
# EXPLORATORY ANALYSIS: EDUCATION IN INDIA

**Data Analytics**

**A PROJECT REPORT**

***Submitted by***

**Rishabh Singh Bais 18BCD7009**

**Under the Guidance of**

**Dr. Awnish Kumar**

**Associate Professor, CSE,**

**VIT-AP**

**TABLE OF CONTENTS**

# Chapter -1

## Introduction

Education is one of the most vital part of our life. Today, India is the second largest higher education network in the world. Universities in India are set up by the Central or the State Governments by the means of legislation, while colleges are established by either State Governments or private bodies / trusts. All colleges are affiliated to some university. But the system of education in India is not up to the mark. 25% of the Indian population is illiterate. Only 7% of the population that goes to school managed to graduate and only 15% of those who enroll manage to make it to high school and achieve a place in the higher education system.

Immediately after independence from British rule in 1947, the Department of Education was set up under the Ministry of Human Resource Development (MHRD), with a mandate to increase both access to education and quality, leading to the first National Policy on Education in 1968. Initial expansion of the education sector was limited by India's economic growth but continued steadily until the end of the 20th century. Since committing to the Millennium Development Goals in 2000, India has made great progress towards achieving universal primary education. The World Bank reports that between 2000 and 2017, elementary school enrolment increased by more than 33 million: from 156.6 million in 2000–01 to 189.9 million in 2017–18. 2 While achievement varies greatly between India's 29 states and seven union territories, two-thirds of these have claimed to have achieved universal primary enrolment.

# Chapter -2

## Background Study

This is inspired from the study published by British council in July 2019 by Jason Anderson and Amy Lightfoot, their study further covers the main government initiatives since independence and also provides a comparative study of the major Indian national boards of school education with global ones such as the International Baccalaureate and the India has made phenomenal progress since independence in the field of education. 2015 target of universal primary education for all children aged 6-10 years in 2007-08. The present education system in India is guided by different objectives and goals but is based around the policies of yesteryears. Immediately after independence, a Department of Education under the Ministry of Human Resource Development was set up on August 29, 1947 with a mandate to expand the educational facilities. Policy on Education was formulated in 1968. Over subsequent years, several policies have been formulated by the Indian government to ensure that the literacy level is gradually increased with a close monitoring of the quality of education as well. Retention of children in schools was of paramount importance in the years that followed. With several educational reforms, school drop-out rates have registered a decline with the gender gap of education also showing a dipping figure2. More recently, two prominent policies of the growth of enrollment in secondary education accelerated from 4.3 percent per year during the 1990s to 6.27 percent per year in the decade ending 2009–10. Education continues to remain a top priority for the Government of India with rising budgetary allocations.

# Chapter -3

## Problem Definition

To do the exploratory analysis on education in India and various factors affecting literacy rate in India, gender literacy rate difference, what makes a state best in education and bottom in education in India. Things could be done to improve the female literacy rate and literacy rate in rural areas. Total no. of schools in the country. Comparing primary, secondary and hr secondary education in each state. Types of school we have in our country. Drop out ratio in primary and secondary education and many more inferences that can be made from the dataset.

# Chapter -4

## Objective

The main objective of this project is to collect data of Education in India . The system of education in India is not up to the mark. 25% of the Indian population is illiterate. In this project we are going to collect the state wise data of their secondary education level. Showing the data literacy rate and growth rate of each state and finding their relation. It will help in analyzing the number of schools and male-female literacy rate in each state. This project will help in understanding the causes of low literacy rate in different regions of the country. Total no. of schools in the country. Comparing primary, secondary and hr secondary education in each state. Types of school we have in our country. Finding the differences in top and bottom states on literacy rate. Our ultimate goal is to find causes and resolve them to improve education standards in India.

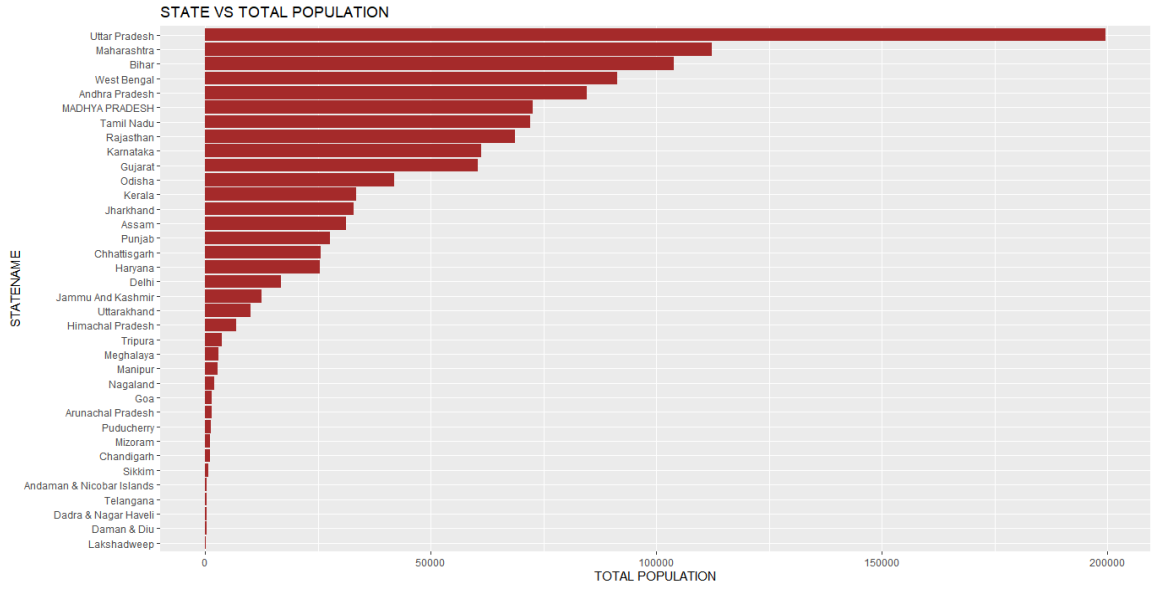
# Chapter -5

## Methodology/Procedure

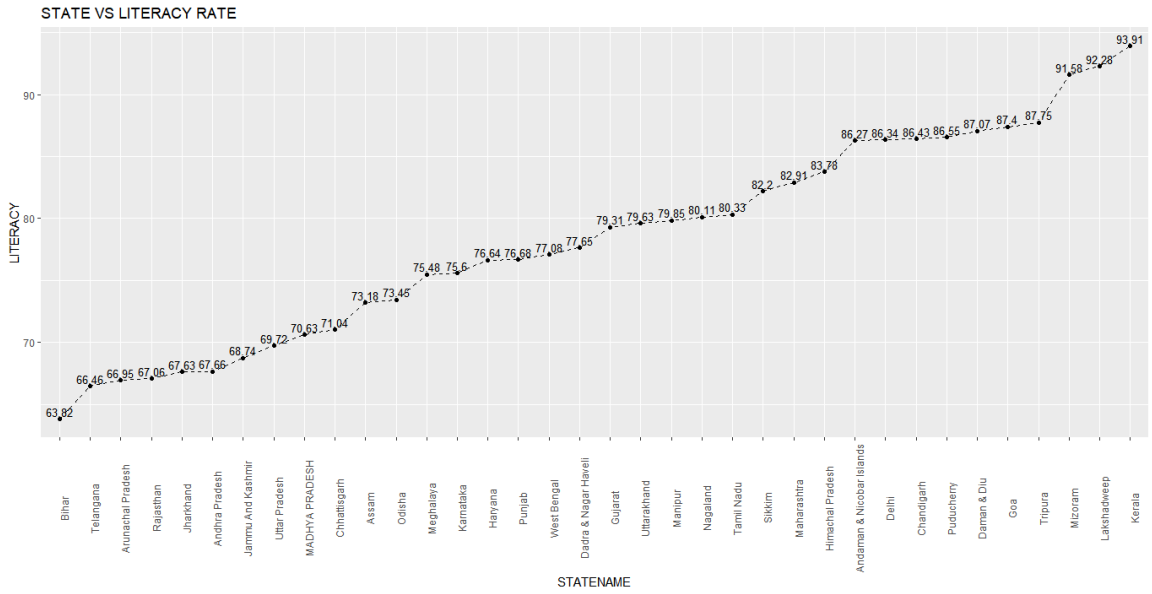
In this analysis we will collect data sets related to education in India mainly primary and secondary data and try to merge them to know what factors are affecting literacy rates in different parts of the country. Here we will analyze the data using R programming and its statistical libraries and visualization libraries to create the statistical data of the literacy rate and the place where literacy rate is quite low to find out it's causes.

# Chapter -6

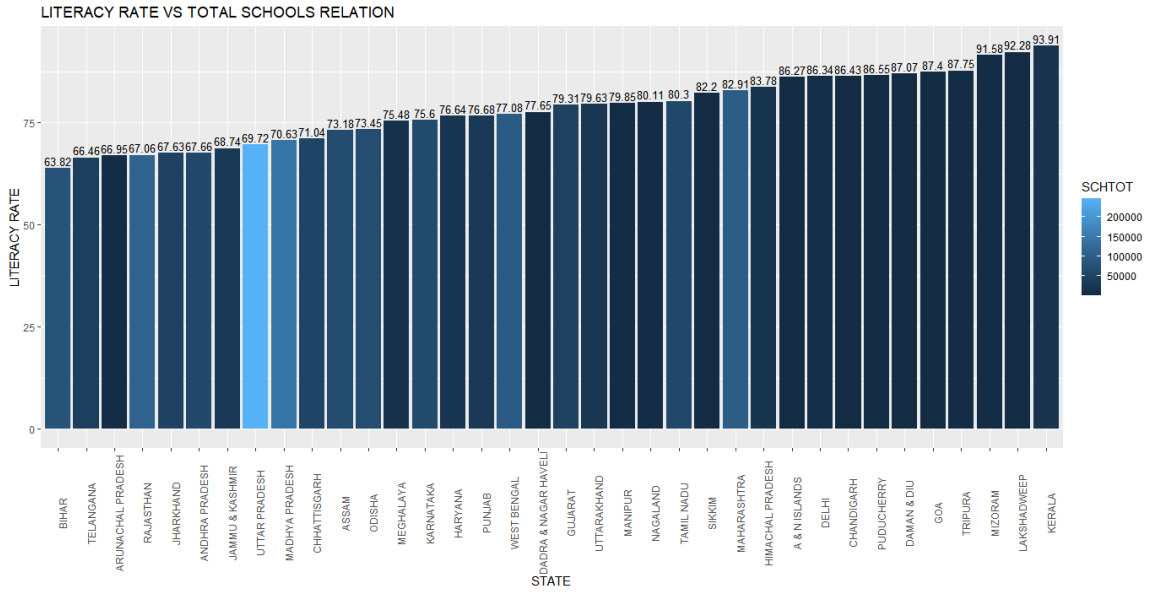
## Results and Discussion



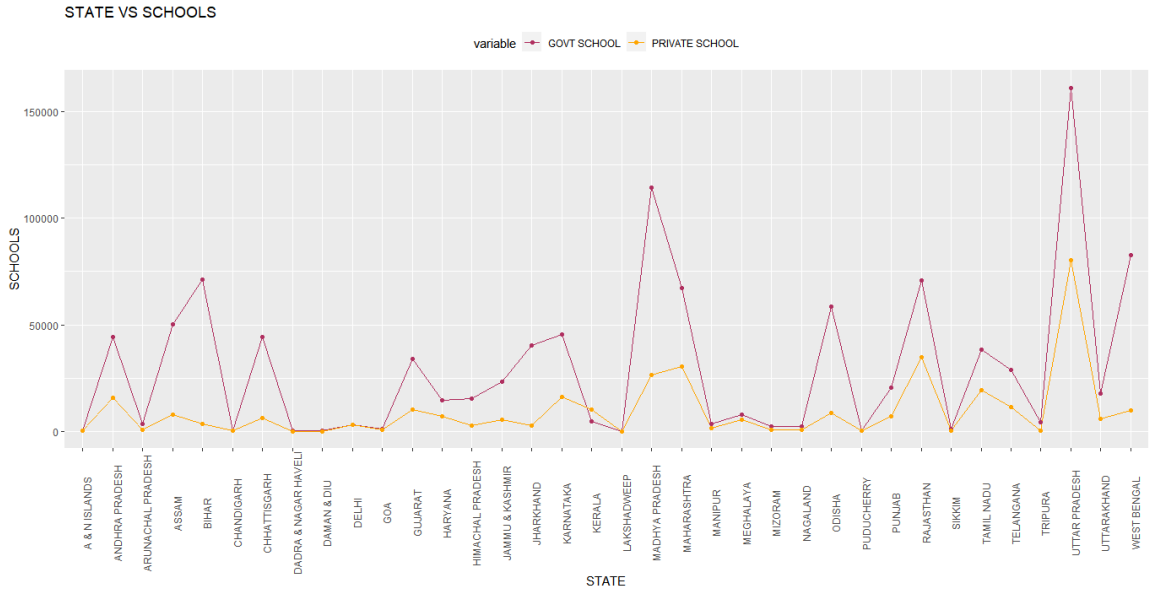

This state vs total population provides an overview as to where we can focus our resources.



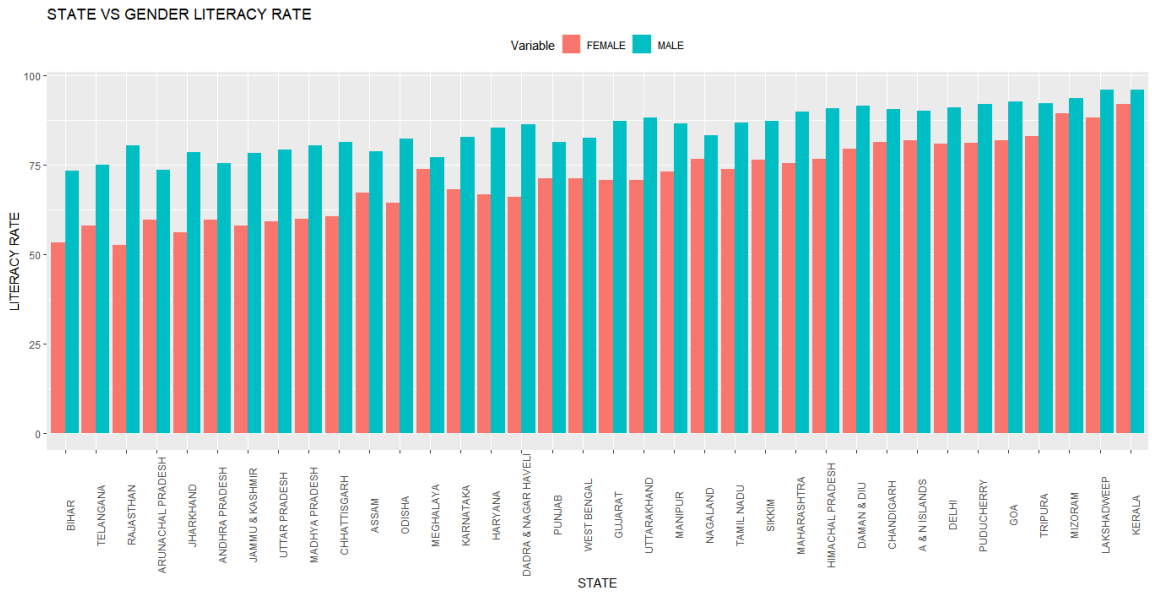

We can see almost all in India states have more than 50% literacy rate




We see a relation i.e states with high literacy rate have more no. of schools. As it is obvious, the more no. of school will be there more people will educate and higher will be the literacy rate




No. of government and private school comparison in each state. Lakshwadeep has zero private schools.




Male and Female literacy rate comparison

We made a new column showing the differences in literacy rates between men and women.

This will help us see the difference in male and female literacy rate for each state easily.

National average difference in literacy rates is 13.89361

States with the least male and female literacy rates difference is

| | |
|---|---|
| MEGHALAYA | 3.39 |
| KERALA | 4.04 |
| MIZORAM | 4.32 |
| NAGALAND | 6.60 |
| LAKSHADWEEP | 7.86 |

States with the most male and female literacy rates difference

| | |
|---|---|
| RAJASTHAN | 27.85 |
| JHARKHAND | 22.24 |
| CHHATTISGARH | 20.86 |
| DADRA & NAGAR HAVELI | 20.53 |
| MADHYA PRADESH | 20.51 |

We check out how the North-East Indian states perform compared to the National Average

The avg in diff in literacy rate for north-eastern states (9.1475) is much less than the national avg (13.89361111111111).

The states with the highest overall literacy rates are: Kerala, Lakshwadeep, Mizoram, Tripura and Goa.

The states with the lowest overall literacy rates are: Bihar, Telngana, Arunachal Pradesh, Rajasthan and Jharkhand.

Dadra and Nagar Haveli, Rajasthan, Chattisgarh, Jharkhand and MP have the maximum difference between Male and Female literacy rates.

Though the top states in male and female literacy remain the same but when we see the bottom states, they change. Meghalaya has the 5th highest no. of illiterate males. The avg female literacy rate for Meghalaya though is higher than the avg female literacy rate.

Also, the states with the highest overall literacy rates have really little differences between male and female literacy rates in general. We also see that states in North east have on an average lesser difference between male and female literacy rates when compared with the average of the country.

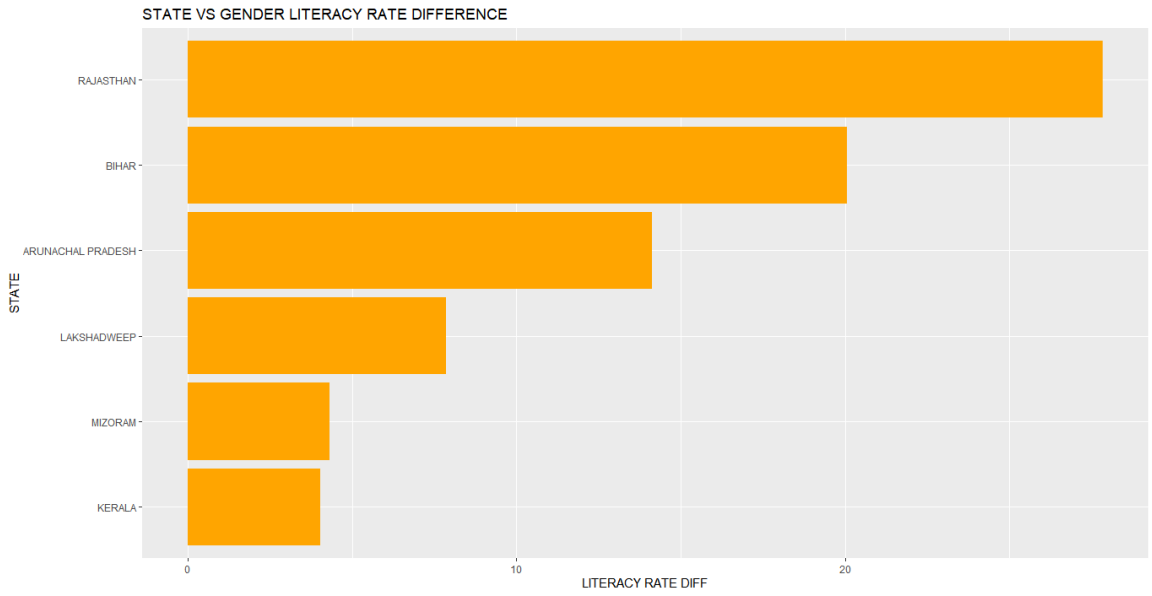
We will be creating top_bottom, a dataframe that contains only the top 3 and bottom 3 states w.r.t. overal literacy rate. This will make plotting and analysis easy for us as we just have to explore these 6 states

Finding the population density

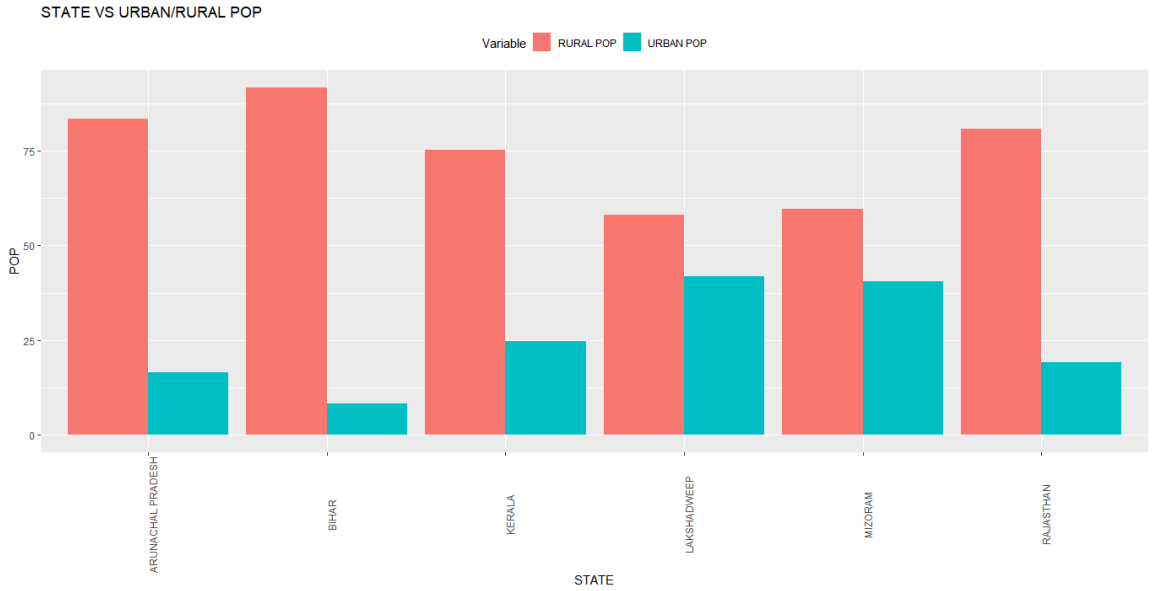
| | |
|---|---|
| KERALA | 859.12050 |
| LAKSHADWEEP | 2000.00000 |
| MIZORAM | 51.75276 |
| RAJASTHAN | 200.50608 |
| ARUNACHAL PRADESH | 16.51481 |
| BIHAR | 1102.39691 |

So population density isn't a factor at all in differentiating between the above two groups. At first we thought that maybe having higher population density created more pressure on the govt to cater to the needs of the people and that it might be difficult for states of really high population density to do so. But the govt of that state should be well equipped enough to

make sure that people get access to education in that state irrelevant of how dense the population is.

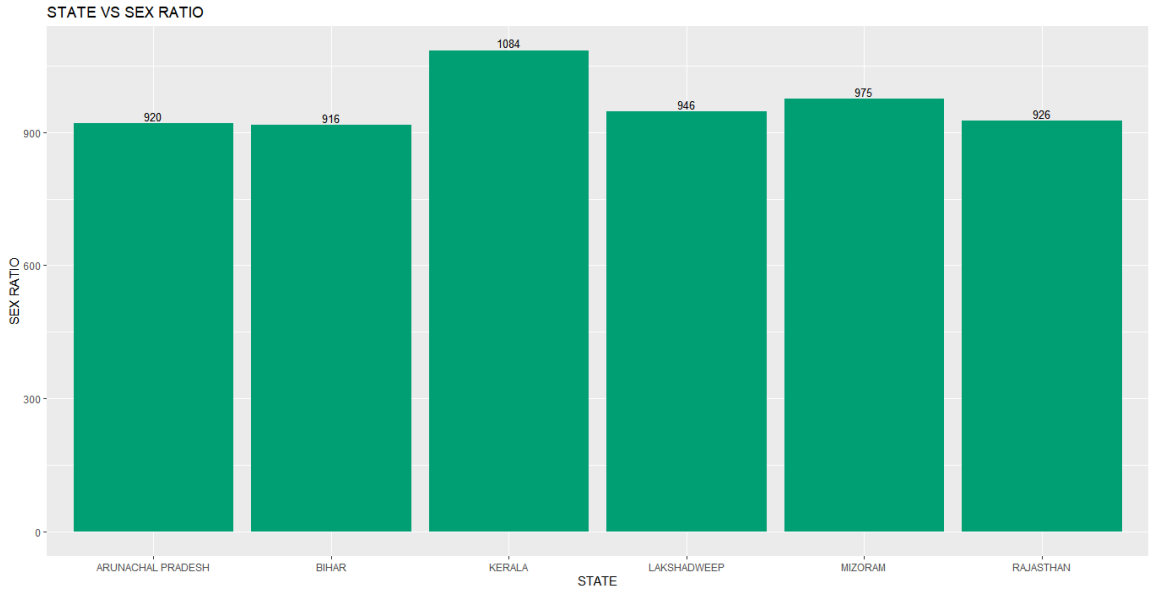



The differences are really high in states with low overall literacy rates. So even if the bottom most states have good male literacy rates, female literacy rates are really low and that takes their overall literacy rate down. Thus these states really need to work on educating their females and increasing their literacy rate.



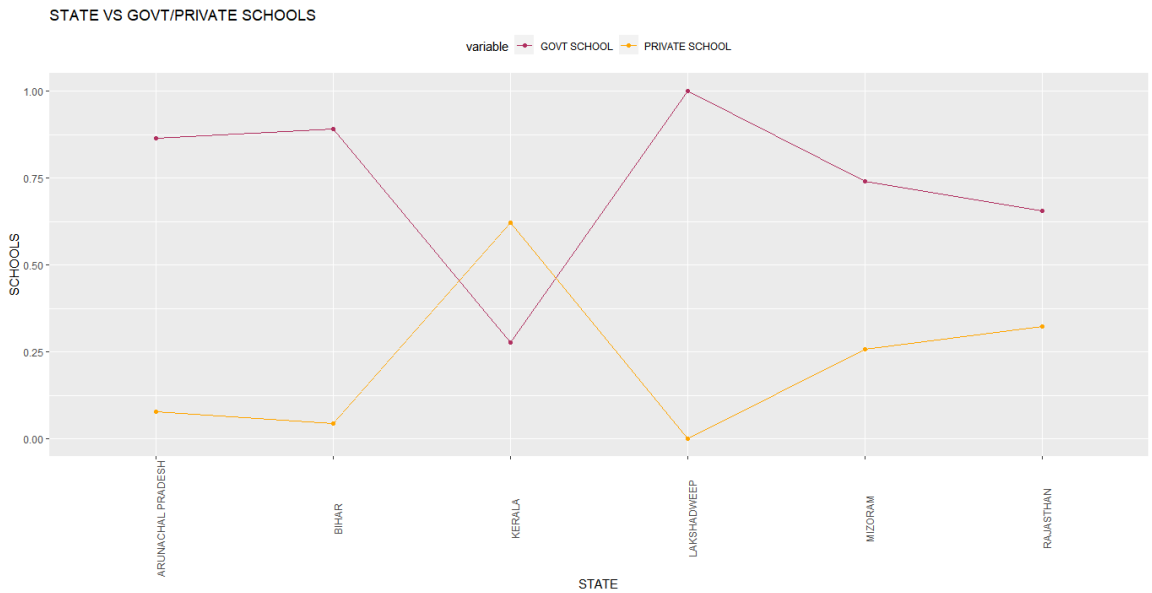

We have created rural population which is 100 – urban population proportion. This will help us in comparing the rural and urban population proportions. Here we see a pretty obvious but important component that differentiates our top 3 and bottom 3 states. The difference between rural and urban population is much much bigger in the bottom 3.

The rural population percentage in the bottom 3 states is much more than the rural population percentage in the top 3. That's an important factor to note. People living in rural areas lead a very different life compared to the people living in urban areas. There is less motivation to go to school in rural areas as a lot of people tend to take up their parents profession or business. Also, often children are made to skip school and work at the farm.




Clearly the sex ratio has nothing to do with the literacy level. The graphs don't show sex ratio affecting the literacy rate.

Now we will try to analyze some school related features and see if they affect the overall literacy rate.



We can also take a look at the different types of schools in these states. From here we can clearly note the following:

Lakshadweep has only govt schools.

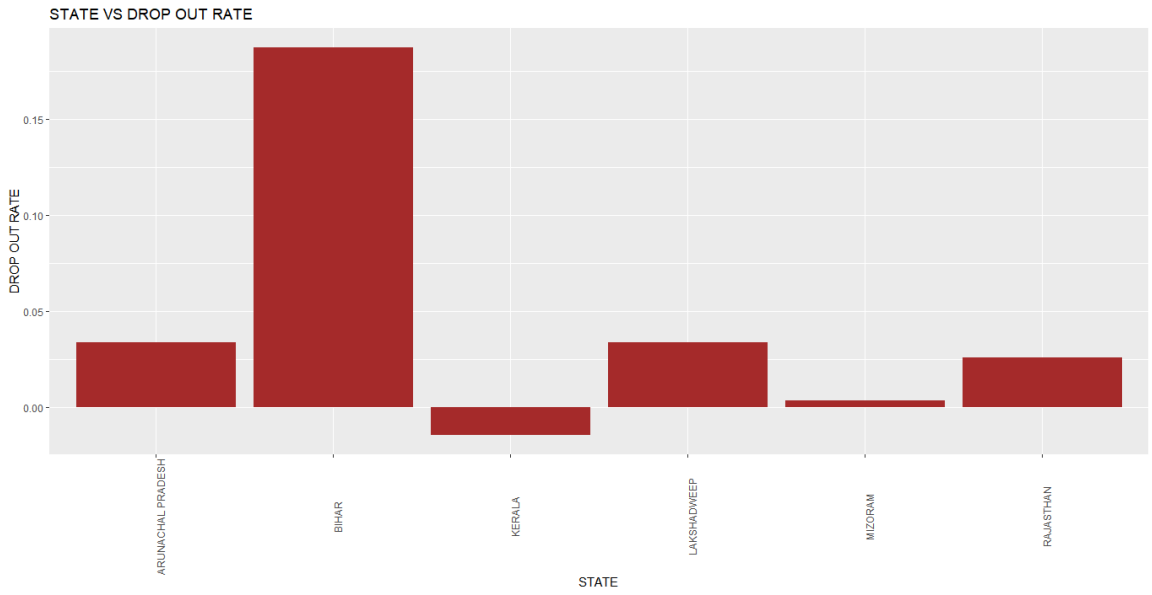
Kerala has the highest number of Pvt schools share in the total schools and is the only state here which has more Pvt schools than govt schools.

Bihar and Arunachal Pradesh have really less pvt schools compared to the govt schools. Their % of govt school is more than the national avg.

Rajasthan has around 35% pvt schools which is large compared to its no. of govt schools.

All of this just shows us that even the share of pvt and govt schools isn't related to the literacy rate.

Dropout rate between primary to secondary education



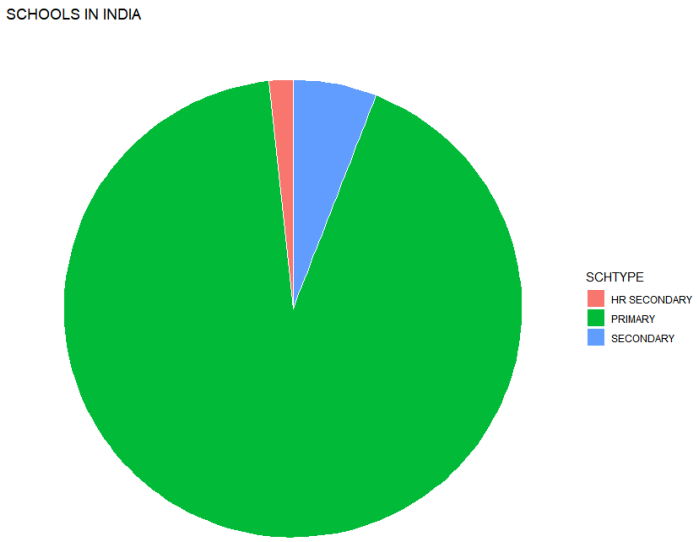

We find the drop out rate from 5th to 6th(primary to secondary) a pretty important feature in distinguishing between the top 3 and bottom 3 states. It is an obvious reason but no one knew that there would be this much of a difference. The top 3 states have more admissions in 6th than they had in 5th than the bottom 3 states.

States should try to increase their female literacy rate by making it safer for girls to travel, making girl only schools with female teachers, providing special incentives to the families to get their girl child to complete her schooling etc.

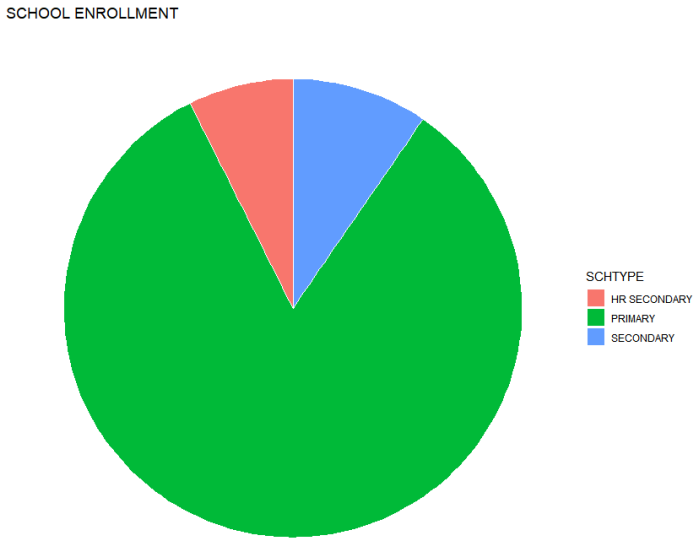
People living in rural areas lead a very different life compared to the people living in urban areas. There is less motivation to go to school in rural areas as a lot of people tend to take up

their parent's profession or business. Also, often children are made to skip school and work at the farm. The state should provide more incentives to ensure that families living in rural areas get their children to complete their schooling.

Dropout rate from 5th to 6th class should be decreased and schemes should be made to ensure that this dropout rate declines. Causes for this dropout rate should be looked into and be dealt with.

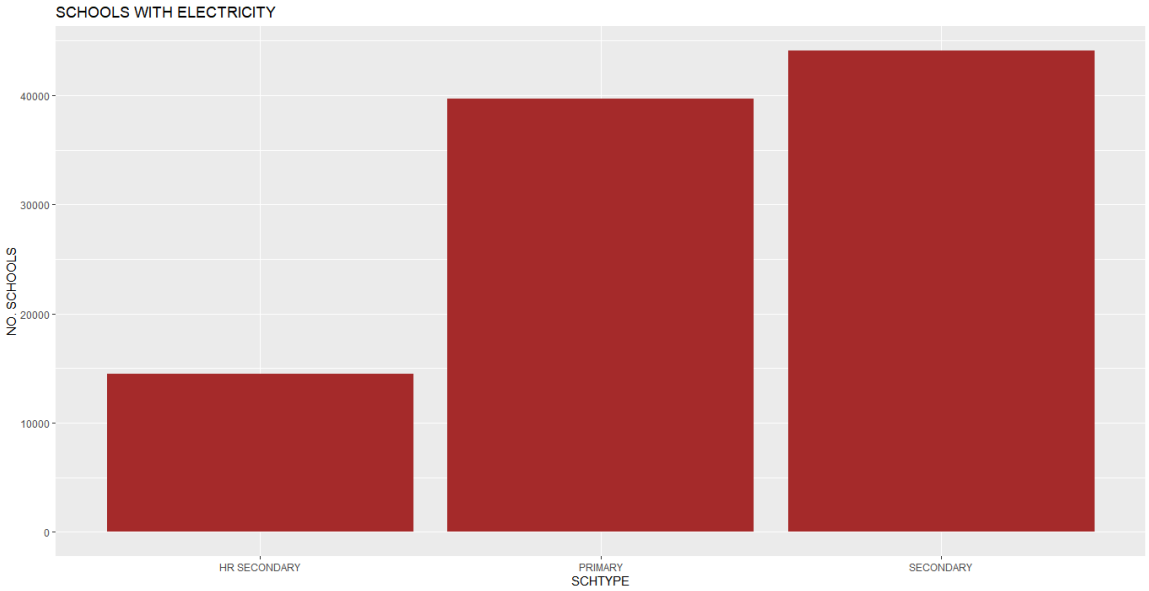



If we compare the total no. of primary, secondary and hr secondary schools in India, primary are most in number and hr secondary are least.




We found out that primary schools have highest no. of enrollment then secondary and hr secondary have least. It is due to two main reason because no. of primary schools is huge and

second is many kids drop out after primary as their family want them work in agriculture field or business.




We can see hr secondary has least, secondary has max no. of schools with electricity and primary is moderate.

# Chapter -7

## Conclusion and Future Scope

We started out with a comparison of the literacy rates in the different states and union territories in India. We compared the Male and Female literacy rates in these states and saw that there was a huge difference in the states with a low overall literacy rate. The top 3 states were Kerala, Lakshadweep and Mizoram and the bottom 3 were Bihar, Telngana and Arunachal Pradesh.

Next we saw that difference in male and female literacy rates, rural population proportion and dropout rates from 8th to 9th class played a huge role in separating the top 3 and bottom 3 states.

Finally, the dropout rates in 5th and 6th classes were explored and while the dropout rate for 6th class was really high, more students had enrolled in class 5th than had dropped out.

The literacy rate in India has been improving but there are some key issues that need to be tackled aggressively in order to improve the state of education in India. This is not just the job of the government, but it is the duty of each and everyone living in the country. Hope to see the stats showing a much better India in the future.

## References


1. https://towardsdatascience.com/exploratory-data-analysis-in-r-for-beginners-fe031add7072
2. https://cran.r-project.org/web/packages/dlookr/vignettes/EDA.html
3. https://blog.datascienceheroes.com/exploratory-data-analysis-in-r-intro/
4. https://www.geeksforgeeks.org/exploratory-data-analysis-in-r-programming/
5. http://www.sthda.com/english/wiki/ggplot2-essentials

# Appendix – A

## Team Work and Work Management

Transparent working environments have been found to make teams more accountable, happy and creative. Transparent environments help to develop a feeling of mutual respect between team members and team leaders. Via open and consistent communication, transparent and authentic workplaces help to feel secure in their positions. In turn, team members feel freer to contribute ideas and suggestions, enhancing creativity. We know that the basis of a cooperative and productive team is good communication.

The aim is to create an environment in which team leaders feel able to provide honest and constructive feedback, and team members feel confident to voice concerns and communicate with one another. For teams with members working remotely, meeting platforms like teams and google meet can provide an ideal way to ensure some face-to-face time is achieved. Providing feedback to each other is one of the best ways you can support them to develop professionally and personally.

Even if you have no negative feedback to give, make sure to hold regular opportunities to check-in. This way, you can provide advice on how you feel your team members are progressing and could grow further. If there are any areas of work that you feel could be improved, these discussions also provide a good opportunity to share your constructive feedback.

Inevitably, your team members will be happier if they can get along well with one another. As an added bonus, they'll perform better too. To achieve this, encourage your team members to collaborate. On your team, there will likely be a whole bunch of diverse skills. Make sure these different skill sets are utilized by ensuring everyone is aware of ongoing projects. That way, team members can jump in to collaborate wherever they feel they can bring value.

Know exactly how you are spending your time. You need to know which tasks are stealing your time. Then you can do something about it. Setting a time limit for a task can be fun. In fact, it can be like a game. Some companies actually divide employees into groups, and the group that completes a project or task first receives a reward. Indeed, a list allows you to stay focused and motivated, focused on feeling that sweet satisfaction every time you check off a task off your list. Lists also allow you to see and track your progress.

Even if you're surrounded by distractions, your list will keep you on track. Planning ahead is an essential part of time management. Ideally, you should plan ahead for the week or at least the day before. When you know exactly what needs to be done for the day or the week, you stay organized and focused. You can spread tasks over several days to see in advance how much time is needed to complete a project. Avoid doing half the work, which means giving up your current task and doing something else. An example of a half-job is writing a report, then suddenly checking your email for no reason and writing responses. This is not only poor time management but also poor concentration.

# Appendix – B

## Coding and Snap Shot

```
df_sec <- read.csv("2015_16_Statewise_Secondary.csv")
df_sec
View(df_sec)
names(df_sec)

df_ele <- read.csv("2015_16_Statewise_Elementary.csv")
df_ele
View(df_ele)
names(df_ele)

sum(is.na(df_sec))
sum(is.na(df_ele))

sum(duplicated(df_sec))
sum(duplicated(df_ele))

length(df_ele[df_ele==0])
length(df_sec[df_sec==0])

df_sec[df_sec == 0] <- NA
for(i in 1:ncol(df_sec)){
  df_sec[is.na(df_sec[,i]), i] <- mean(df_sec[,i], na.rm = TRUE)
}

df_ele[df_ele == 0] <- NA
```

```
for(i in 1:ncol(df_ele)){
  df_ele[is.na(df_ele[,i]), i] <- mean(df_ele[,i], na.rm = TRUE)
}

df_ele$TOTPOPULAT[19]
df_ele$TOTPOPULAT[19]<-991348/10
#outlier removal

library("ggplot2")
library(dplyr)
library(tidyr)

plot1 <- ggplot(df_sec, aes(x=tot_population, y=reorder(statname, tot_population))) +
  geom_bar(stat = "identity",fill='brown')+  # Y axis is explicit. 'stat=identity'
  ggtitle("STATE VS TOTAL POPULATION") +
  xlab("TOTAL POPULATION") + ylab("STATENAME")
print(plot1)

#The plot provides an overview as to where we can focus our resources.
ggplot(data=df_sec, aes(x=reorder(statname, literacy_rate), y=literacy_rate, group=1)) +
  geom_line(linetype = "dashed")+
  geom_point()+
  theme(axis.text.x = element_text(angle = 90))+
  ggtitle("STATE VS LITERACY RATE") +
  xlab("STATENAME") + ylab("LITERACY")+
  geom_text(aes(label=literacy_rate), vjust=-0.3, size=3.5)

#We can see almost all in India states have more than 50% literacy rate
```

```
ggplot(data=df_ele, aes(x=reorder(STATNAME,OVERALL_LI), y=OVERALL_LI, fill=SCHTOT))
+
  geom_bar(stat="identity")+
  theme(axis.text.x = element_text(angle = 90))+
  ggtitle("LITERACY RATE VS TOTAL SCHOOLS RELATION") +
  xlab("STATE") + ylab("LITERACY RATE")+
  geom_text(aes(label=OVERALL_LI), vjust=-0.3, size=3.5)

#We see a relation i.e states with high literacy rate have more no. of schools
df <- df_ele %>%
  select(STATNAME, SCHTOTG, SCHTOTP) %>%
  gather(key = "variable", value = "value", -STATNAME)

df
ggplot(df, aes(x=STATNAME, y=value, group=variable)) +
  geom_line(aes(color=variable))+
  geom_point(aes(color=variable))+
  theme(legend.position="top",axis.text.x = element_text(angle = 90))+
  scale_color_manual(labels = c("GOVT SCHOOL", "PRIVATE SCHOOL"), values = c("maroon",
"orange"))+
  ggtitle("STATE VS SCHOOLS") +
  xlab("STATE") + ylab("SCHOOLS")
#No. of govt an private school comparison

#We will now see a Male vs Female vs Literacy rate comparison
df <- df_ele %>%
  select(STATNAME, MALE_LIT, FEMALE_LIT) %>%
  gather(key = "variable", value = "value", -STATNAME)
```

```
df


ggplot(data=df, aes(x=reorder(STATNAME, value), y=value, fill=variable)) +
  geom_bar(stat="identity", position=position_dodge())+
  theme(legend.position="top",axis.text.x = element_text(angle = 90))+
  scale_fill_discrete(name = "Variable", labels = c("FEMALE", "MALE"))+
  ggtitle("STATE VS GENDER LITERACY RATE") +
  xlab("STATE") + ylab("LITERACY RATE")


df_ele$diff_lit=df_ele$MALE_LIT-df_ele$FEMALE_LIT
df_ele$diff_lit
mean(df_ele$diff_lit)
# States with the least male and female literacy rates difference
head(df_ele[order(df_ele$diff_lit),c("STATNAME","diff_lit")],5)
# States with the most male and female literacy rates difference
head(df_ele[order(-df_ele$diff_lit),c("STATNAME","diff_lit")],5)
#We now check out how the North-East Indian states perform compared to the National Average


D=list('NAGALAND','MANIPUR','MIZORAM','ASSAM','TRIPURA','ARUNACHAL
PRADESH','MEGHALAYA','SIKKIM')
D
for(i in D){
  z<-df_ele[df_ele$STATNAME==i,"diff_lit"]
  print(z)
}
message("male female literacy diff of north eastern states ",mean(z))
message("male female literacy diff national avg ",mean(df_ele$diff_lit))
```

```
df_ele[order(-df_ele$OVERALL_LI),"STATNAME"]
#remove telangana as it has lack of data
top_bottom=df_ele[order(-df_ele$OVERALL_LI),]
View(top_bottom)
top_bottom=top_bottom[c(1,2,3,33,34,36),]  #removing telangana as it is newly formed
View(top_bottom)
#Now we will go over different features that we think affect our state's overall literacy rate

den=list()
den$state=top_bottom$STATNAME
den$density=(top_bottom$TOTPOPULAT/top_bottom$AREA_SQKM)*1000
den=as.data.frame(den)
den

ggplot(top_bottom, aes(x=diff_lit, y=reorder(STATNAME, diff_lit))) +
  geom_bar(stat = "identity",fill="orange")+
  ggtitle("STATE VS GENDER LITERACY RATE DIFFERENCE") +
  xlab("LITERACY RATE DIFF") + ylab("STATE")

top_bottom$P_PUR_POP=100-top_bottom$P_URB_POP
df <- top_bottom %>%
  select(STATNAME, P_URB_POP, P_PUR_POP) %>%
  gather(key = "variable", value = "value", -STATNAME)

df

ggplot(data=df, aes(x=reorder(STATNAME, value), y=value, fill=variable)) +
```

```
  geom_bar(stat="identity", position=position_dodge())+
  theme(legend.position="top",axis.text.x = element_text(angle = 90))+
  scale_fill_discrete(name = "Variable", labels = c("RURAL POP", "URBAN POP"))+
  ggtitle("STATE VS URBAN/RURAL POP") +
  xlab("STATE") + ylab("POP")
#Here we see a pretty obvious but important component that differentiates our top 3 and bottom 3 states. The difference between rural and urban population is much much bigger in the bottom 3.


#Sex ratio is the no. of females per 1000 males
ggplot(top_bottom, aes(x=STATNAME, y=SEXRATIO)) +
  geom_bar(stat = "identity",fill="#009E73")+
  geom_text(aes(label=SEXRATIO), vjust=-0.3, size=3.5)+
  ggtitle("STATE VS SEX RATIO") +
  xlab("STATE") + ylab("SEX RATIO")


cor(top_bottom$OVERALL_LI, top_bottom$SEXRATIO)
#Clearly the sex ratio has nothing to do with the literacy level.The graphs dont show sex ratio affecting the literacy rate. Also the correlation's too weak.


ps<-top_bottom["STATNAME"]
ps$schtotpp<-top_bottom$SCHTOTP/top_bottom$SCHTOT
ps$schtotgp<-top_bottom$SCHTOTG/top_bottom$SCHTOT
ps[,]
ps$schtotpp[5]<-0.0787  #outlier removal


df <- ps %>%
  select(STATNAME, schtotgp, schtotpp) %>%
  gather(key = "variable", value = "value", -STATNAME)


df
```

```
ggplot(df, aes(x=STATNAME, y=value, group=variable)) +
  geom_line(aes(color=variable))+
  geom_point(aes(color=variable))+
  theme(legend.position="top",axis.text.x = element_text(angle = 90))+
  scale_color_manual(labels = c("GOVT SCHOOL", "PRIVATE SCHOOL"), values = c("maroon",
"orange"))+
  ggtitle("STATE VS GOVT/PRIVATE SCHOOLS") +
  xlab("STATE") + ylab("SCHOOLS")

tbd<-top_bottom["STATNAME"]
a<-top_bottom$C5_B+top_bottom$C5_G
b<-top_bottom$C6_B+top_bottom$C6_G
tbd$drop<-(a-b)/a
tbd

ggplot(tbd, aes(x=STATNAME, y=drop)) +
  geom_bar(stat = "identity",fill="brown")+
  theme(legend.position="top",axis.text.x = element_text(angle = 90))+
  ggtitle("STATE VS DROP OUT RATE")+
  xlab("STATE") + ylab("DROP OUT RATE")



q=sum(df_ele$ENR1)
w=sum(df_sec$enr_5)
e=sum(df_sec$enr_7)

SCHTYPE <- c("PRIMARY", "SECONDARY","HR SECONDARY")
```

```
values = c(q,w,e)
percents <- round((values/sum(values))*100,1)
df<-as.data.frame(SCHTYPE)
df<-cbind(df,percents)

ggplot(df, aes(x = "", y = percents, fill = SCHTYPE)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)+
  theme_void()+
  ggtitle("SCHOOL ENROLLMENT")

#school enrollment

q=sum(df_ele$SCH1)
w=sum(df_sec$sch_5)
e=sum(df_sec$sch_7)

SCHTYPE <- c("PRIMARY", "SECONDARY","HR SECONDARY")
values = c(q,w,e)
percents <- round((values/sum(values))*100,1)

df<-as.data.frame(SCHTYPE)
df<-cbind(df,percents)

ggplot(df, aes(x = "", y = percents, fill = SCHTYPE)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)+
  theme_void()+
  ggtitle("SCHOOLS IN INDIA")
```

```
#type of schhols in india


a<-cbind(df_ele['SELE1'],df_sec[c('statname','electric_5','electric_7')])
z<-colSums(a[,c(1,3,4)])
z
z<-as.data.frame(z)
z$names<-c("PRIMARY","SECONDARY","HR SECONDARY")
z$z[1]<-z$z[1]/10 #outlier removal
ggplot(z, aes(x=names, y=z)) +
  geom_bar(stat = "identity",fill='brown')+  # Y axis is explicit. 'stat=identity'
  ggtitle("SCHOOLS WITH ELECTRICITY") +
  xlab("SCHTYPE") + ylab("NO. SCHOOLS")
#schools wih electricity in india
```