Applied Data Science Specialization
Applied Data Science Capstone Project: The Battle of Neighborhoods
Setting-up Educational Institutions in India
By Smriti Goyal
August 10, 2020

Introduction

Literacy in India is a key for socio-economic progress. Despite government programs, increase in India's literacy rate has been very slow. The 2011 census, indicated a 2001–2011 decadal literacy growth of 9.2%, which is slower than the growth seen during the previous decade. One of the main factors contributing to this relatively low literacy rate is usefulness of education and availability of schools in vicinity in rural areas. There is a shortage of classrooms to accommodate all the students in 2006–2007. In addition, there is no proper sanitation in most schools. The study of 188 government-run primary schools in central and northern India revealed that 59% of the schools had no drinking water facility and 89% no toilets. In 600,000 villages and multiplying urban slum habitats, 'free and compulsory education' is the basic literacy instruction dispensed by barely qualified 'para teachers'. The average pupil teacher ratio for all India is 42:1, implying a teacher shortage. Such inadequacies resulted in a non-standardized school system where literacy rates may differ. Furthermore, the expenditure allocated to education was never above 4.3% of the GDP from 1951 to 2002 despite the target of 6% by the Kothari Commission. This further complicates the literacy problem in India.

Business Problem

The objective of this project is to analyze and select the best locations in India to open schools. This project is mainly focused on the analysis of 2011 census data to understand where there is a need to start- up educational institutions. Using data science methodology and machine learning techniques like clustering and data visualization, this project aims to provide solutions to low literacy rates in India.

Data

To solve the problem, we will need the following data:

- 2011 district census data. This dataset contains data like literacy rate, population, sex-ratio and growth rate.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

Sources of data and the methods to extract the data

I got the 2011 district census data from the official Indian census website. I have used web scraping techniques to extract the data from the census page, with the help of Python requests and beautiful soup packages. Then we can get the latitude and longitude coordinates

of the neighborhoods using Python Geocoder package. After that, I have used the Foursquare API to get the venue data for those neighborhoods.
Foursquare API will provide many categories of the venue data, and we are particularly interested in the residence category in order to help us solve the problem. This is a project that will make use of many data science skills, from web scraping (census data), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).