**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).

**Methodology**

First step is to get a list of districts and their census data for the year 2011. This dataset can be acquired from the official Indian census website:

https://www.census2011.co.in/district.php

To extract the data, we will do scraping using Python Requests and Beautiful soup package and convert it into a data frame. This provides us with a list of districts and literacy rate, growth rate, state, population and sex ratio data pertaining to each district. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Python Geocoder package that will allow us to convert the address into geographical coordinates in the form of latitude and longitude. Once we have obtained the latitude and longitude coordinates for all the places, we need to merge the coordinates into the original data frame. After gathering the data, we can start using data visualizing and clustering tools. Now we can utilize the foursquare API. The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which you can send requests, so there's really nothing to download onto your server.

**Analysis of data and Result**

This is our main dataset:

Out[33]:

| | # | District | State | Population | Growth | Sex-Ratio | Literacy |
|---|---|---|---|---|---|---|---|
| 0 | 1 | Thane | Maharashtra | 11060148 | 36.01 % | 886 | 84.53 |
| 1 | 2 | North Twenty Four Parganas | West Bengal | 10009781 | 12.04 % | 955 | 84.06 |
| 2 | 3 | Bangalore | Karnataka | 9621551 | 47.18 % | 916 | 87.67 |
| 3 | 4 | Pune | Maharashtra | 9429408 | 30.37 % | 915 | 86.15 |
| 4 | 5 | Mumbai Suburban | Maharashtra | 9356962 | 8.29 % | 860 | 89.91 |
| 5 | 6 | South Twenty Four Parganas | West Bengal | 8161961 | 18.17 % | 956 | 77.51 |
| 6 | 7 | Barddhaman | West Bengal | 7717563 | 11.92 % | 945 | 76.21 |
| 7 | 8 | Ahmadabad | Gujarat | 7214225 | 24.03 % | 904 | 85.31 |
| 8 | 9 | Murshidabad | West Bengal | 7103807 | 21.09 % | 958 | 66.59 |
| 9 | 10 | Jaipur | Rajasthan | 6626178 | 26.19 % | 910 | 75.51 |
| 10 | 11 | Nashik | Maharashtra | 6107187 | 22.30 % | 934 | 82.31 |
| 11 | 12 | Surat | Gujarat | 6081322 | 42.24 % | 787 | 85.3 |

The basis of our problem are the low literacy rates so to analyze the data I decided to focus on literacy rate and how it effects the other properties like population, sex ratio, growth rate etc. I started with calculating the average literacy rate amongst the districts and used that figure as a marker: considering literacy rates below it to be low and above it to be decent.

```
In [39]: avg_Literacy = df_data['Literacy'].mean()
         avg_Literacy

Out[39]: 72.30842187499996
```

Considering the districts with low literacy rates:

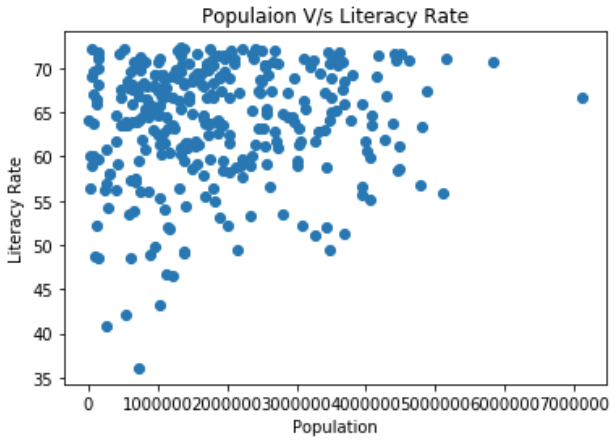
```
In [40]: df_literacy72 = df_data[df_data.Literacy <= 72.30]
         df_literacy72
```

Out[40]:

| | # | District | State | Population | Growth | SexRatio | Literacy |
|---|---|---|---|---|---|---|---|
| **8** | 9 | Murshidabad | West Bengal | 7103807 | 21.09 | 958 | 66.59 |
| **14** | 15 | Patna | Bihar | 5838465 | 23.73 | 897 | 70.68 |
| **18** | 19 | East Godavari | Andhra Pradesh | 5154296 | 5.16 | 1006 | 70.99 |
| **19** | 20 | Purbi Champaran | Bihar | 5099371 | 29.43 | 902 | 55.79 |
| **21** | 22 | Guntur | Andhra Pradesh | 4887813 | 9.47 | 1003 | 67.40 |
| **23** | 24 | Muzaffarpur | Bihar | 4801062 | 28.14 | 900 | 63.43 |
| **25** | 26 | Moradabad | Uttar Pradesh | 4772006 | 25.22 | 906 | 56.77 |
| **29** | 30 | Azamgarh | Uttar Pradesh | 4613913 | 17.11 | 1019 | 70.93 |
| **35** | 36 | Jaunpur | Uttar Pradesh | 4494204 | 14.89 | 1024 | 71.55 |
| **36** | 37 | Madhubani | Bihar | 4487379 | 25.51 | 926 | 58.62 |
| **37** | 38 | Sitapur | Uttar Pradesh | 4483992 | 23.88 | 888 | 61.12 |
| **38** | 39 | Bareilly | Uttar Pradesh | 4448359 | 22.93 | 887 | 58.49 |
| **39** | 40 | Gorakhpur | Uttar Pradesh | 4440895 | 17.81 | 950 | 70.83 |
| **40** | 41 | Agra | Uttar Pradesh | 4418797 | 22.05 | 868 | 71.58 |
| **41** | 42 | Gaya | Bihar | 4391418 | 26.43 | 937 | 63.67 |
| **43** | 44 | Visakhapatnam | Andhra Pradesh | 4290589 | 11.96 | 1006 | 66.91 |
| **44** | 45 | Samastipur | Bihar | 4261566 | 25.53 | 911 | 61.80 |

Plotting the district literacy rates against district population, I realized that districts with low literacy rates have low population densities:

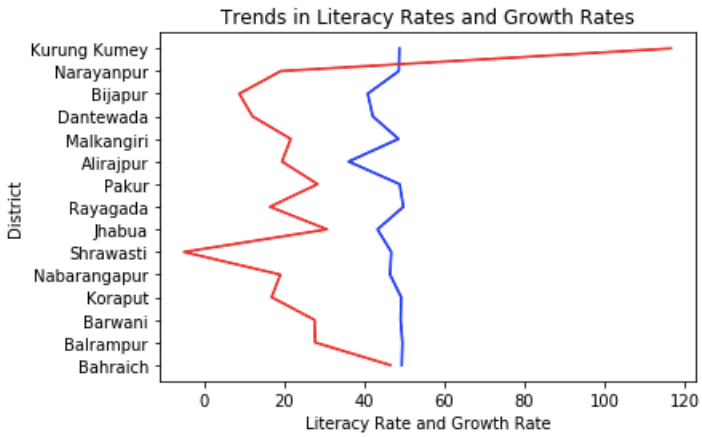
```
In [41]: plt.scatter(df_literacy72.Population, df_literacy72.Literacy)
         plt.xlabel("Population")
         plt.ylabel("Literacy Rate")
         plt.title("Populaion V/s Literacy Rate")
         plt.show()
```




Districts with literacy rates less than the avg have low population densities.

I further reduced the dataset to districts with literacy rates less than 50% to understand how literacy rate effects the growth rate of districts. I found that districts with low literacy rates also have low growth rates.

```
In [43]: plt.plot(df_literacy50.Literacy, df_literacy50.District, color = 'blue')
         plt.plot(df_literacy50.Growth, df_literacy50.District, color = 'red')
         plt.ylabel("District")
         plt.xlabel("Literacy Rate and Growth Rate")
         plt.title("Trends in Literacy Rates and Growth Rates")
         plt.show()
```




Districts with low literacy rates also have low growth rates.

To understand how literacy rates and sex ratios of the districts are related, I created a subset of districts with high sex ratio from the original dataset.  I found out that most districts with high sex ratios have above average literacy rates.

```
In [44]: avg_SexRatio = df_data['SexRatio'].mean()
         avg_SexRatio
```

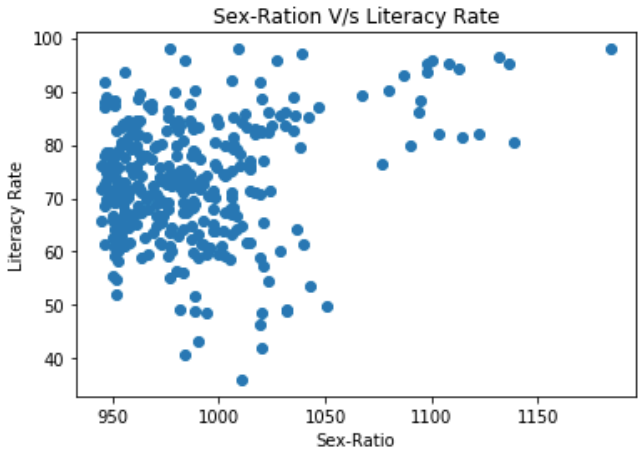
```
Out[44]: 945.4328125
```

```
In [45]: df_SexR945 = df_data[df_data.SexRatio >= 945]
         df_SexR945
```

Out[45]:

| | # | District | State | Population | Growth | SexRatio | Literacy |
|---|---|---|---|---|---|---|---|
| **1** | 2 | North Twenty Four Parganas | West Bengal | 10009781 | 12.04 | 955 | 84.06 |
| **5** | 6 | South Twenty Four Parganas | West Bengal | 8161961 | 18.17 | 956 | 77.51 |
| **6** | 7 | Barddhaman | West Bengal | 7717563 | 11.92 | 945 | 76.21 |
| **8** | 9 | Murshidabad | West Bengal | 7103807 | 21.09 | 958 | 66.59 |
| **13** | 14 | Paschim Medinipur | West Bengal | 5913457 | 13.86 | 966 | 78.00 |
| **15** | 16 | Hugli | West Bengal | 5519145 | 9.46 | 961 | 81.80 |
| **16** | 17 | Rangareddy | Andhra Pradesh | 5296741 | 48.16 | 961 | 75.87 |
| **17** | 18 | Nadia | West Bengal | 5167600 | 12.22 | 947 | 74.97 |
| **18** | 19 | East Godavari | Andhra Pradesh | 5154296 | 5.16 | 1006 | 70.99 |
| **21** | 22 | Guntur | Andhra Pradesh | 4887813 | 9.47 | 1003 | 67.40 |
| **24** | 25 | Belgaum | Karnataka | 4779661 | 13.41 | 973 | 73.4 |
| **27** | 28 | Nagpur | Maharashtra | 4653570 | 14.40 | 951 | 88. |

```
In [46]: plt.scatter(df_SexR945.SexRatio, df_SexR945.Literacy)
         plt.xlabel("Sex-Ratio")
         plt.ylabel("Literacy Rate")
         plt.title("Sex-Ration V/s Literacy Rate")
         plt.show()
```




Most districts with high Sex Ratios have above average Literacy Rates.

I further decided to use the df.groupby function to get the states with districts that have literacy rates lower than the average value.

The states with literacy rates lower than the average are:

1. Andhra Pradesh
2. Arunachal Pradesh
3. Assam
4. Bihar
5. Chhattisgarh
6. Gujarat
7. Haryana
8. Jammu and Kashmir
9. Jharkhand
10. Karnataka
11. Madhya Pradesh
12. Orissa
13. Punjab
14. Rajasthan
15. Uttar Pradesh

```
In [47]: df_group = df_literacy72.groupby(['State', 'Literacy'])
         df_group.groups
```

```
('Andhra Pradesh', 67.4): Int64Index([21], dtype='int64'),
('Andhra Pradesh', 68.9): Int64Index([126], dtype='int64'),
('Andhra Pradesh', 70.99): Int64Index([18], dtype='int64'),
('Andhra Pradesh', 71.53): Int64Index([46], dtype='int64'),
('Arunachal Pradesh', 48.75): Int64Index([616], dtype='int64'),
('Arunachal Pradesh', 52.19): Int64Index([612], dtype='int64'),
('Arunachal Pradesh', 56.46): Int64Index([638], dtype='int64'),
('Arunachal Pradesh', 59.0): Int64Index([632], dtype='int64'),
('Arunachal Pradesh', 59.8): Int64Index([597], dtype='int64'),
('Arunachal Pradesh', 59.99): Int64Index([636], dtype='int64'),
('Arunachal Pradesh', 60.02): Int64Index([623], dtype='int64'),
('Arunachal Pradesh', 63.8): Int64Index([621], dtype='int64'),
('Arunachal Pradesh', 64.1): Int64Index([639], dtype='int64'),
('Arunachal Pradesh', 66.46): Int64Index([611], dtype='int64'),
('Arunachal Pradesh', 67.07): Int64Index([620], dtype='int64'),
('Arunachal Pradesh', 68.18): Int64Index([599], dtype='int64'),
('Arunachal Pradesh', 69.13): Int64Index([629], dtype='int64'),
('Assam', 58.34): Int64Index([239], dtype='int64'),
('Assam', 63.08): Int64Index([461], dtype='int64'),
('Assam', 63.55): Int64Index([546], dtype='int64'),
```

**Discussion**

Literacy in India is a key for socio-economic progress. Despite government programs, increase in India's literacy rate has been very slow. One of the main reasons for this relatively low literacy rate is the insufficient education and availability of schools in vicinity in rural areas. Low literacy rates lead to low sex ratios and growth rates. This is a cause for concern as this leads to a wide gender disparity in the literacy rate in India. The low female literacy rate has a dramatically negative impact on family planning and population stabilization efforts in India. Studies have indicated that female literacy is a strong predictor of the use of contraception among married Indian couples, even when women do not otherwise have economic independence. Severe caste disparities also exist. Discrimination of lower castes has resulted in high dropout rates and low enrollment rates. The National Sample Survey Organization and the National Family Health Survey collected data in India on the percentage of children completing primary school which are reported to be only 36.8% and 37.7% respectively. This further complicates the literacy problem in India.

**Conclusion**

There is a need to put focus of on this issue of literacy. Investment should be made in educating lower caste and rural communities as well as there is a need to focus on increasing female literacy in India.