Simplifying 2011 Census Data

Parvez Khan, Adeeba Yusuf, Tanay Pathode

1. Introduction

1.1 Overview

We have explored socio - economic data for districts in India in which we have tried to dive into the unique narrative of each region of India, uncovering patterns in population growth, literacy rates, gender ratios, and more. Socio-economic data presents a rich story, where each district's data serves as a "character" with its own background and unique attributes. This dataset allows us to visualize and understand these dynamics, turning raw numbers into insights that can shape policies and drive development.
Studying this dataset on socio-economic factors across districts in India offers a window into the regional diversity and development patterns within the country. This data-driven understanding can help identify regions that may require targeted interventions, support equitable resource allocation, and enhance policy planning.

1.2 Project Motivation & Objective

Here is our hypothetical *Problem Statement* for this project:

The Indian government launched a scheme to provide free education for people upto class 12th. Now, after 10 years, the government wants to assess whether the scheme is truly helping people in order to decide if it should be continued, improved, or abolished as it may be a waste of government funds. The Government also aims to find which districts are thriving well.

We as data Scientists have collected data of several socio economic factors from the census that could tell about the effectiveness of scheme in different districts and made a data visualization app to discuss which districts had really benefited from the scheme.

1.3 Questions We Imposed

Here are a few questions laid down by us looking at the dataset:

- What is the relationship between overall literacy rates and population density?
- Does a higher literacy rate correlate with a more balanced sex ratio?
- How do literacy rates and sex ratios compare across different states?
- Is there a correlation between literacy rates and population growth across different states or regions?
- Are there districts where the disparity between male and female literacy is particularly high?

1.4 Libraries we used in this project

- rvest
- tidyverse
- dplyr
- ggplot2

- sf
- leaflet
- leaflet.extras
- shiny
- shinydashboard
- shinydashboardPlus

{r include=FALSE} library(rvest) library(tidyverse) library(dplyr) library(ggplot2) library(sf) library(leaflet) library(leaflet.extras)

2. Dataset

{r include=FALSE} district_data<- read.csv("district_data.csv") attach(district_data)

state_data <- read.csv("state_data.csv")

shp_data <- st_read("India-Districts-2011Census.shp")

shp_data <- merge(district_data, shp_data[,5:6], by.x = "District.Code", by.y = "censuscode") %>% relocate(District.Code, .after = State) %>% arrange(State,District) %>% st_as_sf()

2.1 Source and method of obtaining dataset

The dataset used for the analysis has been taken from 2011 census of India which is a collection of demographic, economic, and social information about the population of India. We have particularly used the dataset where the information of districts were provided.

2.2 Describing dataset variables

The dataset we obtained was having data for 640 different districts of India. We scraped the data of 10 variables from the website. These variables are described as follows:

District Contains the name of the district.

State It gives the name of state in which district is located.

Population Gives the figure of the population of the district.

Population density Indicates the number of people living in per square kilometer.

Growth Tells the change in the number of individuals living in a particular area.

Literacy This variable gives the percentage of the adult population aged 15 years and over which are literate (The Census defines a person as a literate if he/she is able to read and write a sentence in any language with proper understanding of what they are reading or writing).

Male literacy This variable gives the percentage of the adult male population aged 15 years and over which are literate .

Female literacy This variable gives the percentage of the adult female population aged 15 years and over which are literate .

Child Sex ratio This gives the number of females per 1000 males in the age group 0–6 years.

Sex Ratio This gives the number of females per 1000 males.

3. Obtaining the dataset

We have initialized two vectors for the extracted tables and district URLs for the 22 pages of the Census website. A loop runs through all 22 pages of the Census website, scraping data from each page and first table has been extracted and stored iin variable

Then the district URLs are extracted and are combined in vector. The individual tables from each page are then combined into a single table.

A second loop extracts the population density, child sex ratio, and literacy rates data from the third table on each district's specific page. The values are parsed from the relevant rows, and the extracted data is stored in the pre-defined vectors. We then downloaded a district code list from an external source and merged with the current. We then converted it into a csv file and saved the dataset.

4. Bias in the dataset

1. The Indian census employs a door-to-door survey method so due to logistical issue the data from regional and rural can be underrepresented.
2. Factors such as migration, homelessness, and difficulty in accessing certain populations (e.g., in conflict zones) can lead to under counting.
3. The data provided by a person could be false .
4. In Metropolitan cities people keep moving with a very fast pace which makes it impossible to collect the accurate data.

5. Important Visualizations

We now try to answer the imposed question by visualizing the data

What is the relationship between overall literacy rates and population density? {r include=FALSE} # There seems to have some outliers in in Population Density column, so we remove the first.

upper_bound <- quantile(Population.Density, 0.75) + IQR(Population.Density) new_data <- district_data[Population.Density < upper_bound,]

{r echo=FALSE, message=FALSE, warning=FALSE, results='hide'} library(ggplot2)

```
ggplot(new_data, aes_string(y = new_data$Literacy, x = new_data$Population.Density)) +
  geom_point(color = "blue", size = 2, alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(x = "Literacy", y="Population density",
        title = "Relationship between Literacy and population density") +
  theme(plot.background = element_rect(color = "black", fill = NA, size = 0.5))
```

From the plot, it can be seen that there is absolutely no correlation between Literacy and Population Density. This can potentially mean there may be other factors which are affecting Literacy or Population Density.

Does a higher literacy rate correlate with a more balanced sex ratio? {r echo=FALSE, message=FALSE, warning=FALSE, results=‘hide’} library(ggplot2) ggplot(district_data, aes_string(x = Literacy, y = Sex.Ratio)) + geom_point(color = “blue”, size = 3, alpha = 0.6) + geom_smooth(method = “lm”, color = “red”, se = TRUE) + labs(x = “Literacy”, y=“Sex Ratio”, title = “Correlation between Literacy and Sex Ratio”) + theme(plot.background = element_rect(color = “black”, fill = NA, size = 0.5))

We see that as literacy increases sex ratio become more balanced.

As literacy rates increase, there is often a shift in societal attitudes towards gender equality. Education tends to promote more progressive values, reducing discriminatory practices such as gender-based preferences that can lead to an imbalanced sex ratio. Literacy also fosters awareness of gender equality, reducing gender biases that might favor one gender over another, especially in family planning thus reducing female infanticide.

How do literacy rates and sex ratios compare across different states? {r echo=FALSE} pal <- colorNumeric(“Greens”, domain = shp_data[[“Literacy”]], na.color = “transparent”)

line1 <- paste0(toupper(district_data$District), ”, ”, toupper(district_data$State)) line2 <- paste0(“Literacy”, “:”, get(“Literacy”)) popup <- paste(line1, line2, sep = “”)

leaflet(shp_data) %>% addProviderTiles(providers$OpenStreetMap.Mapnik) %>% setView(lat = “22.512”, lng = “80.329”, zoom = 4) %>%

```
  addPolygons(fillColor = ~pal(get("Literacy")),
              color = "black",
              weight = 1,
              fillOpacity = 0.7,
              popup = ~popup
  ) %>%
  addLegend("bottomright", pal = pal, values = shp_data[["Literacy"]],
            title = "Literacy Heatmap",
            opacity = 0.7)
```

We can infer from the heatmap that Kerala have highest literacy and some districts of Mizoram have greater literacy rates

{r echo=FALSE} pal <- colorNumeric(“Greens”, domain = shp_data[[“Sex.Ratio”]], na.color = “transparent”)

line1 <- paste0(toupper(district_data$District), ”, ”, toupper(district_data$State)) line2 <- paste0(“Sex.Ratio”, “:”, get(“Sex.Ratio”)) popup <- paste(line1, line2, sep = “”)

leaflet(shp_data) %>% addProviderTiles(providers$OpenStreetMap.Mapnik) %>% setView(lat = “22.512”, lng = “80.329”, zoom = 4) %>%

```
  addPolygons(fillColor = ~pal(get("Sex.Ratio")),
              color = "black",
              weight = 1,
              fillOpacity = 0.7,
              popup = ~popup
  ) %>%
  addLegend("bottomright", pal = pal, values = shp_data[["Sex.Ratio"]],
            title = "Sex Ratio Heatmap",
            opacity = 0.7)
```

From this heatmap we observe that Kerala have higher sex ratio as compared to other state and also Pauri Garwal district in Uttarakhand have high sex ratio.

Is there a correlation between literacy rates and population growth across different states or regions? {r echo=FALSE, message=FALSE, warning=FALSE, results='hide'} ggplot(district_data, aes_string(x = Literacy, y = Population.Growth)) + geom_point(color = "blue", size = 3, alpha = 0.6) + geom_smooth(method = "lm", color = "red", se = TRUE) + labs(x = "Literacy", y="Population growth", title = "Correlation between Literacy and Population growth") + theme(plot.background = element_rect(color = "black", fill = NA, size = 0.5))

We infer from the graph that as literacy rate increases population growth decreases.This is expected as literate individuals, especially women, are more likely to be aware of family planning methods and have access to information about contraception, enabling them to make informed choices about family size.

Are there districts where the disparity between male and female literacy is particularly high? First we check state wise disparity:

{r echo=FALSE, message=FALSE, warning=FALSE, results='hide'} rel_dif <- (state_data$Male_{L}iteracy - state_{d}ata$Female_Literacy)/state_data$Male_Literacy

ggplot(state_data, aes(x = reorder(State, rel_dif), y = rel_dif)) + geom_bar(stat = "identity", fill = "skyblue") + labs(x = "State", y="Literacy Disparity", title = "Plot of Disparity between Male and Female Literacy") + theme(plot.background = element_rect(color = "black", fill = NA, size = 0.5)) + coord_flip()

It can be seen from the plot that Rajasthan has the highest disparity between Male and Female literacy followed by Jharkhand and Bihar.

Now, we check district wise disparity in a geographical heatmap:

{r include=FALSE} shp_data$Literacy.Disparity <- (Male.Literacy - Female.Literacy)/Male.Literacy attach(shp_data)

{r echo=FALSE} pal <- colorNumeric("Reds", domain = shp_data[["Literacy.Disparity"]], na.color = "transparent")

line1 <- paste0(toupper(district_data$District), ", ", toupper(district_{d}ata$State)) line2 <- paste0("Literacy Disparity", ":", get("Literacy.Disparity")) popup <- paste(line1, line2, sep = "")

leaflet(shp_data) %>% addProviderTiles(providers$OpenStreetMap.Mapnik) %>% setView(lat = "22.512", lng = "80.329", zoom = 4) %>%

```
addPolygons(fillColor = ~pal(get("Literacy.Disparity")),
            color = "black",
            weight = 1,
            fillOpacity = 0.7,
            popup = ~popup
) %>%
addLegend("bottomright", pal = pal, values = shp_data[["Literacy.Disparity"]],
          title = "Literacy Disparity Heatmap",
          opacity = 0.7)
```

From The heatmap it can be seen that most districts of Rajasthan, and some districts of Bihar, Jharkhand, Chattisgarh and Jammu & Kashmir have relatively high disparity between literacy rates of male and female, thereby suggesting more gender focused education initiatives in those states.

5. Conclusion

In analyzing the socio-economic data of India's districts, we have uncovered valuable insights into literacy, population density, sex ratio, and other demographic factors. This data analysis and visualization effort highlights distinct patterns and disparities across the country, enabling a nuanced understanding of regional strengths and areas requiring focused intervention.

From the visualizations, we observe several notable trends:

Literacy and Population Density: Our analysis reveals no clear correlation between literacy and population density. This suggests that factors beyond population density—such as infrastructure, economic opportunities, and policy focus—may more directly influence literacy.

Literacy and Sex Ratio: There is a positive correlation between literacy and a balanced sex ratio. Regions with higher literacy rates often exhibit a more balanced sex ratio, likely due to increased awareness around gender equality and family planning.

State Comparisons: Literacy rates and sex ratios vary considerably by state. Kerala, for example, stands out for high literacy and a balanced sex ratio, while some districts in Rajasthan and Bihar exhibit significant disparities in male and female literacy rates.

Literacy and Population Growth: Our data supports the idea that increased literacy, particularly among women, correlates with lower population growth rates, likely due to enhanced awareness of family planning.

Gender Disparities in Literacy: There is a significant disparity between male and female literacy rates in specific regions, especially in Rajasthan, Bihar, and Jharkhand. These findings underscore the need for targeted educational programs to bridge gender gaps in literacy, particularly in these states.

6. Resources

We would like to express our gratitude to following websites for providing valuable resources and data that were essential to the completion of this project. The information available on the platform greatly contributed to the insights and findings presented here.

Data scrapping websites:

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

District boundary data: https://github.com/datameet/maps/tree/master/Survey-of-India-Index-Maps/Boundaries

To know about how census is done: http://new.census.gov.in/census.website/node/325