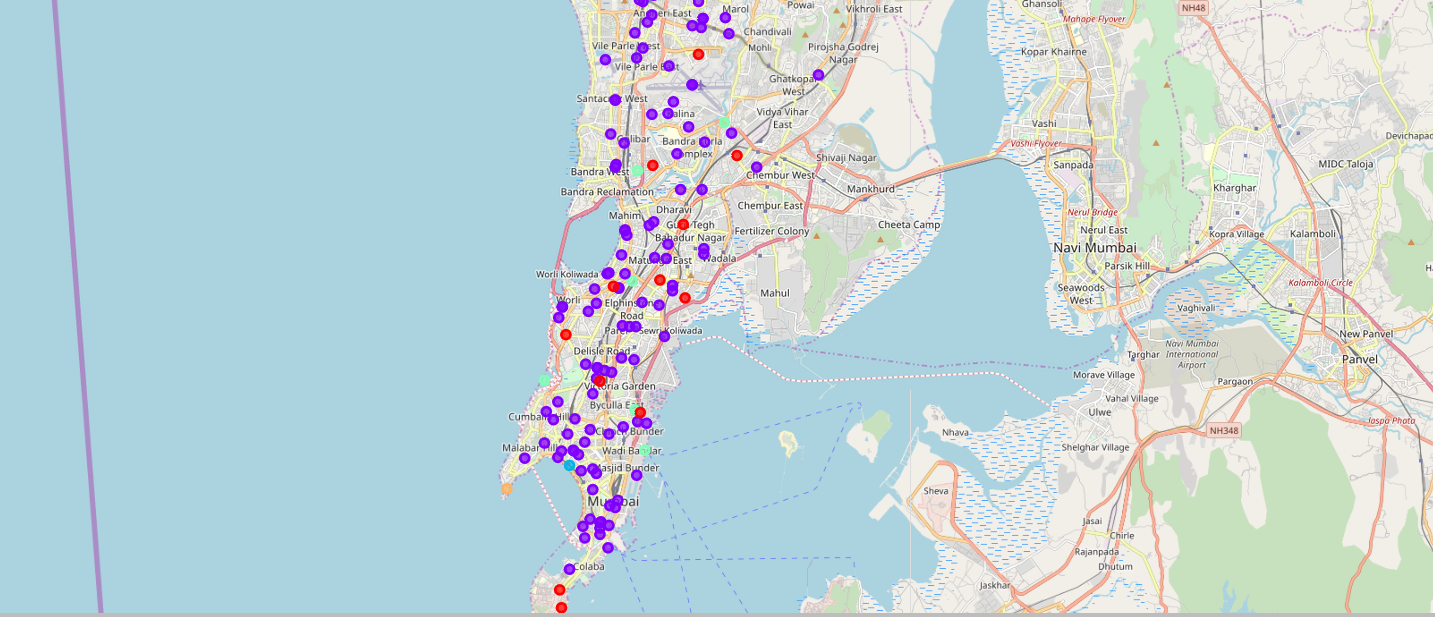
Analysing Mumbai data to get most common venue for each area.



**K-mean Cluster Mumbai**

Using K-mean to clustering data area with most common venue





**Analyse Pune**

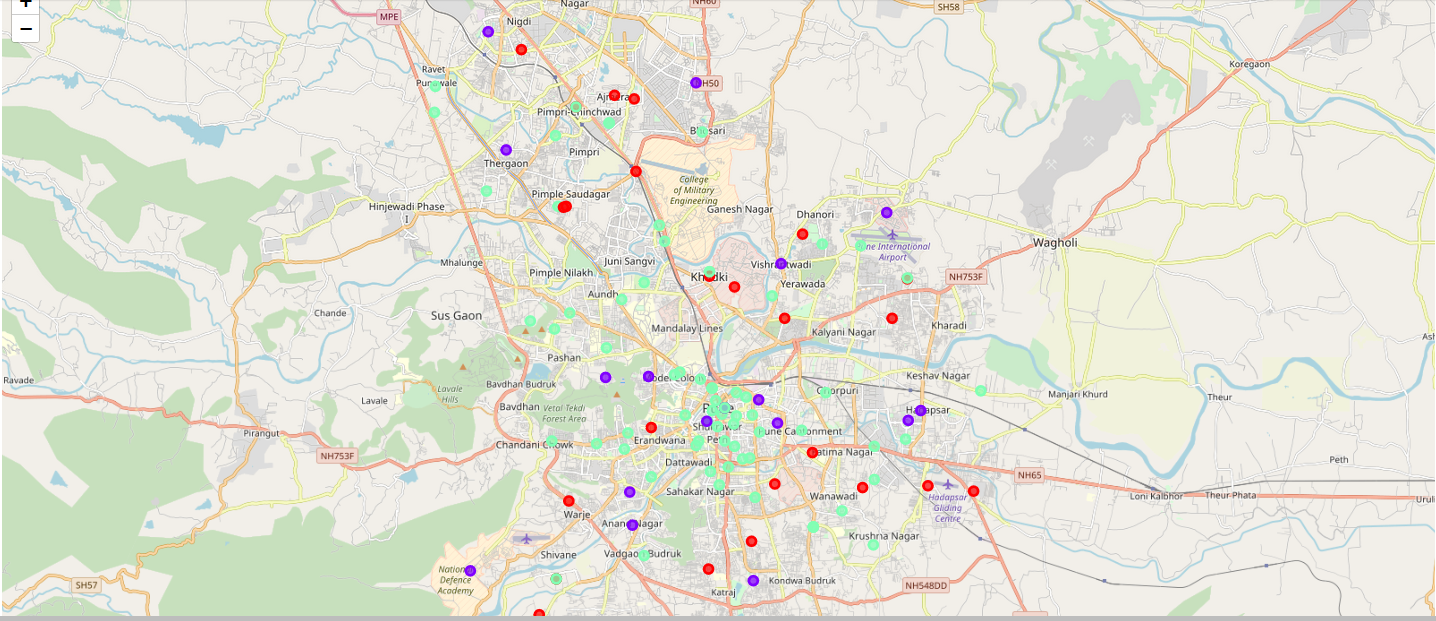
Analysing Pune data to get most common venue for each area.



**K-mean Cluster Pune**

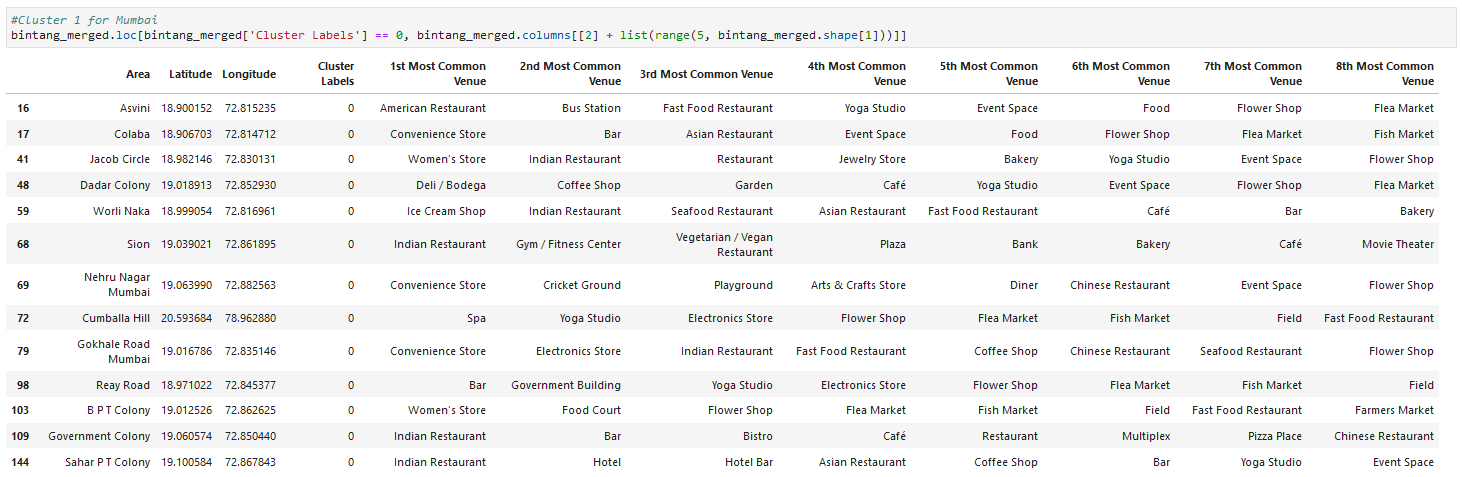
Using K-mean to clustering data area with most common venue



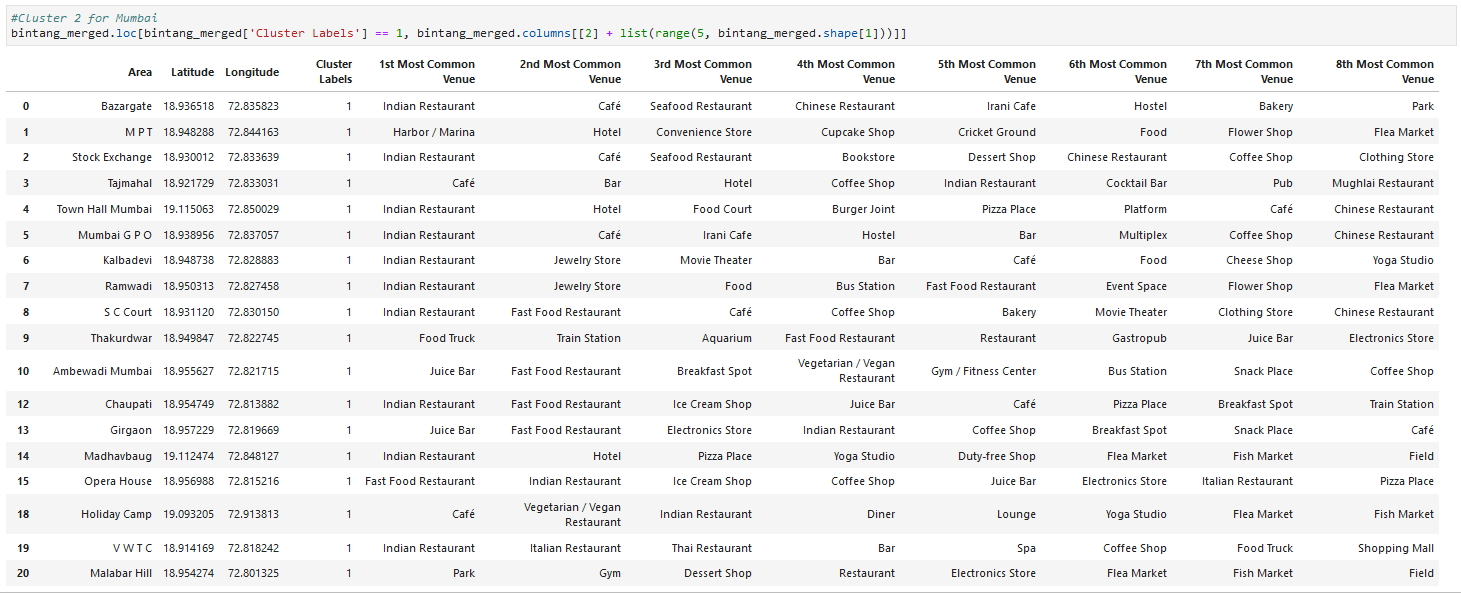


**Result**

Cluster 1 for Mumbai



Cluster 2 for Mumbai

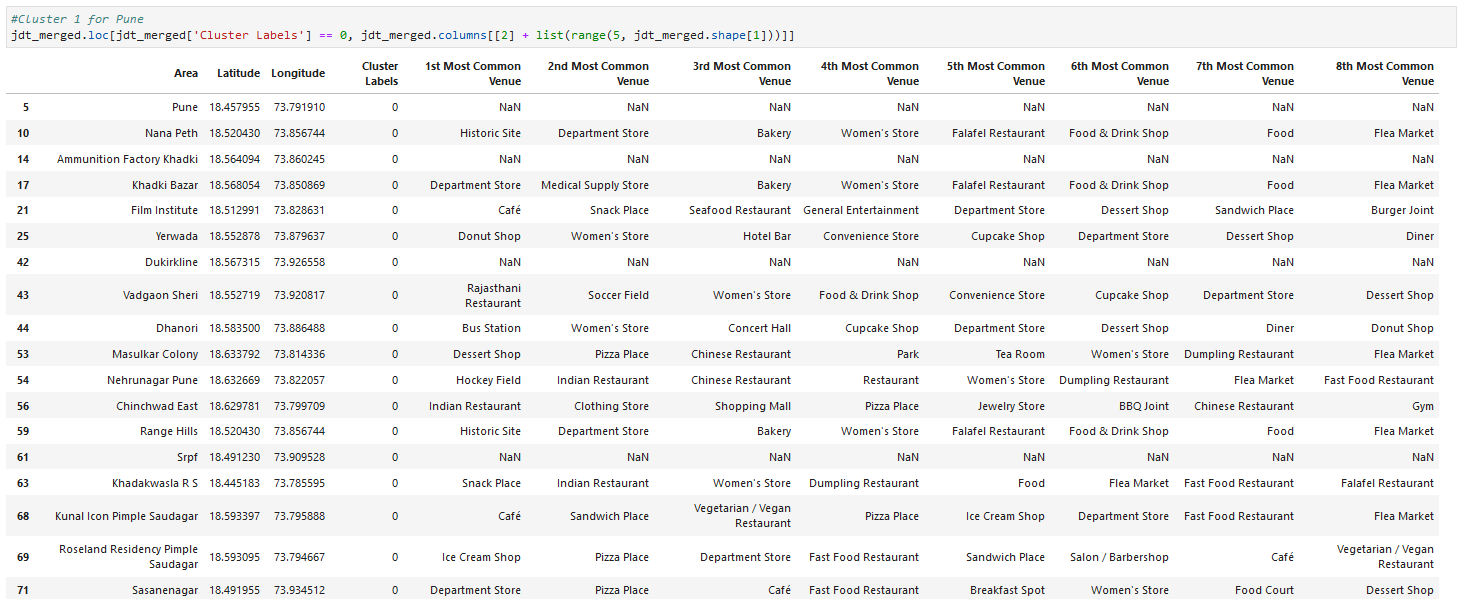




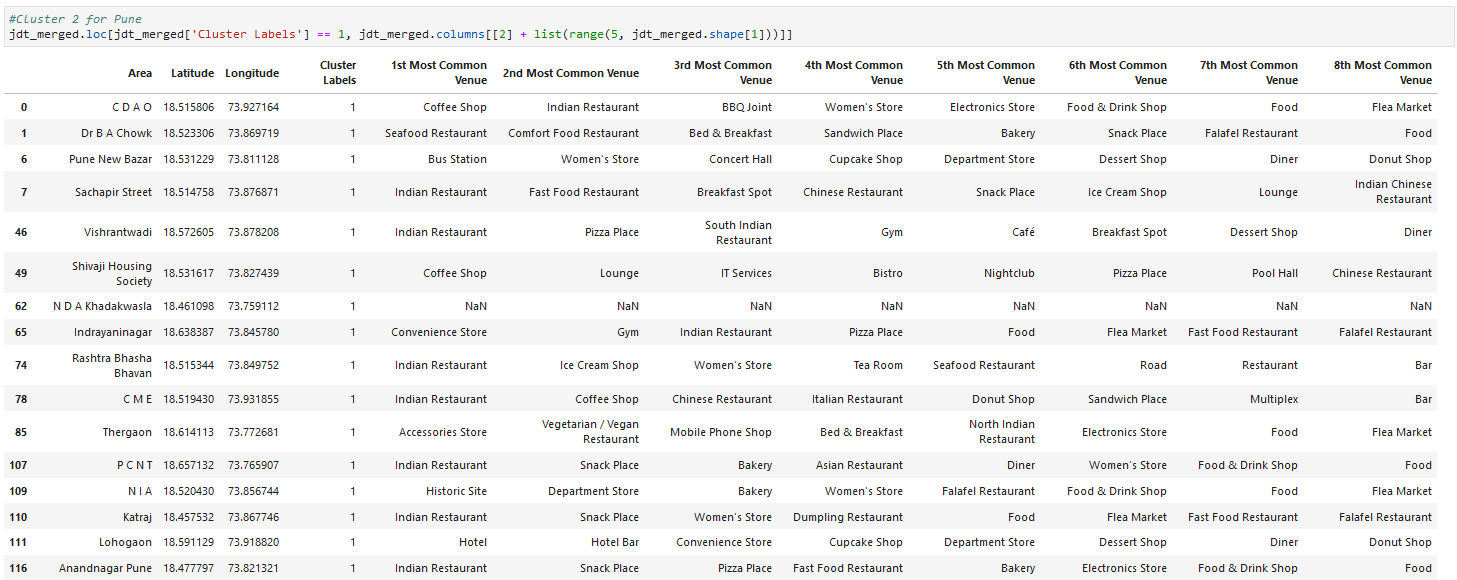
Cluster 3 for Mumbai



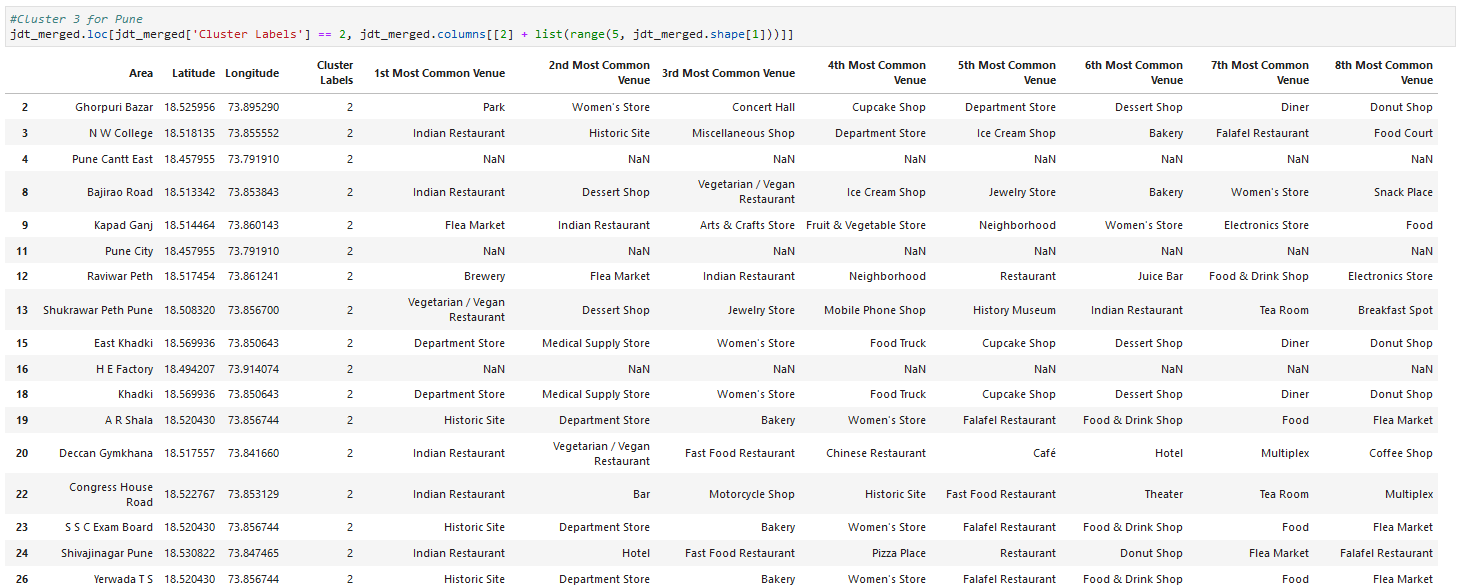
Cluster 1 for Pune



Cluster 2 for Pune



Cluster 3 for Pune



**Discussion**

Based on cluster for each city above, we believe that classification for each cluster can be done better with calculation of venues categories (most common) in each cities. Referring to each cluster, we can't determine clearly what represent in each cluster by using Foursquare - Most Common Venue data.

However, for the sake of this project we assumed each cluster as follow:

Cluster 1: Mumbai: Shopping

Cluster 2: Mumbai: Mix

Cluster 3: Mumbai: Food

Cluster 1: Pune: Tourism

Cluster 2: Pune: Residential

Cluster 3: Pune: Mix

What is lacking at this point is a systematic, quantitative way to identify and distinguish different district and to describe the correlation most common venues as recorded in Foursquare. The reality is however more complex: similar cities might have or might not have similar common venues. A further step in this classification would be to find a method to extract these common venues and integrate the spatial correlations between different of areas or district.

We believe that the classification we propose is an encouraging step towards a quantitative and systematic comparison of the different cities. Further studies are indeed needed in order to relate the data acquired, then observe it to more meaningful and objective results.

**Conclusion**

Using Foursquare API, we can captured data of common places all around the world. Using it, we refer back to our main objectives, which is to determine;

the similarity or dissimilarity of both cities classification of area located inside the city whether it is residential, tourism places, or others In conclusion, both cities Mumbai and Pune are the center of attraction among Indian. However, to declare both cities are similar or dissimilar base on common venues visited is quite difficult. Both cities is similar in some venues also dissimilar in certain venues. And for classification based on common venues, again we must have more systematic or quantitative way to identify and declare this. Comparison can be made, but no such method or quantitative data to determine this. We hope in the future, a method to determine it can be establish and explore for references.