

# PROJECT REPORT SUBMITTED TO Symbiosis Institute of Geoinformatics

## FOR PARTIAL FULFILLMENT OF THE M. Sc. DEGREE

By Dilpreet Kaur PRN-22070243017

M.Sc. (Data Science and Spatial Analytics)

BATCH-2022-24

## **TOPIC-LITERACY RATE OF INDIA: 1991-2011**

Symbiosis International (Deemed University)
5th Floor, Atur Centre
Gokhale Cross Road
Model Colony
Pune – 411016
Maharashtra
India

## **ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to Dr. Moushmi Das Gupta for their invaluable guidance, encouragement, and support throughout the project. This project would not have been possible without her expertise and unwavering support.

I would like to extend my gratitude to Symbiosis Institute of Geoinformatics for their invaluable contributions, support, and cooperation. I would also like to acknowledge my teachers for their support and assistance during the course of this project.

Lastly, I would like to extend my sincere thanks to all my friends for their cooperation and support. I take complete responsibility for any errors or omissions in this report

### **DECLARATION**

I, Dilpreet Kaur (PRN: 22070243017), student of M.Sc. Data Science and Spatial Analytics 2022-2024 certify that this project is my own work, based on my personal study and/or research and that I have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication. I also certify that this project has not previously been submitted for assessment in any academic capacity, and that I have not copied in part or whole or otherwise plagiarised the work of other persons. I confirm that I have identified and declared all possible conflicts that I may have.

Dilpreet Kaur

PRN:22070243017

Date: 31 Jan 2023

### **ABSTRACT**

The literacy rate in India has been a topic of concern for policymakers and researchers for many years. This report aims to provide an overview of the current literacy rate in India, as well as the value benefit of conducting a study on this topic. The report will examine data on literacy rates at the national and state, and will also explore the factors that contribute to variation in literacy rates across different regions/ zones of the country. Additionally, the report will discuss the potential benefits of increasing literacy rates in India, including improved economic growth and development, increased access to education and job opportunities, and a reduction in poverty and inequality. We will get to know the regions with low literacy rates and distribution of male-females literacy across the nation. Overall, this report aims to demonstrate the importance of investing in literacy programs and policies in order to promote sustainable development and improve the lives of all citizens in India.

### INTRODUCTION

Literacy is a fundamental human right and plays a crucial role in the social and economic development of a nation. In India, despite various efforts by the government and non-government organizations, the literacy rate has remained a challenge. This report aims to provide a comprehensive analysis of the literacy rate in India over the last two decades, from 1991 to 2011.

The need for this study arises from the importance of understanding the trends and patterns in literacy rate and identifying the areas where further efforts are required to improve the same. The report will also provide an overview of the government initiatives and other interventions aimed at improving literacy in India.

The findings of this report will provide a deeper insight into the current literacy scenario in India and highlight the challenges faced by the population in terms of accessing education and achieving literacy. Based on the data analysis, the report will make recommendations for future efforts to improve the literacy rate in the country. This report is intended to be a valuable resource for policymakers, researchers, and other stakeholders concerned with education and literacy in India.

This report provides an in-depth analysis of the literacy rate in India over the last two decades. The data analysed covers a period from 1991-2011 and highlights the changes, trends and challenges faced by the Indian population in terms of accessing education and achieving literacy. The report will shed light on the efforts made by the government and non-government organizations to improve the literacy rate and provide education to all. This report aims to provide a comprehensive overview of the literacy scenario in India and highlight the key findings and recommendations for further improvement.

#### **ABOUT DATA**

The literacy rate is an important indicator of the level of education and human development in a country. By choosing a dataset on literacy rate of India, one can analyse the trends and patterns of literacy rate over time and across different regions of India and identify areas and communities with low literacy rate, which can be targeted for improvement through policy interventions.

Overall, the literacy rate dataset is useful for understanding the current state of education and formulating plans for improvement.

I choose a dataset on the literacy rate of India for two decades i.e., 1991-2011. It consisted of the total literacy rate over the nation for all the three years along with male female distribution for the same. The data even consisted of the rural urban distribution for 2011 and male-female distribution for urban rural regions of all the states/UT's of India which helped me demonstrate trends in literacy rate across nation.

### **METHODOLOGY**

The purpose of this project is to understand the literacy rate of India. The literacy rate is an important indicator and having accurate information on the same can be useful for researchers and policy makers to make improvements. In this project, we are using a literacy rate dataset to understand the trends for three years across the nation.

#### **DATA COLLECTION**

The data consisted for the literacy rate for 2 decades for the states/UT's. Data taken from Statistical Year Book of India - 2013. The 1991 data refer to Census of India. Data for 1991 2001 is from Office of the Registrar General of India.

Link: https://data.gov.in/resource/literacy-rate-india-nsso-and-rgi#

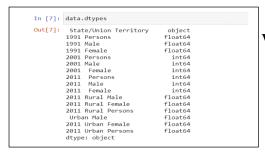
#### **DATA EXPLORATION**

The first step we did was exploring the data. We loaded the data from a csv file and explored various features.

We loaded the csv file and explored the features where we can see that the dataset consists of data of 3 years total literacy rate and also the male-female distribution. The data also had rural-urban distribution for 2011. The data consisted of some missing values which will be treated in the next steps.

[3]:	data	data.head()															
[3]:		State/Union Territory	1991 Persons	1991 Male	1991 Female	2001 Persons	2001 Male	2001 Female	2011 Persons	2011 Male	2011 Female	2011 Rural Male	2011 Rural Female	2011 Rural Persons	Urban Male	2011 Urban Female	2011 Urban Persons
	0	India	52.0	64.0	39.0	65	75	54	74	82	65	NaN	NaN	NaN	NaN	NaN	NaN
	1	Andhra Pradesh	44.0	55.0	33.0	61	70	50	68	76	60	70.0	52.0	61.0	86.0	75.0	81.0
	2	Arunachal Pradesh	42.0	52.0	30.0	54	64	44	67	74	60	69.0	54.0	62.0	89.0	79.0	85.0
	3	Assam	53.0	62.0	43.0	63	71	55	73	79	67	77.0	64.0	70.0	92.0	86.0	89.0
	4	Bihar	38.0	51.0	22.0	47	60	33	64	73	53	72.0	51.0	62.0	84.0	72.0	79.0

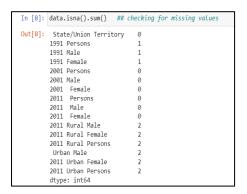
The data consisted of **36 rows** i.e., all the states and union territories of India. It consisted of **16 columns** for the distribution of the literacy rate across the nation for years 1991,2001 and 2011.



We can see here our data is in **numeric format** as it is the literacy rate distribution. We can analyse the changes and trends on the basis of this distribution.

#### **DATA PRE-PROCESSING**

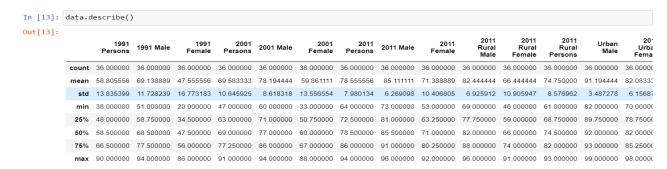
The data collected was subjected to several pre-processing steps to ensure its quality and suitability for analysis. The first step was to check for **missing** or incorrect values and handle them in an appropriate manner.



The missing values in the dataset for the numeric variable were filled with the **mean** of the column as it will help to maintain the normal distribution of the data.

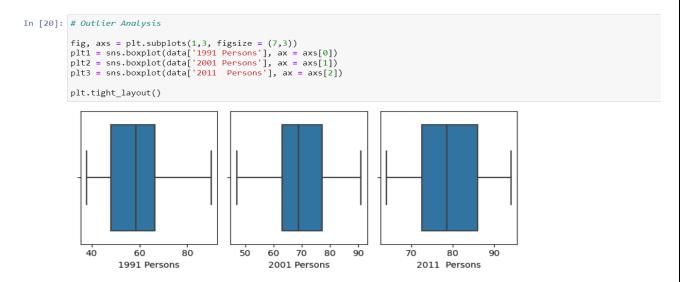


Though **description** of the data comes under data exploration but since data consisted of some missing values and it could have affected the statistical summary thus we treated the missing values first before describing the data.



Through this **statistical summary** we can understand the maximum and minimum literacy rate distribution across the nation for 2 decades. Also we can find the mean or average literacy rate for the states/UT's.

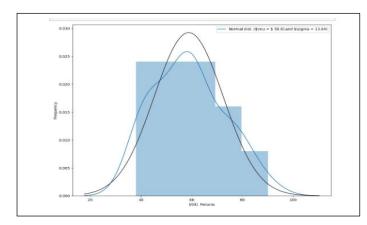
As the data comprised of total literacy rate across nation so we analysed the **outliers** for each year. And we can see that the data does not consists of any outliers.

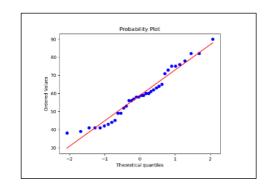


#### **DATA ANALYSIS**

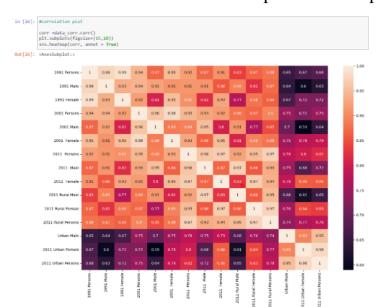
We checked the **distribution** of the data. We created a histogram visualization of the distribution of the '1991 Persons' column in the data frame to check if the column is normally distributed. The data was normally distributed. Normal distribution is important for linear regression models as many of their underlying assumptions are based on a normal distribution.

The probability scatter plot is also used with the help of stats.probplot function that shows the observed values of the data against their theoretical quantiles if they follow a normal distribution. A line is drawn to show the agreement between the observed values and the theoretical quantiles.





**Correlation** between the features is necessary to depict the strength of the linear relation between pair of variables. Hence, we plotted a correlation map between the variables with the help of a heat map.



Here we can see that the literacy rate for different years is almost related to each other. Major differences can be seen along the urban rural literacy rates across the nation for 2 decades.

We tried perform **advanced analysis** with the help of Linear Regression using Stats Model. OLS (**Ordinary Least Squares**) regression analysis, which is a type of linear regression was performed to estimates the relationship between independent variables and a dependent variable. In this case, the dependent variable is "1991 Persons" and the independent variables were "1991 Male" and "1991 Female"

The summary table provides information about the **fit of the regression model** to the data. The values in the table are used to evaluate the goodness of fit and the statistical significance of the individual coefficients (the estimates of the relationships between the independent and dependent variables).

```
In [79]: import statsmodels.api as sm

# Define the dependent variable (y) and independent variable(s) (X)
y = data['1991 Persons']
X = data[['1991 Male', '1991 Female']]

# Add a constant to the independent variable(s) for the intercept ter
X = sm.add_constant(X)

# Fit the linear regression model
model = sm.OLS(y, X).fit()

# Print a summary of the model's results
print(model.summary())
```

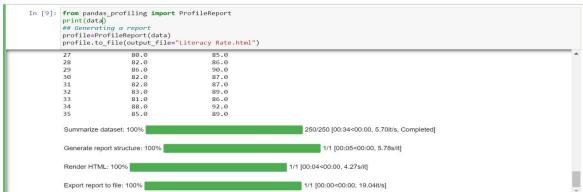
Dep. Variable:		1991 Persons	R-squar	red:	0.999		
Model:		OLS	Adj. R-	squared:	0.999		
Method:		Least Squares	F-stati	stic:	1.523e+04		
Date:	Sun	, 12 Feb 2023	Prob (F	-statistic):	1.16e-49		
Time:		10:51:40	Log-Lik	elihood:	-22.237		
No. Observation	S:	36	AIC:		50.47		
Of Residuals:		33	BIC:			55.22	
Of Model:		2					
Covariance Type	:	nonrobust					
				- 1.1			
		std err				0.975]	
const						2.363	
1991 Male	0.5056	0.019	26.615	0.000	0.467	0.544	
1991 Female	0.4843	0.013	36.458	0.000	0.457	0.511	
 Omnibus:		0.724	 -Durbin	Watson:		2.448	
Prob(Omnibus):		0.696	Jarque-	Bera (JB):	0.729		
Skew:			Prob(JE	, ,		0.694	
(urtosis:	2.319	Cond. N	lo.		838.		

- **R\_squared** is the coefficient of determination, which is a measure of how well the model fits the data. We have the value as 0.9 which indicates it as good fit.
- 1991 Male/1991 Female in the summary table explains the coefficients for the independent variables "1991 Male" and "1991 Female", respectively. These coefficients represent the estimated changes in the dependent variable for a one-unit change in the independent variable, holding all other variables constant.
- The **t-statistic**, which is used to test the statistical significance of each coefficient.
- **P**>|**t**| where the p-value is associated with the t-statistic, indicates the probability of observing the t-statistic under the assumption that the true coefficient is 0. A low p-value indicates that the coefficient is significantly different from 0.
- [0.025 0.975] indicates the 95% confidence interval for the coefficient, which provides a range of values that are likely to contain the true coefficient with 95% confidence.

Hence the summary helps us to determine the significance of the variables and also their relationship

Since the data was normally distributed hence, we performed the analysis on the small portion of the data.

Exploratory Data Analysis(EDA) was performed using **profile\_reporting** to gain insights of the data.



Here attached is the report:



Literacy Rate.html

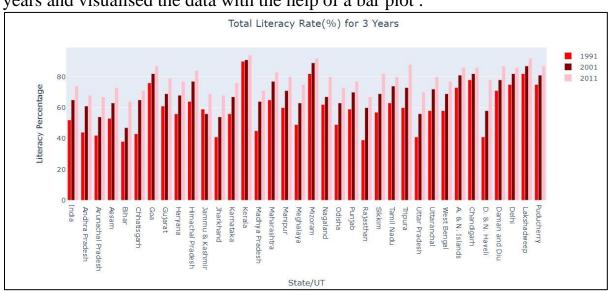
The data is 100% unique and distinct with no missing values.

The data is highly correlated.

#### **DATA VISUALIZATION**

## **Total Literacy rate across Nation for 1991,2001 and 2011**

The dataset consisted the data of 3 years i.e., 1991,2001 and 2011 for each state/UT. We analysed the data on the basis of the total literacy rate % for the 3 years and visualised the data with the help of a bar plot.



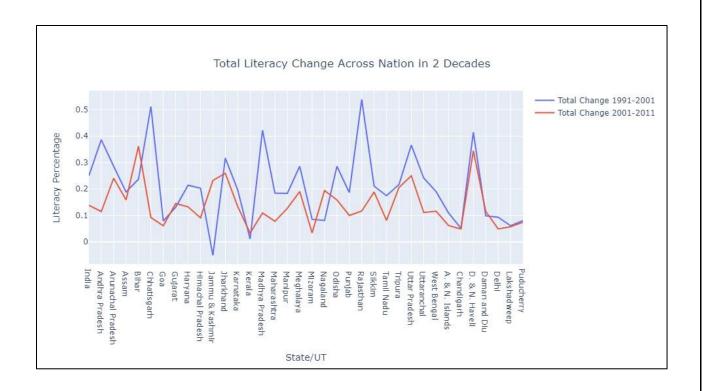
- Literacy percentage has increased over time across the nation
- Kerela has highest literacy among all three years where as Bihar has the lowest literacy rate.
- For the 3 years all the states/UTs had an increase in their literacy rate but the major changes can be seen in Chhattisgarh, Meghalaya, Rajasthan, D&N Haveli and Uttar Pradesh.

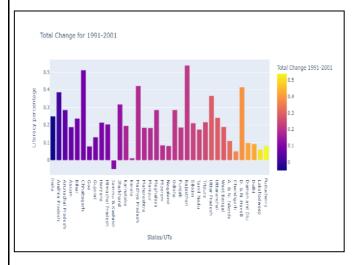
## **Changes in the Literacy Rate for 2 Decades**

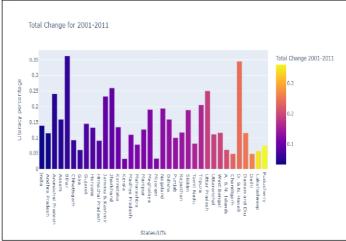
To analyse the changes between the two decades we performed the calculations to get the percentage of change that took place between 1991-2001 and 20012011. Since our data also consisted of the urban-rural distribution hence we also calculated the changes for urban-rural for 2011.



As we have the data of 2 decades so on the basis of the analysis between the decades comparison was done among the changes in the literacy rate for both the decades i.e., 1991-2001 and 2001-2011.

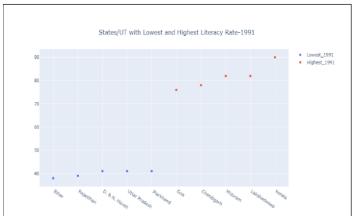






- Major fluctuations can be seen for the decade 1991-2001 whereas some slight changes in 2001-2011.
- Andhra Pradesh had increase in the literacy rate for 19-01 but it declined for the year 01-11 whereas Arunachal Pradesh had a decline in the literacy for 01-11.
- Bihar had decline in literacy rate for 91-01 but the literacy rate took a toll in 01-11 with an increase 13.17%.
- States such as Assam, Goa, Mizoram, Punjab have a declined literacy rate for the 2 decades. Whereas states such as Jharkhand, Madhya Pradesh, Uttar Pradesh and D&N Haveli have shown an increase.
- Jammu Kashmir had negative declination in the 1991-2001 but showed major increase of 74.04 % for 2001-2011 decade.

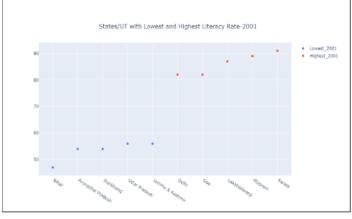
## **Lowest and Highest Literacy Rate States/UT**

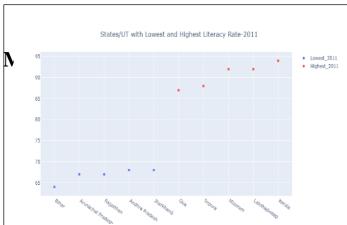


#### 1991

Kerela, Lakshadweep, Mizoram, Chandigarh, Goa have the highest literacy rate.

Bihar, Rajasthan, D&N Haveli, Uttar Pradesh and Jharkhand have the lowest literacy rate.





#### 2001

Mizoram replaced Lakshadweep and became the 2<sup>nd</sup> most literate state as compared to 1991. Delhi came under top five while literacy rate of Chandigarh declined.

Uttar Pradesh had an increase and took over Jharkhand to be the 4<sup>th</sup> lowest literate state where as there was decline in literacy rate of Arunachal Pradesh and J&K

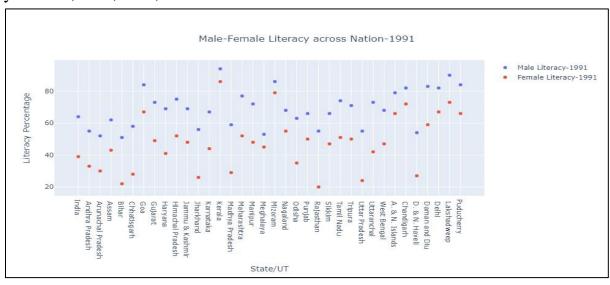
#### 2011

While Kerela remained on top, Tripura entered among the 5 top literate states by taking over Goa and Delhi.

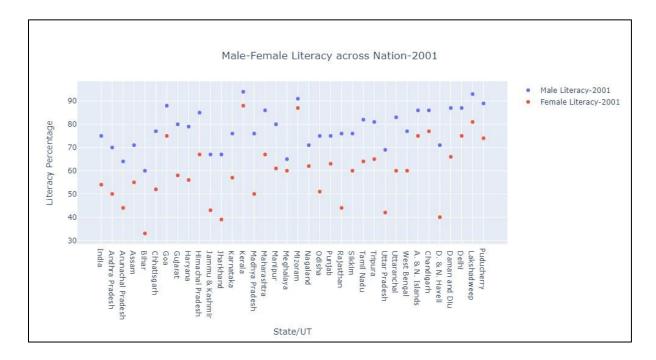
Bihar remained the lowest literate state while Andhra Pradesh and Rajasthan entered the list of lowest 5 literate state by taking over Uttar Pradesh and J&K

## **Distribution of Male-Female Literacy across Nation**

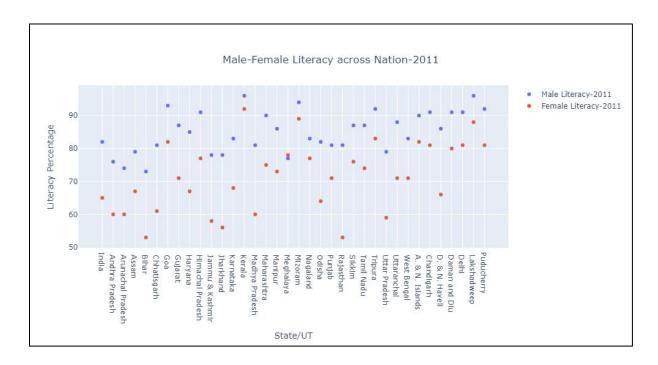
Our data consisted of male-female literacy percentage for all the States/UT's. The data is visualised with the help of scatter plots for all the 3 years i.e., 1991, 2001, and 2011.



- Kerela has the highest male-female literacy while Bihar has the lowest.
- Less female literacy can be seen in areas such as Chhatisgarh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and D&N haveli as compared to their male literacy.
- Goa, Kerela and Mizoram have the highest male-female literacy.

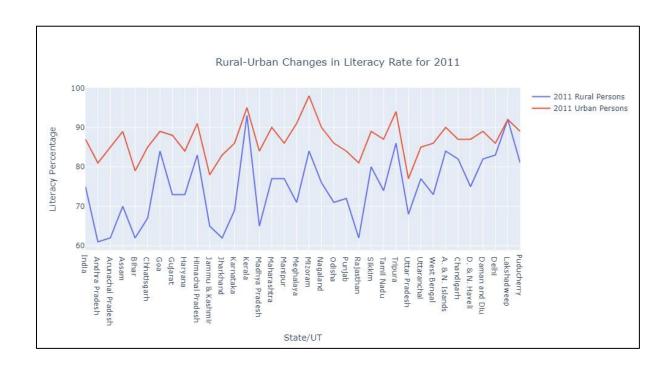


- Female literacy of states such as Rajasthan, Uttar pradesh and Chhatisgarh increased.
- Female literacy of lakshadweep increased and took over Goa to come under the top 3 literate states/UT for male-female literacy.
- In Year 2001 total 13 States/Union Territories had lesser literacy rate than overall Indian literacy rate



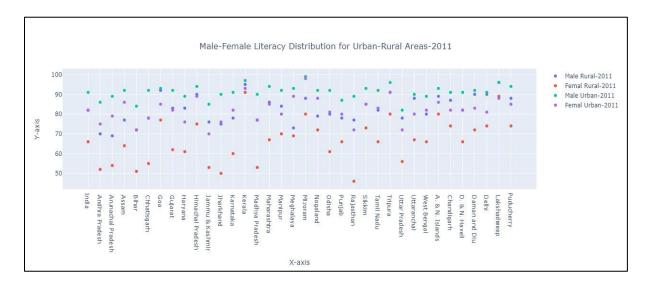
- Most diversification can be seen for the year 2011. States such as Chhattisgarh, Himachal Pradesh, Rajasthan and Tamil Nadu shows increase in the male literacy rate.
- Female literacy of states such as Tripura and Meghalaya increased significantly. While there was decline in the female literacy of Rajasthan.
- In Year 2011 total 11 States/Union Territories had lesser literacy rate than overall Indian literacy rate. Meghalaya and D & N Haveli managed to increase their literacy rate.
- Tripura managed to increase it's literacy rate to 5 most literate states along with Kerala, Lakshadweep, Mizoram and Goa in 2011.
- Mizoram, Kerala, Chandigarh, Delhi and Ponducherry have least percentage increase in literacy rate as it has remained constant

## **Changes in Rural-Urban Literacy for 2011**



- Bihar, Jharkhand, Madhya Pradesh, Meghalaya, Rajasthan, Uttar Pradesh have less rural literacy rate as comapred to urban literacy rate.
- Kerela, Tripura and Lakshdweep have the least difference in their ruralurban literacy rate.
- Assam, Haryana, Chhattisgarh, Sikkim shown an increase in their rural literacy.
- Mizoram, Tripura, Kerela, Lakshadweep and Himachal Pradesh have the highest urban literacy.

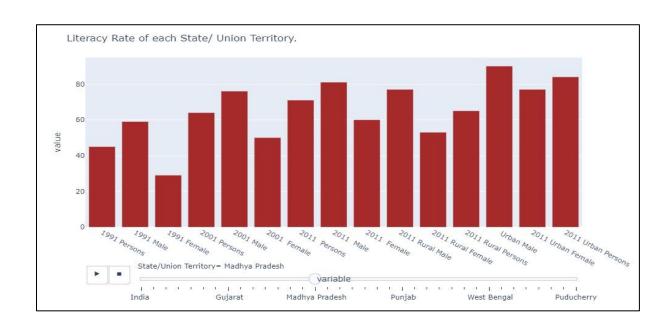
## **Male-Female Distribution for Urban-Rural Areas-2011**

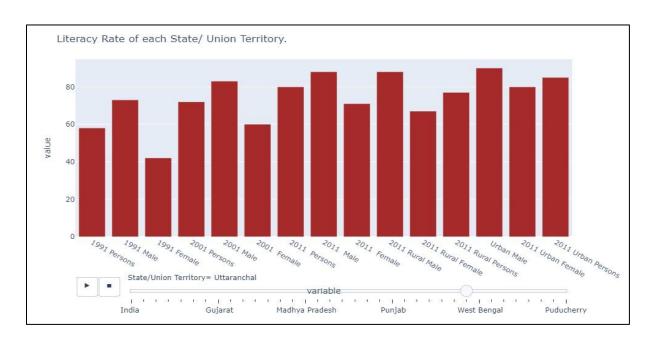


## **Insights:**

- Females in the rural areas are very less literate as compared to the urban females.
- Mizoram has the highest urban male-female literacy while Rajasthan has the least rural female literacy and Andhra Pradesh the least rural male literacy.
- Kerela has the least percentage of change in the urban-rural male female literacy rate.

## **Literacy Rate of Each State/UT**

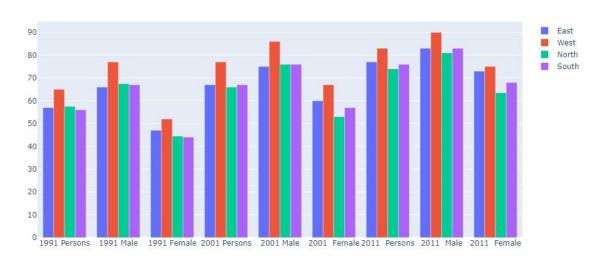


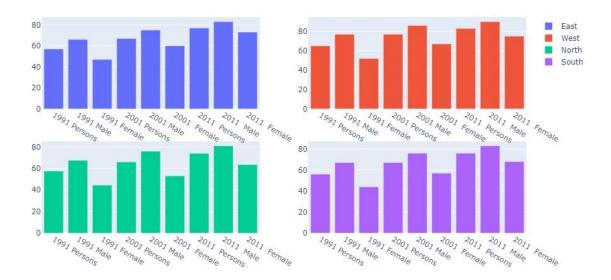


- Since it is a melting graph it is not possible to display the values for all states/UT's.
- Here we have taken Uttaranchal and Madhya Pradesh. Male literacy is
  higher for each year as compared to the female and total literacy. Female
  literacy has taken significant increase for the decades.

## **Average Literacy Distribution by Zones for 2 Decades**







- West Zone has highest average literacy rate in all sectors(for all 3 years).
- North zone has least average literacy rate in all sectors.
- Average literacy of East zone has significantly increased over the years.
- Male-female literacy increased for all the zones over the years.

### **RESULT**

Northern states such as Bihar, Jharkhand, Uttar Pradesh etc have the lowest literacy rates in all sectors thus policies must be made to uplift such states.

Even after the advancement in the urban-rural literacy among the years, still there is much gap between the urban literacy rate and rural literacy rate which need to be fulfilled.

Many States/UT improved their male-female literacy rate over the years but the female literacy rate is still very less as compared to the male literacy rate. States such as Himachal Pradesh and Rajasthan have shown advancement.

Rural-female literacy is also very less as compared to the rural-male literacy thus advancements can be made.

#### CONCLUSION

In conclusion, the literacy rate in India has improved significantly over the years but still faces many challenges in achieving complete literacy for all. Despite the efforts of the government and various non-government organizations, the literacy rate in rural areas is still low compared to urban areas and also female literacy is less than the male literacy. Factors such as poverty, lack of educational infrastructure, and cultural beliefs contribute to the low literacy rate.

It is imperative that the government and society continue their efforts to address these issues and provide equal opportunities for education to all, especially for women and underprivileged communities. The implementation of programs aimed at improving the quality of education and making it accessible to all, regardless of their socio-economic background, can help India achieve its goal of complete literacy.

In conclusion, the literacy rate in India remains a crucial aspect that needs constant attention and efforts to improve. With sustained efforts, India can ensure a bright future for its citizens through education and empowerment.