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 southern most 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. The southern India is spread over an area of 635,780 Sq Km 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 and largest city 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 passage of the States Reorganisation Act, by combining Malayalam-speaking regions. Spread over 38,863 km2 (15,005 sq mi), 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 southern most states of India. We will use Foursquare data as well as machine learning algorithms such as segmentations and clustering to help us complete this study. The reason of 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 also 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 a much 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.