The Battle of South Indian States - Tamil Nadu and Kerala

Applied Data Science Capstone by IBM/Coursera

In this assignment, we are going to explore, segment, and cluster the two southernmost states in India, namely, Tamil Nadu and Kerala.

Introduction

In this project, we will look at two southern states in India, namely Tamil Nadu and Kerala. There are some of the large districts in these states for residential, tourism, culture, etc. Southern India is spread over an area of 635,780 km^2 and is filled with a lot of culture and traditions. Let's see more information on the two states that we would be focusing on.

Tamil Nadu: formerly Madras State, is one of the 29 states of India. Its capital is Chennai (formerly known as Madras). Tamil Nadu lies in the southernmost part of the Indian subcontinent and is bordered by the union territory of Puducherry and the South Indian states of Kerala, Karnataka, and Andhra Pradesh. It is bounded by the Eastern Ghats on the north, by the Nilgiri Mountains, the Meghamalai Hills, and Kerala on the west, by the Bay of Bengal in the east, by the Gulf of Mannar and the Palk Strait on the southeast, and by the Indian Ocean on the south. The state shares a maritime border with the nation of Sri Lanka. (https://en.wikipedia.org/wiki/Tamil_Nadu)

Kerala: is a state on the southwestern Malabar Coast of India. It was formed on 1 November 1956, following the passage of the States Reorganisation Act, by combining Malayalam-speaking regions. Spread over 38,202 km^2, Kerala is the twenty-second largest Indian state by area. It is bordered by Karnataka to the north and northeast, Tamil Nadu to the east and south, and the Lakshadweep Sea to the west. Malayalam is the most widely spoken language and is also the official language of the state. (https://en.wikipedia.org/wiki/Kerala)

Business Problem

In this project, we will analyze the two southernmost states of India. We will use Foursquare data as well as machine learning algorithms such as segmentation and clustering to help us complete this study. The reason for the undertaking is to group the most well-known venues in both these Indian states. This project can help us identify the similarities as well as the dissimilarities for the two states and to classify the various districts in these states.

Data

The data needed for this project has been obtained from Wikipedia as well as Census 2011. From Wikipedia, we can get the Districts, Population and Area details of both the states. From Census 2011, we can obtain the data on the literacy percentage of all the districts. We have also used the geocoder to derive the latitude and longitude of the districts in Tamil Nadu and Kerala. All these acquired data are processed, so as to get clean data to make the analysis process easier and then is stored in a single data frame. The webpages from where the required data has been collected are:

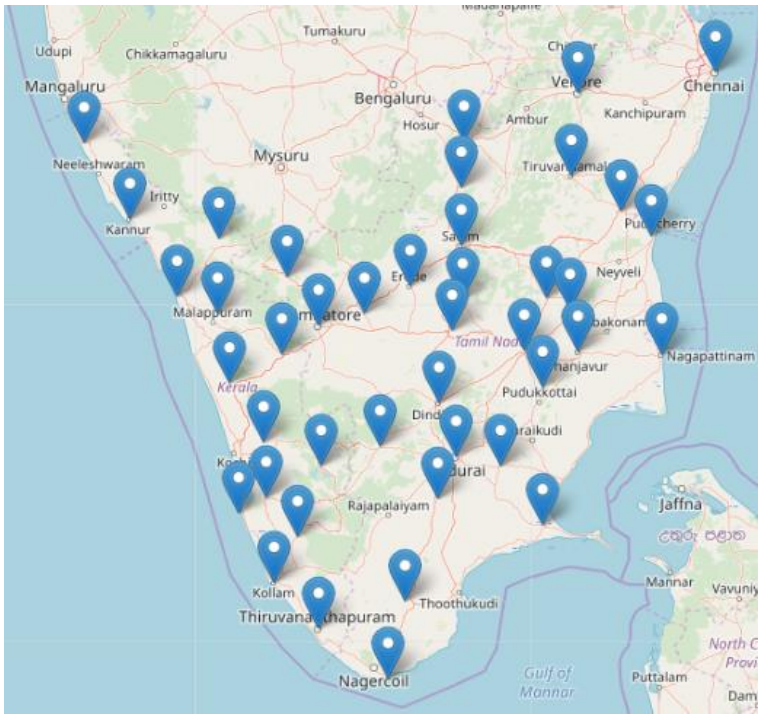
Tamil Nadu Districts (https://simple.wikipedia.org/wiki/List_of_districts_in_Tamil_Nadu)
Tamil Nadu Districts (https://www.census2011.co.in/census/state/districtlist/tamil+nadu.html)
Kerala Districts (https://en.wikipedia.org/wiki/List_of_districts_in_Kerala)
Kerala Districts (https://www.census2011.co.in/census/state/districtlist/kerala.html)

In this analysis, we would take some help from the Foursquare API as it is major data gathering source as it has a database of more than 105 million places. We have also used information from Foursquare to derive the venues in all the districts. It can also help us to find the most common locations in these districts.

	State	District	Population	Area_sq_km	Literacy	Latitude	Longitude
0	Tamil Nadu	Ariyalur	754894	1949	71.34	11.135771	79.072320
1	Tamil Nadu	Chennai	4646732	426	90.18	13.080172	80.283833
2	Tamil Nadu	Coimbatore	3458045	4723	83.98	11.001812	76.962842
3	Tamil Nadu	Cuddalore	2605914	3678	78.04	11.742694	79.750306
4	Tamil Nadu	Dharmapuri	1506843	4497	68.54	12.134799	78.158986

Methodology

In this project, we will use various methods to do the analysis of the metrics available to us. We have transformed and kept only the required metrics that could help us in this project. The unwanted metrics are removed using python codes. Using the latitude and longitude data, we can explore all the districts in Tamil Nadu and Kerala with the help of a folium map. Folium library in python helps to visualize data on an interactive leaflet map.



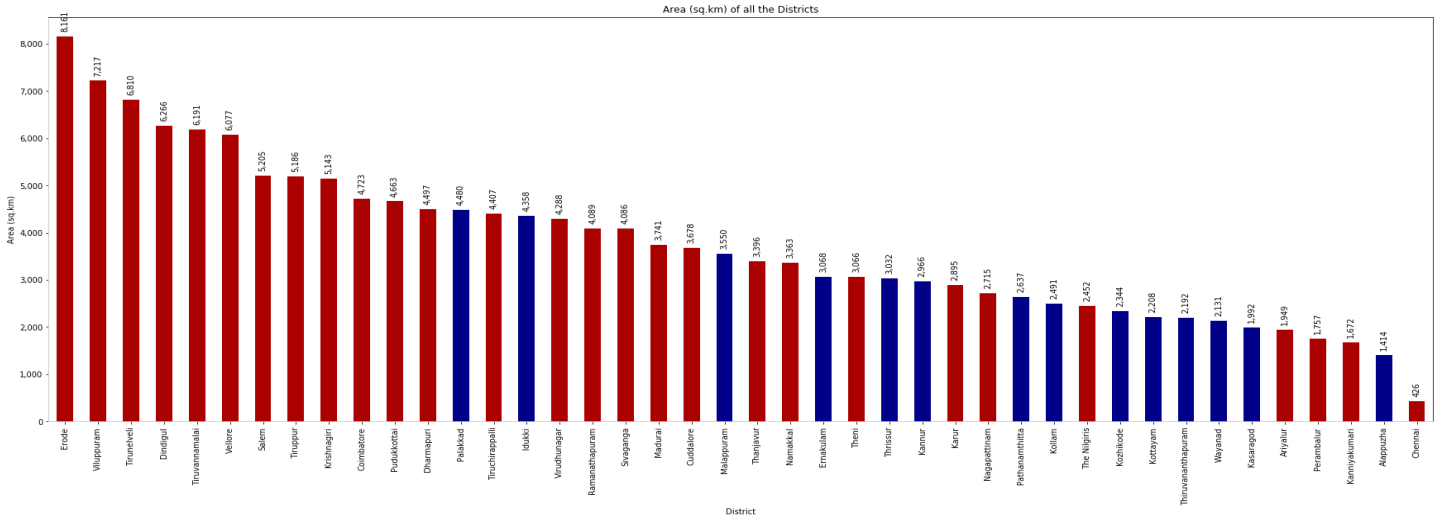
We can see that there are 42 districts in total, where Tamil Nadu has 28 districts and Kerala with 14 districts.

These are the districts in Tamil Nadu: Ariyalur, Chennai, Coimbatore, Cuddalore, Dharmapuri, Dindigul, Erode, Kanyakumari, Karur, Krishnagiri, Madurai, Nagapattinam, Namakkal, Perambalur, Pudukottai, Ramanathapuram, Salem, Sivaganga, Thanjavur, The Nilgiris, Theni, Thiruvannamalai, Thirunelveli, Tiruppur, Trichirappalli, Vellore, Villupuram and Virudhunagar.

The following are the districts in Kerala: Alappuzha, Ernakulam, Idukki, Kannur, Kasaragod, Kollam, Kottayam, Kozhikode, Malappuram, Palakkad, Pathanamthitta, Thiruvananthapuram, Thrissur and Wayanad.

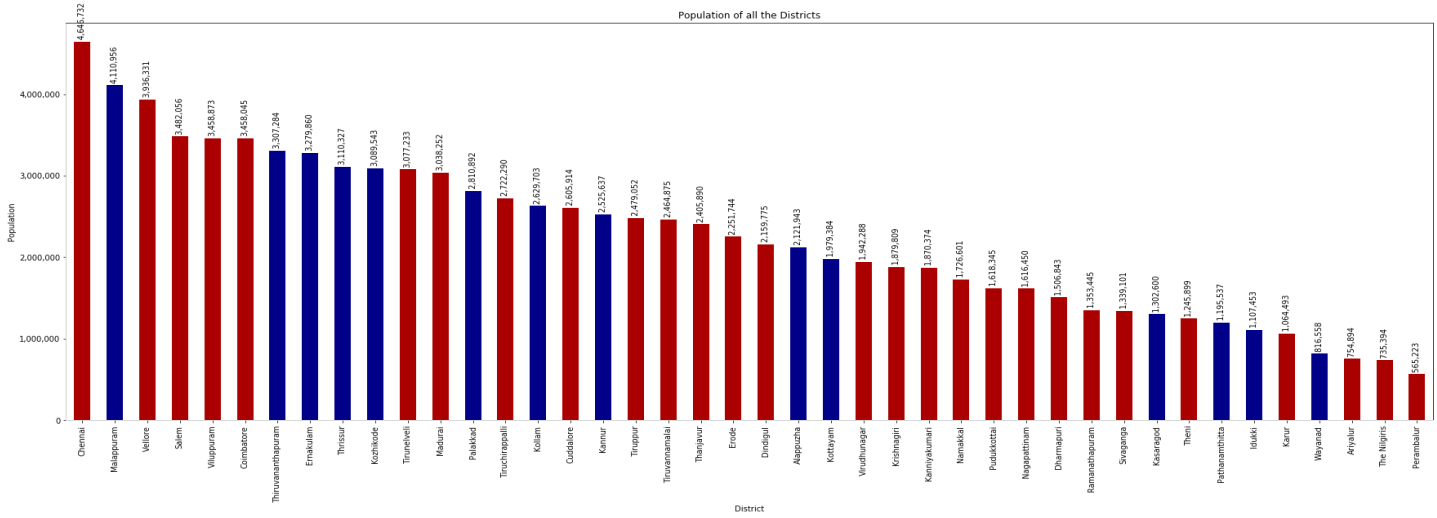
These are the districts that we would be analyzed to determine the similarities and dissimilarities between the two states.

Analysis on the Area: Tamil Nadu is larger than Kerala based on area, as Tamil Nadu spreads over 118,119 km^2 and Kerala spreads over 38,202 km^2. To further see into the area of each district, I have used the data available to us to be plotted into a bar graph to better visualization.



From the above bar graph, we can see that Erode from Tamil Nadu is the largest with an area of 8,161 km^2 and Chennai from Tamil Nadu with an area 426 km^2 is the smallest among all the districts. The largest district in Kerala is Palakkad and smallest district is Alappuzha.

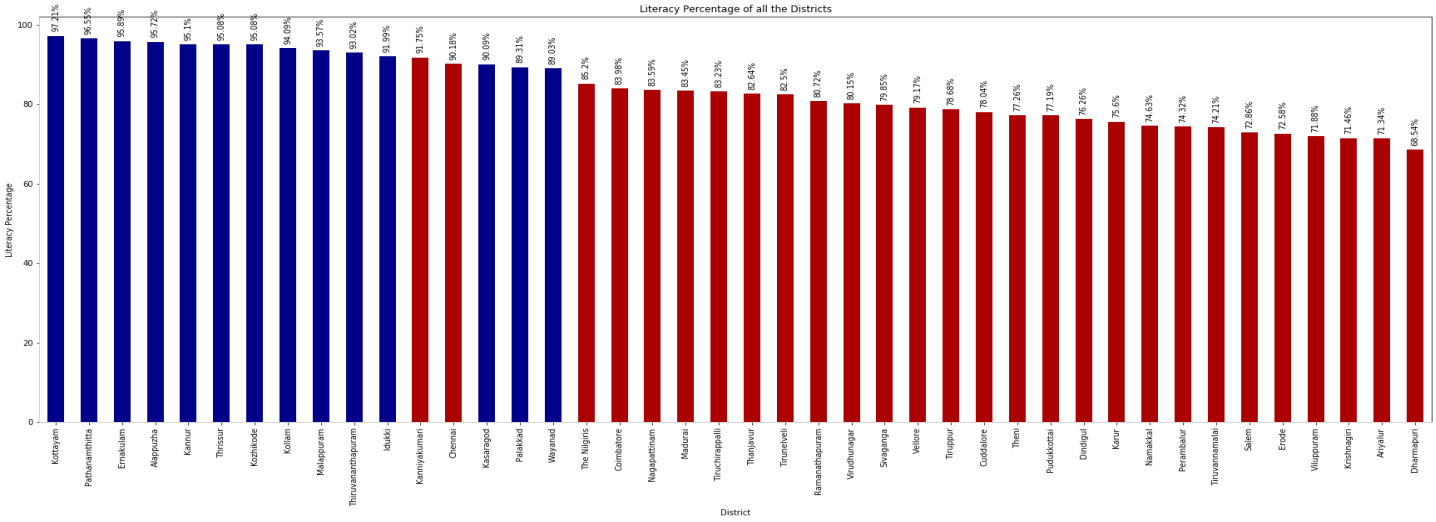
Analysis of the Population: Tamil Nadu is more populated than Kerala. Tamil Nadu has 61.41 million population and Kerala has 33.39 million population. To further see into the population of each district, I have used the data available to us to be plotted into a bar graph to better visualization.



From the above bar chart, we can see that Chennai from Tamil Nadu is the most populated with a population of 4.65 million and the least populated district is Parambalur from Chennai with 565,223 population. Wayanad is the least populated district in Kerala and Malappuram with the most population in Kerala.

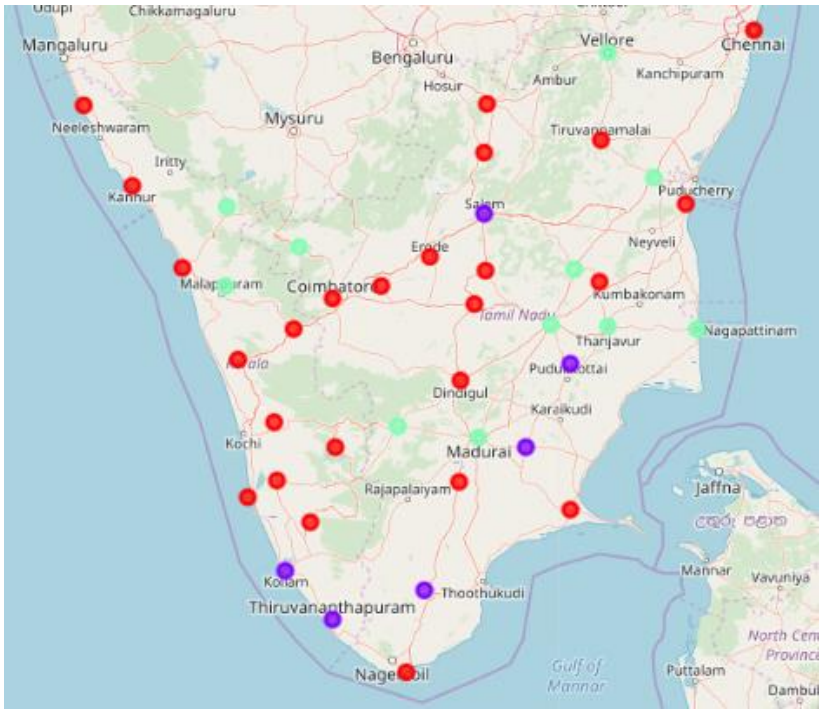
We should also note that Chennai is the capital and it has the least area but has the highest population. So, we can say clearly that Chennai is overpopulated.

Analysis of the Literacy Percentage: From the data, we can say that people in Kerala are more literate than Tamil Nadu. Let's look at the literacy percentage of all the districts in both the states by plotting the data on a bar graph.



The above graph shows that overall Kottayam in Kerala has the highest literacy percentage of 97.21% and Dharmapuri in Tamil Nadu has the least literacy percentage of 68.54% among all the 42 districts. Kanyakumari has the highest literacy percentage in Tamil Nadu and Wayanad has the least literacy percentage in Kerala.

Clustering of the Districts: Foursquare API allows application developers to interact with Foursquare platform and data. Using Foursquare API, we can explore the venues of all the districts in, Tamil Nadu and Kerala, then most common top 10 venues in each district. Later, we use this feature to group the districts into K-means clusters to segment the districts. We have again used folium maps to visualize the clusters for better understanding.



Result

Using k-means, we have three clusters grouping all the districts. Each of the clusters is categorized based on the most common venues in them.

Cluster 1 for Tourism and Metropolitan Districts: This cluster includes metropolitan districts with tourism industries because of its ability to draw large crowds with quality attraction and entertainment. These are districts with significant culture, hotels, resorts, shopping, restaurants, and art.

Tamil Nadu: Ariyalur, Chennai, Coimbatore, Cuddalore, Dharmapuri, Dindigul, Erode. Kanyakumari, Karur, Krishnagiri, Namakkal, Ramanathapuram, Tiruppur, Tiruvannamalai, and Virudhunagar

Kerala: Alappuzha, Ernakulam, Idukki, Kannur, Kasaragod, Kottayam, Kozhikode, Palakkad, Pathanamthitta, and Thrissur

	State	District	Population	Area_sq_km	Literacy	Latitude	Longitude	Cluster Labels	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 Co
0	Tamil Nadu	Ariyalur	754894	1949	71.34	11.135771	79.072320	0	Indian Restaurant	Bus Station	Historic Site	Beach	Hotel	Hotel Bar	Rest
1	Tamil Nadu	Chennai	4646732	426	90.18	13.080172	80.283833	0	Indian Restaurant	Hotel	Café	Ice Cream Shop	Beach	Multiplex	Sh
2	Tamil Nadu	Coimbatore	3458045	4723	83.98	11.001812	76.962842	0	Indian Restaurant	Multiplex	Café	Ice Cream Shop	Dessert Shop	Resort	Rest
3	Tamil Nadu	Cuddalore	2605914	3678	78.04	11.742694	79.750306	0	Indian Restaurant	Hotel	Beach	Café	Train Station	Vegetarian / Vegan Restaurant	
4	Tamil Nadu	Dharmapuri	1506843	4497	68.54	12.134799	78.158986	0	Indian Restaurant	Hotel	Café	Pizza Place	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Depa

Cluster 2 for Mix of Tourism and Residential Districts: This cluster includes the spice of both tourism and residence. It has point of interests for both the locals and the tourists.

Tamil Nadu: Pudukkottai, Salem, Sivaganga and Tirunelveli

Kerala: Kollam and Thiruvananthapuram

	State	District	Latitude	Longitude	Cluster Labels	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 M Comm Ver
15	Tamil Nadu	Pudukkottai	10.500000	78.833333	1	Indian Restaurant	Hotel	Train Station	Café	Shopping Mall	Multiplex	Asian Restaurant	Ice Cream Shop	Pi Pl
17	Tamil Nadu	Salem	11.661201	78.160250	1	Indian Restaurant	Hotel	Ice Cream Shop	Scenic Lookout	Bus Station	Café	South Indian Restaurant	Restaurant	Pi Pl
18	Tamil Nadu	Sivaganga	9.848688	78.487046	1	Indian Restaurant	Bakery	Café	Hotel	Ice Cream Shop	Shopping Mall	Cupcake Shop	History Museum	Fr Chick J
22	Tamil Nadu	Tirunelveli	8.729526	77.685235	1	Indian Restaurant	Hotel	Bakery	Beach	Restaurant	Resort	Shopping Mall	Scenic Lookout	Hist
33	Kerala	Kollam	8.887054	76.590706	1	Resort	Indian Restaurant	Beach	Hotel	Restaurant	Movie Theater	Café	Fast Food Restaurant	Cre Sh

Cluster 3 for Residential and Urban Districts: This cluster includes urban districts which mainly focuses on residences and comfort of the people. In these districts, there are zones for commuting, work opportunities, cafes, restaurants, entertainment, etc.

Tamil Nadu: Madurai, Nagapattinam, The Nilgiris, Perambalur, Thanjavur, Theni, Tiruchirappalli, Vellore, and Viluppuram

Kerala: Malappuram and Wayanad

	State	District	Latitude	Longitude	Cluster Labels	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 M Comm Ven
10	Tamil Nadu	Madurai	9.926115	78.114098	2	Indian Restaurant	Café	Bakery	Hotel	Resort	Bus Station	Restaurant	Shopping Mall	Ice Cre Sh
11	Tamil Nadu	Nagapattinam	10.764795	79.843078	2	Historic Site	Indian Restaurant	Hotel	Beach	Platform	Garden	Train Station	Pool	As Restaur
13	Tamil Nadu	The Nilgiris	11.400000	76.700000	2	Indian Restaurant	Resort	Café	Hotel	Ice Cream Shop	Mountain	Bakery	Dessert Shop	Garc
14	Tamil Nadu	Perambalur	11.235917	78.868811	2	Indian Restaurant	Hotel	Ice Cream Shop	Bakery	Vegetarian / Vegan Restaurant	Historic Site	Train Station	Pizza Place	Bus Stat
19	Tamil Nadu	Thanjavur	10.786027	79.138150	2	Indian Restaurant	Hotel	Train Station	Historic Site	Beach	Asian Restaurant	Multiplex	Ice Cream Shop	Departm St

Discussion

We can see that both the states are really good for residential as well as tourism purposes. These districts in both states are very similar to each other. For a matter of fact, we know that even their languages are quite similar. Both these states are known for their south Indian cuisines, as we can also see from the data that the most popular venue in all the districts in the Indian restaurants. We can for sure say that there is no limitation of food in all these places to enjoy.

From the data, we can say that Tamil Nadu is more populated than Kerala but the literacy rate in Kerala is way higher than Tamil Nadu. This shows that people in Kerala give more importance to education and literacy than Tamil Nadu. This is one of the differences that can be seen between these two states.

Also, some interesting information derived from this analysis is on the capital of Tamil Nadu, that is, Chennai. Chennai is the smallest district among all the districts in both the states with an area of only 426 km^2 but it is also the most populated district among all the districts with a population of 4.65M. This certainly shows that Chennai is bit too overpopulated metropolitan city.

The categories given to each cluster is not crystal-clear. But I have tried to categorize them as much as I could. Below, you can find the overall insights from the analysis.

Overall
High Area: Tamil Nadu - 118,119 sq.km

High Population: Tamil Nadu - 61.4M
High Literacy Percentage: Tamil Nadu
Most common venue: Indian Restaurants

Tamil Nadu
Top District based on Population: Chennai – 4.65M
Top District based on Area: Erode – 8,161 sq.km
Top District based on Literacy: Kanyakumari – 91.75%

Kerala
Top District based on Population: Palakkad – 2.8M
Top District based on Palakkad: Erode –4,480 sq.km
Top District based on Literacy: Kottayam – 97.21%

Conclusion

The main objective of this project was to determine the similarity and dissimilarity of both states. We have successfully compared both the states with our available resources on the basis of area, population, literacy percentage and common venue. Even though both these southern states are quite similar, we can't guarantee it based on only these features. There are various other aspects that need to be considered to come to a definite point. As far as the clustering is done in this project, we can see that both the cities are similar in certain venues and also dissimilar in others. We can surely state that the most common venue in both states is Indian Restaurants. Also, due to the restriction in the radius that can measure in Foursquare API is limited to 100km, this analysis doesn't include the entire area of the districts. So, in the future, I hope that this restriction is removed, and further analysis can be conducted to help us get more insights into these districts.