



**PES University, Bangalore**  
(Established under Karnataka Act No. 16 of 2013)

**UE17CS2  
03**

**B. Tech, Sem III  
Session: Aug-Dec, 2018**

## **UE17CS203 – INTRODUCTION TO DATA SCIENCE**

### **REPORT**

#### **EXPLORATORY ANALYSIS ON (DATA SET NAME)**

<b>DATA SET LINK :</b>	<a href="https://www.kaggle.com/rajanand/education-in-india/kernels">https://www.kaggle.com/rajanand/education-in-india/kernels</a>
<b>TEAM MEMBERS</b>	<b>NAME : SAMHITHA DINESH</b> <b>SRN : PES1201701274</b> <b>EMAIL ID : samhithadinesh13@gmail.com</b> <b>CONTACT NO. : 9686357208</b>
	<b>NAME : ASHIKA MERYL PINTO</b> <b>SRN : PES1201701346</b> <b>EMAIL ID : pintoashika@gmail.com</b> <b>CONTACT NO. : 9110281466</b>

#### **ABSTRACT**

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. It is nothing but “taking a peek” at the data to understand more about what it represents and how to apply it. As a part of this assignment we have done the same with the dataset “EDUCATION IN INDIA-2015-16”

which we accessed from the Kaggle website. The dataset gives us an insight of the education development taking place in the country. We started with cleaning the data by replacing the missing values and the invalid data entries. Next, we identified all the important variables that could come to use in the process of analysis. We picked questions that we thought we could answer using the dataset, checked if what we assumed about these questions was indeed true by visualizing the data using different plotting methods like bar graphs, scatter plots, box plots and histograms.

Iterating through all these steps again and again we concluded 8 facts(inferences) about the data that could come of use in improving the current education status of India.

### **Data Set**

The dataset chosen shows us a bird's-eye view of the education system in India. It has estimated values of the data from which we can deduce the education development taking place in the country. From this extensive dataset, it is possible to draw relations between important contributing factors towards literacy such as population, growth rate, additional facilities, number of villages, enrolment of students, number of schools, gender ratio etc.

This is a large dataset containing 680 rows and columns which are used as variables to obtain the required results.

We have collected this dataset from Kaggle and plotted the co relations between different factors which we think have a major role in contributing to the educational development of the country.

**The sole purpose of scrutinising this dataset is to procure sufficient information regarding the literacy of the country and to come up with efficient solutions to improve the same.**

### **Introduction**

The following dataset we've chosen extensively talks about education in India. It has a variety of information ranging from population, area, male-female literacy rates, sex ratio etc which will help us to draw conclusions and analyse the parallels between these factors and the literacy rate.

To begin with, we chose the dataset from Kaggle and went through all the rows and columns to have a thorough groundwork for our analysis. The sole purpose of scrutinising this dataset is to procure sufficient information regarding the literacy rate of the country and to come up with efficient solutions to improve the same. It is possible to question a number of aspects and obtain suitable answers required to have a better insight into the overall educational system of the country.

## **Processing (Data Cleaning)**

Our dataset has missing values in certain columns. Hence we have written the code that replaces them with the medians of the respective columns.

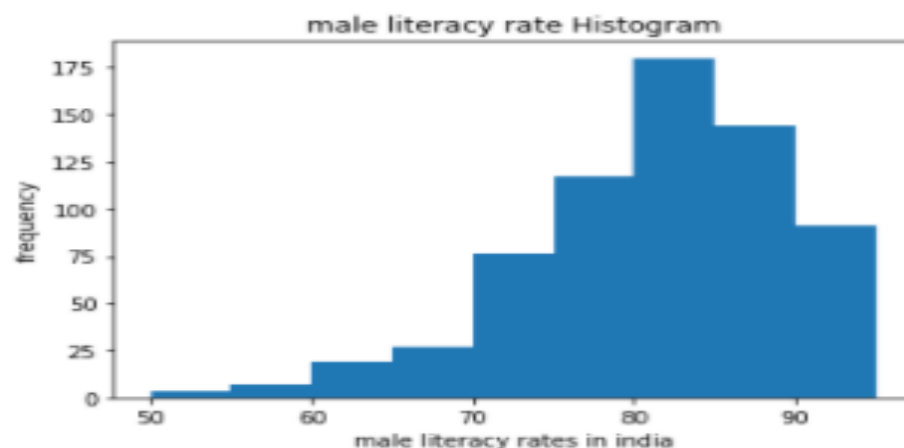
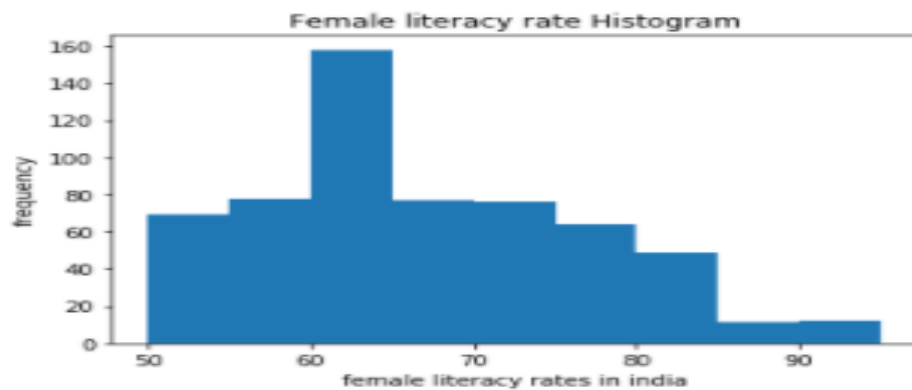
Measures taken:

- 1.>Replacing missing values by the median of the respective columns
- 2.>Replacing invalid entries-> for example: negative values for percentage
- 3.>There are also some outliers in the data set. But we found it best to not replace them as those could be the actual data points, replacing them would affect our data. With the intention to not lose important data, we let the outliers be.

## **EXPLORATORY ANALYSIS**

### **INFERENCES MADE:**

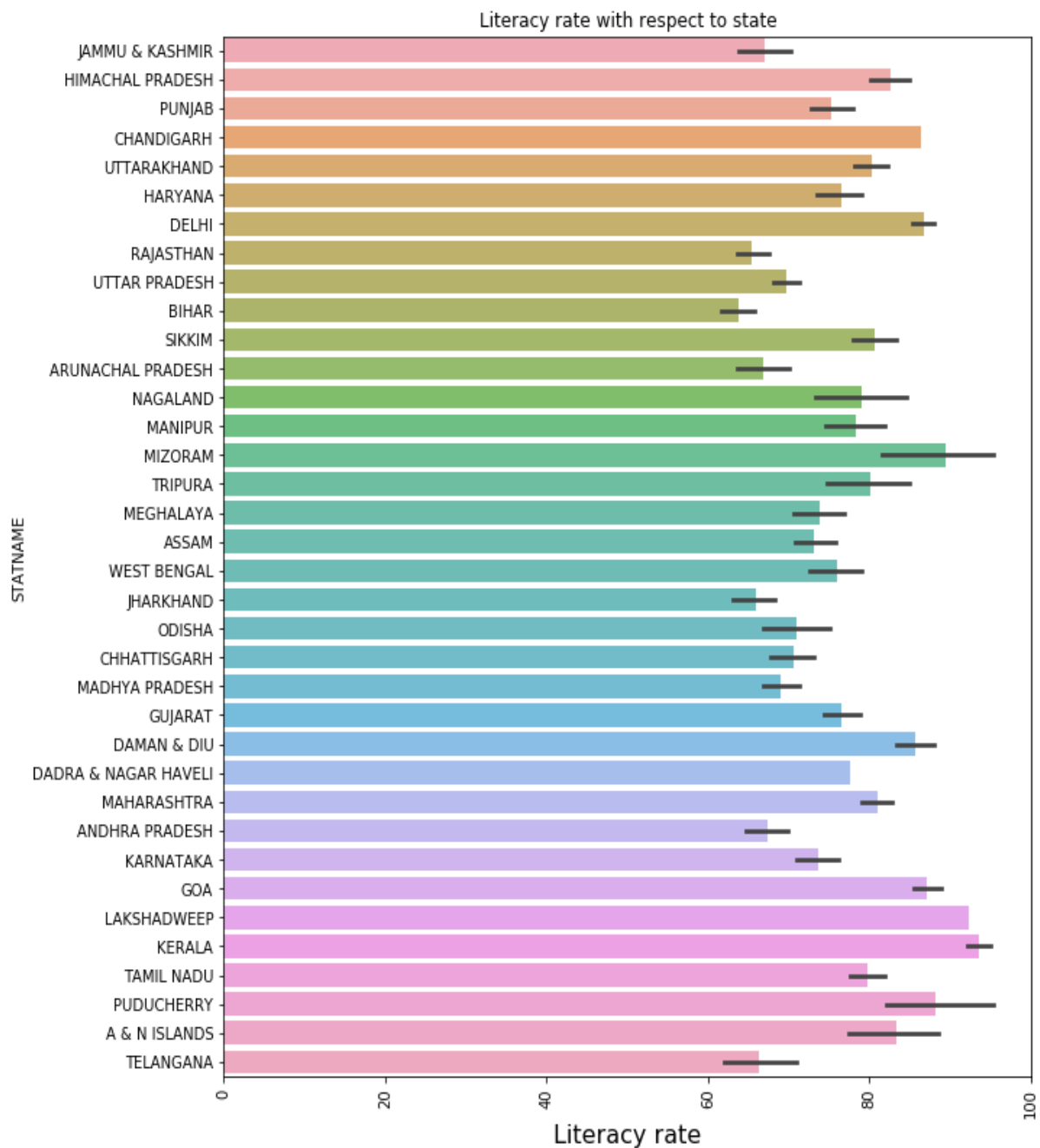
**1.LITERACY AND EDUCATION OF WOMEN IN INDIA-MAJOR DISAPPOINTMENT!!**

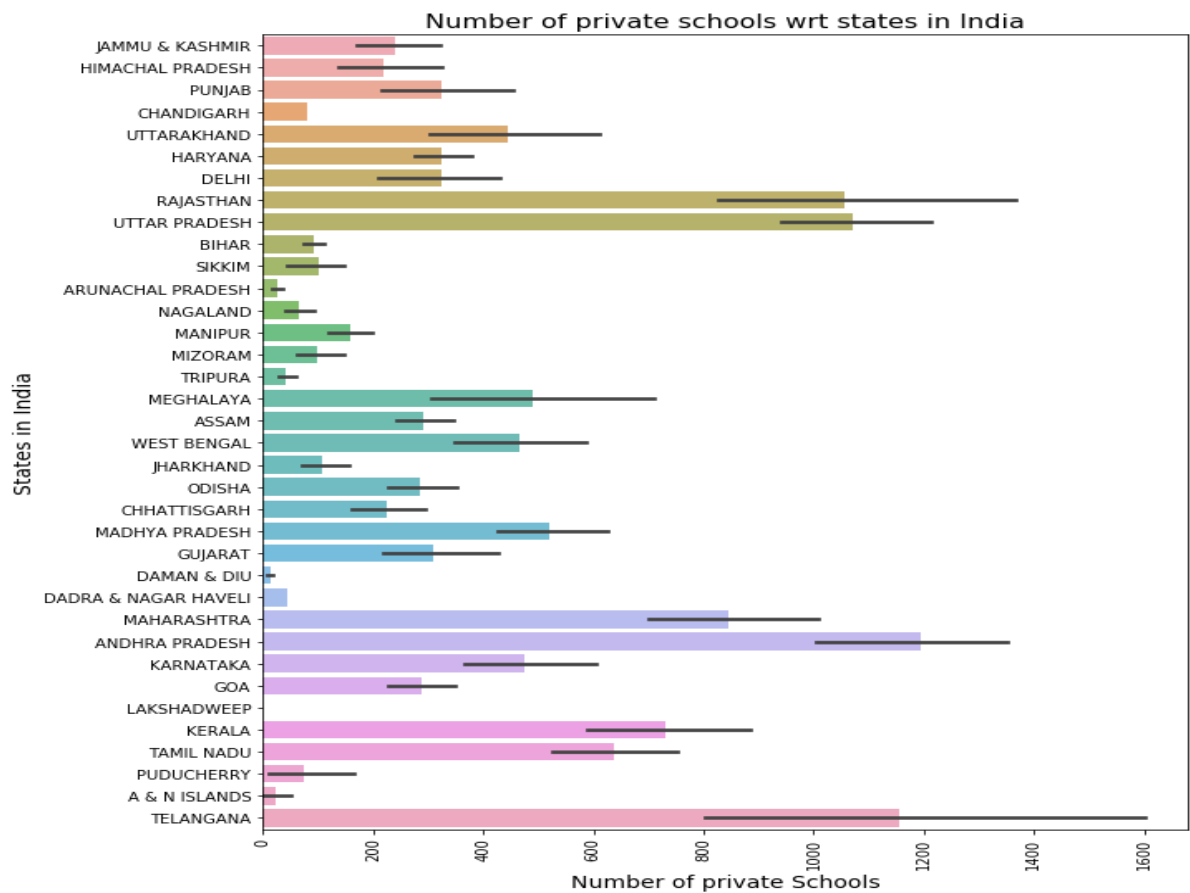
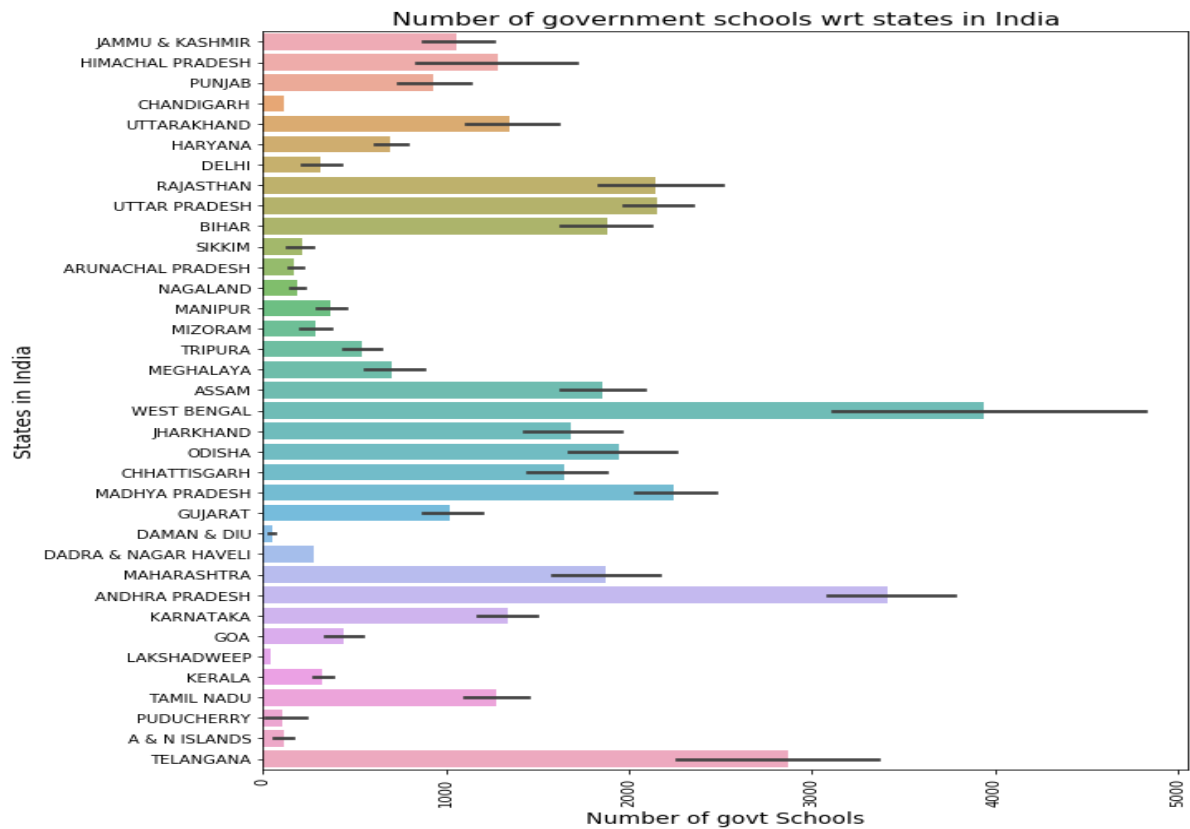


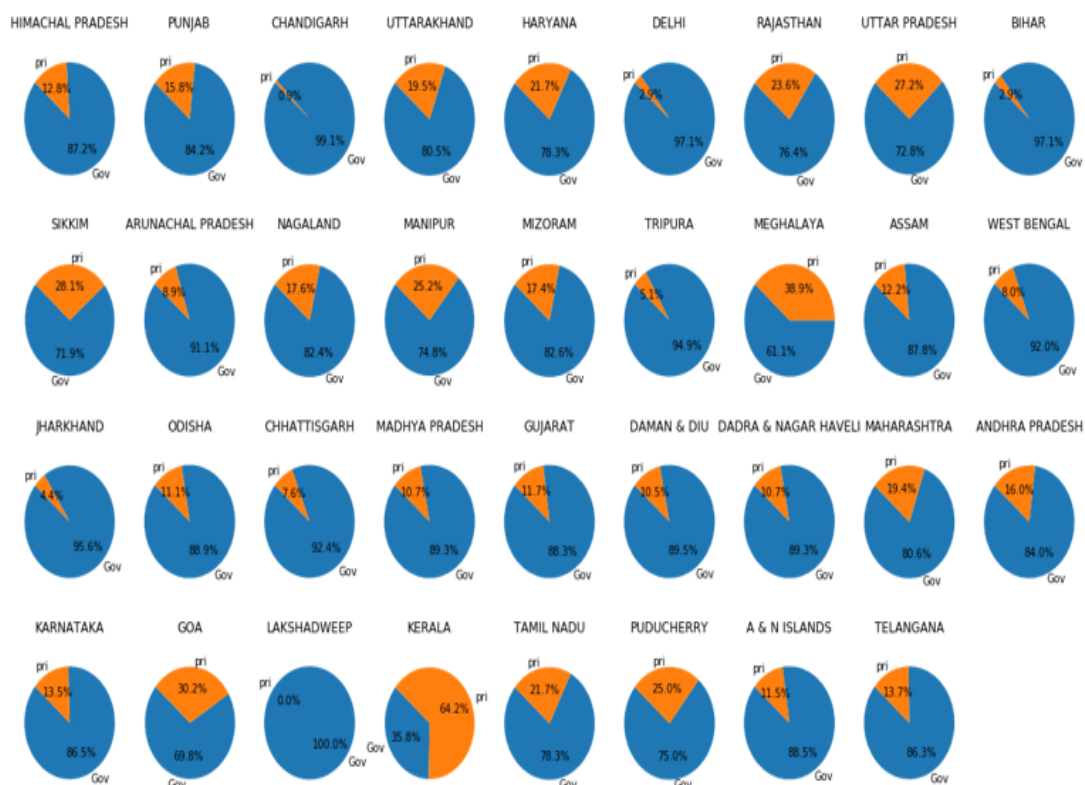
- ▶ graph shows that lesser number of states in India have high female literacy rate
- ▶ Low female literacy rate means an overall sluggish growth of India, as it impacts every arena of the development.
- ▶ India is struggling hard to stabilize its growing population through family planning programs . But if females are illiterate, then this has a direct and negative impact on these initiatives.
- ▶ The negative attitude of parents towards the girl child and her education is one of the major reasons of low female literacy rate in India.
- ▶ Kerala has the highest female literacy rate whereas Rajasthan has the lowest female literacy rate in India. States such as Uttar Pradesh and Bihar that are the most populated states in India show low levels of female literacy. This is directly related to the health and infant mortality. Kerala has the lowest infant mortality whereas states of Bihar and Uttar Pradesh have a high mortality rate.
- ▶ Female literacy rate in urban areas is higher as compared to rural India. In Rajasthan, most of the rural women are illiterate.
- ▶ Poverty is the root cause of many problems in India and also of low female literacy rate. More than one-third of population in India is living below the poverty line. Though government is putting efforts to make the primary education free but still parents are not ready to send their girls to school. To this is connected the accessibility to schools. In most of the rural areas lack of easy accessibility to school

is another reason for low female literacy rate. Parents do not prefer to send girls to schools if these are located at a far distance from their village or home. Even if schools are there then lack of adequate school facilities becomes a hurdle. Some of the schools are really in pathetic conditions and do not have even basic facilities. As per a survey, 54% of schools in Uttar Pradesh do not have water facility and 80% do not have latrine facilities. Even some schools do not have enough rooms to accommodate all the students.

## 2.HIGHER NUMBER OF PRIVATE SCHOOLS IMPLY HIGHER LITERACY RATES

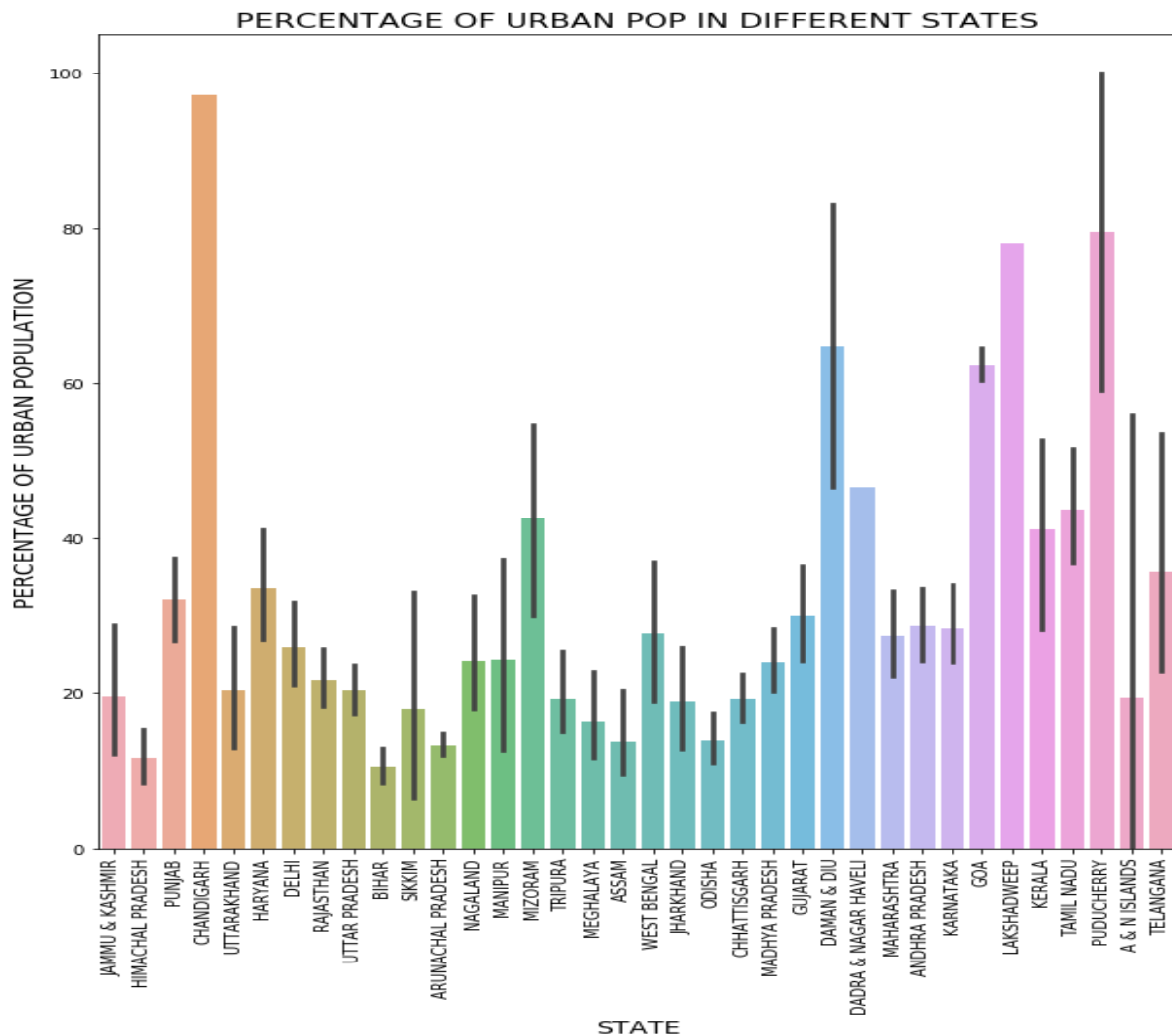






- ▶ 1) The number of governments schools is high in almost all states of India (except Kerala)
- ▶ 2) 60% of the schools in Kerala are private schools, which is unlike any other state ( it is an anomaly)
- ▶ 3) Incidentally Kerala also has the highest literacy rate among the Indian states
- ▶ 4) In Bihar, which has one of the lowest literacy rates only 3% of the schools are private
- ▶ Does this mean that a states literacy depends on the number of private educational institutions? We can safely assume that there is a relation (as little as it may seem)
- ▶ POSSIBLE REASON: Education imparted through private schools is more impactful and sophisticated compared to public schools. The infrastructure and other impeccable facilities provided by these schools ensures a thorough grooming of the students which enables them to be well equipped with the norms of the society. Also, the students get to study in an environment where there's a sense of discipline and dedication. There's a lot of importance given to encourage creativity and innovation among the students. All these provisions enable more number of parents to have a sense of security which ensures them to enroll their wards into private schools. This in turn leads to an increase in the literacy rate.

### 3.GREATER URBAN POPULATION SHOWS BETTER LITERACY RATE

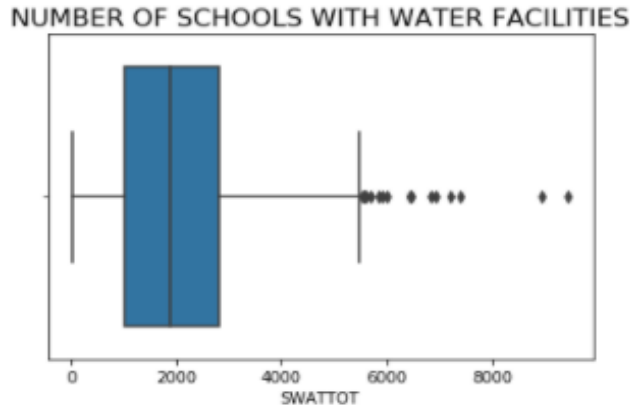
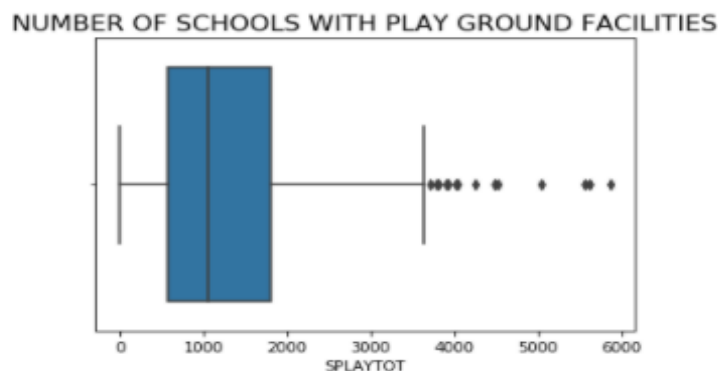
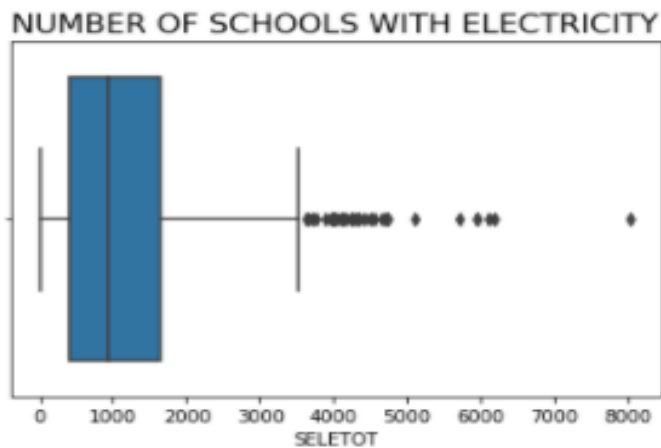


- It is observed that Bihar, which has the least literacy rate has the least percentage of urban population
- All states that have a fairly high percentage of urban population have good literacy rates.
- So it won't be incorrect if we say that the more the urban population in a state, the more is its literacy rate.

POSSIBLE REASON: Due to the availability of good teaching staff, course material, technology, and other teaching aids, urban areas can provide education of higher quality. Urban living offers many job opportunities and several career options. Urban areas have better infrastructure and transport facilities. Lacking good staff, funds, learning material, and educational tools, rural areas cannot provide the kind of education that urban areas can. Government funds as well as relief or rehabilitation programs may take longer to reach the rural areas.

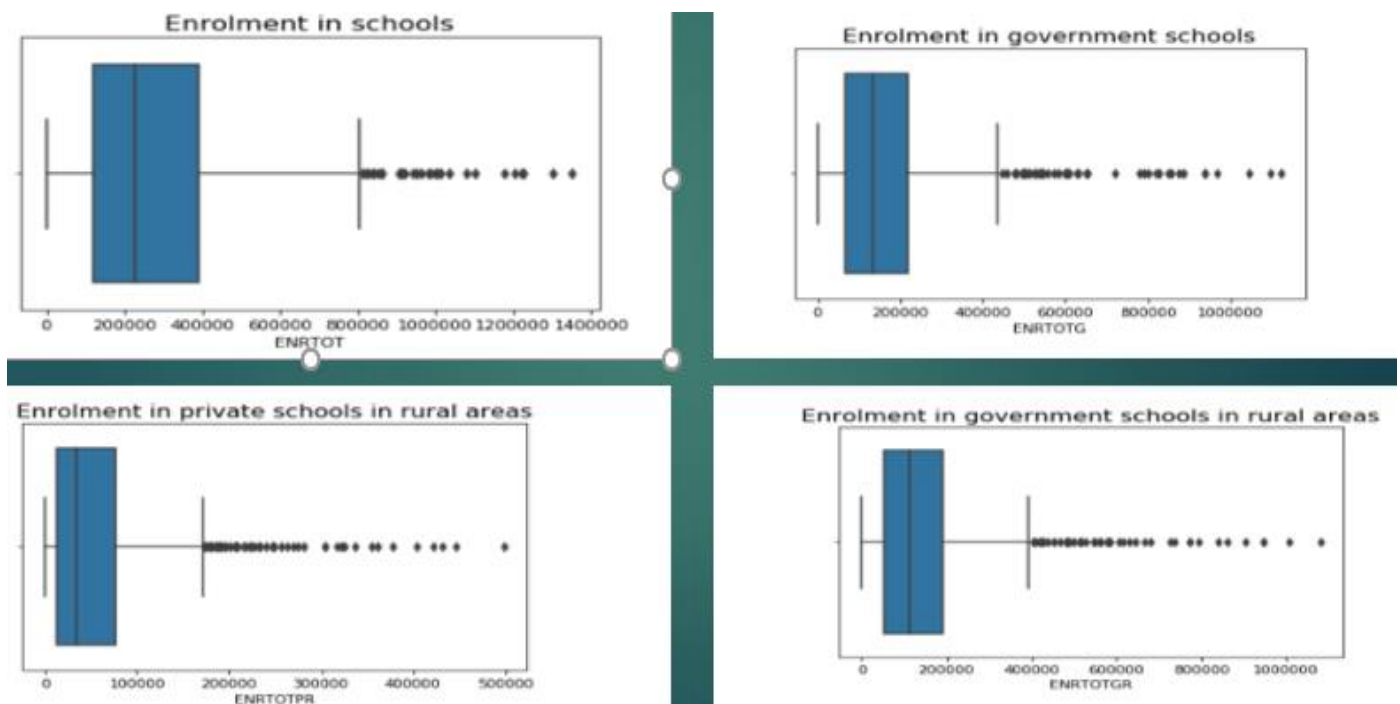
#### 4. LOW NUMBER OF WELL-EQUIPPED SCHOOLS IN INDIA!!





- Plotting box plots for number of schools with water, electricity and playground facilities for children reveals that relatively fewer schools are well equipped in India. All three boxplots turned out to be right skewed.
- When the schools are well facilitated it will surely encourage enrollment of a greater number of students. So, we can infer that by improving the infrastructure and providing better amenities along with quality education, it might be possible to increase the rate of literacy in the country.

## ► 5. COMPARING ENROLMENT IN GOVERNMENT AND PRIVATE SCHOOLS: WHY GOVERNMENT SCHOOLS ARE MORE POPULAR IN RURAL AREAS?

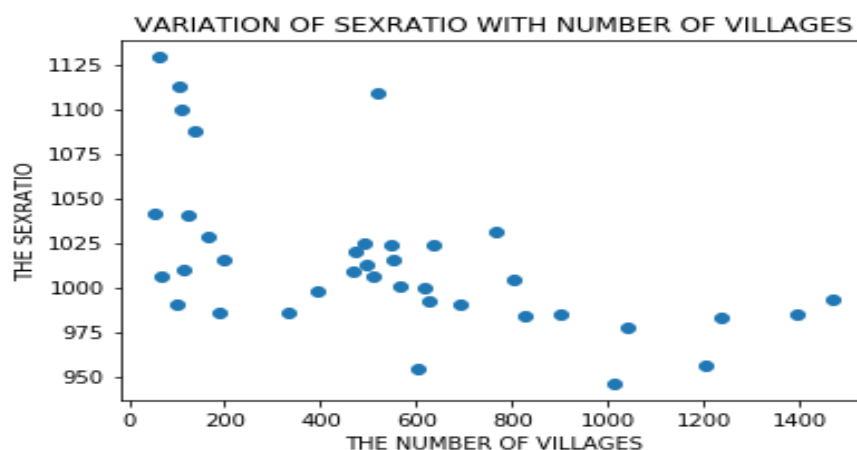


First graph (boxplot of enrolment in schools in India) is right skewed. This infers that more data points lie to the left of the graph. So, we can say that number of schools with high enrolment is relatively low. The second graph for enrolment in government schools in India also turned out to be right skewed.

The boxplots for enrolment in government and private schools in rural areas depict that enrolment in government schools is higher than in private in rural areas (because the graph is more right skewed for private).

**POSSIBLE REASON:** Because the number of private schools is itself few in number in rural areas. And also, the expenses cannot be met by the working class in rural areas and the mindset of the people isn't accustomed to give enough importance to educate their children at any cost. Hence, we can say that enrollment in private schools is less in rural areas.

## 6. DECREASING SEXRATIO WITH INCREASE IN NUMBER OF VILLAGES!



1.>On plotting Sex ratio against the number of villages, it was observed that sex ratio reduces with increase in number of villages.

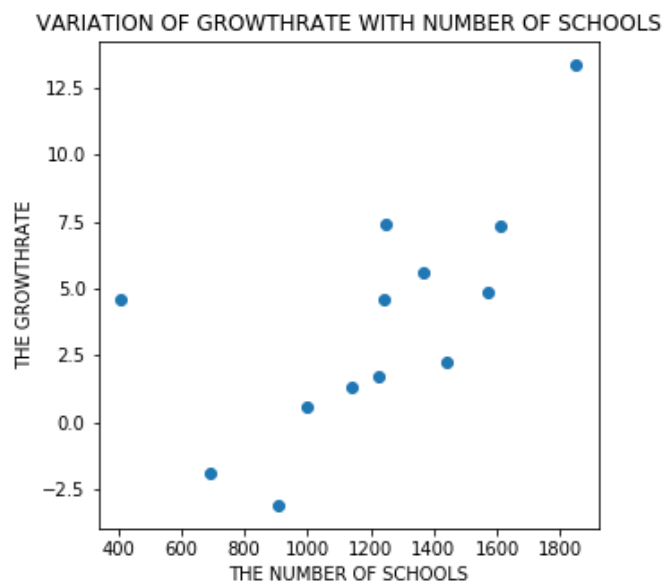
2.>Also carefully looking at the bar plot we can observe that Kerala, the state with the highest literacy rate has the least number of villages.

This serves as a classic backdrop to reflect the irrational ideas of the society. As the rural areas lack educational benefits, people rely heavily upon their superstitious beliefs which results in female infanticide- one of the core stigmas of the Indian society.

The birth of a girl child is often perceived as a huge liability and are nipped in the bud. This is more prevalent in rural areas as they haven't been exposed to rational thinking and this in turn leads to the decline in the sex ratio.

It is important to change these backward perceptions of people which is possible only through education!

## 7. INCREASE IN THE GROWTHRATE WITH INCREASE IN NUMBER OF SCHOOLS

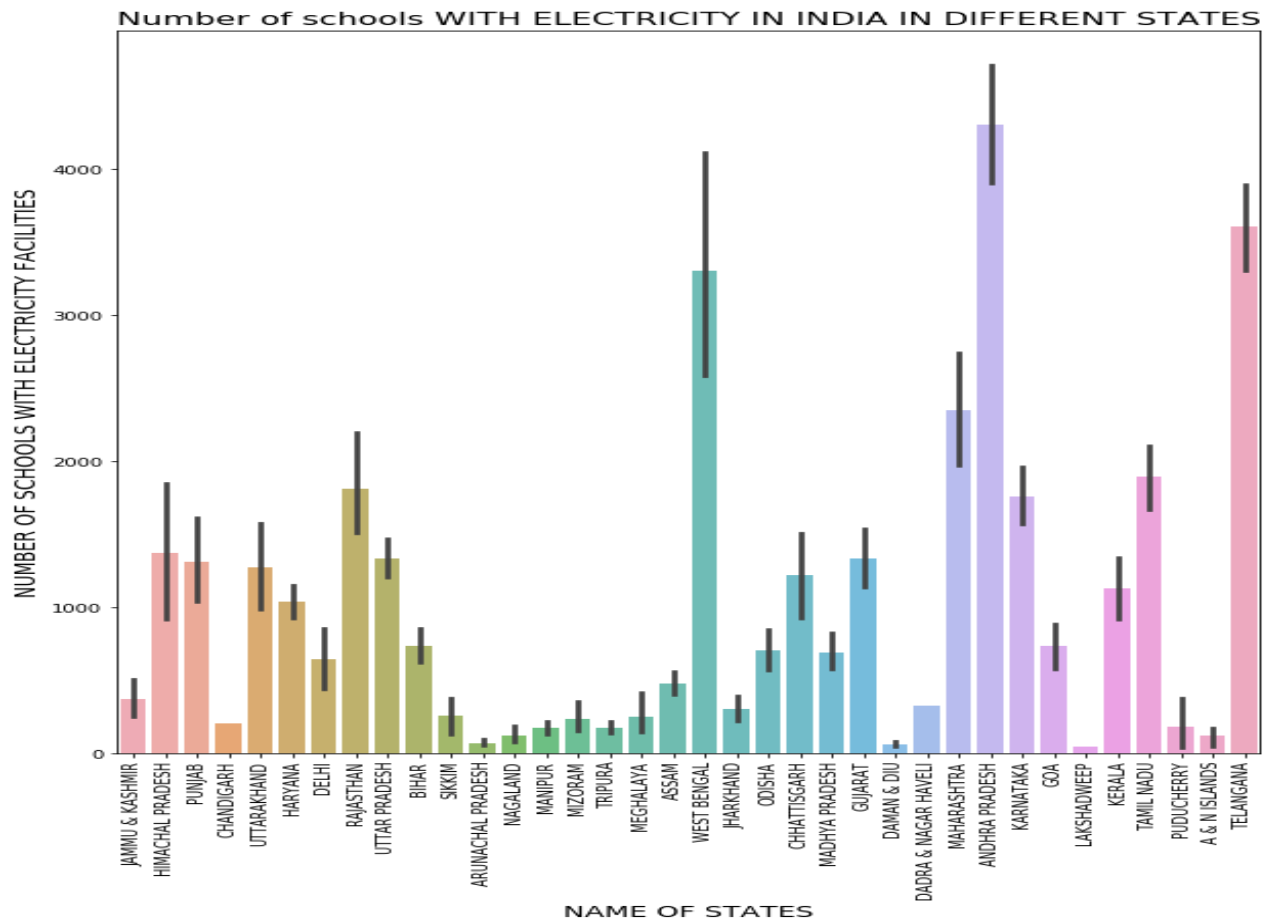


The scatter plot for Growth rate versus the number of schools appears to be a little close to a linearly increasing graph.

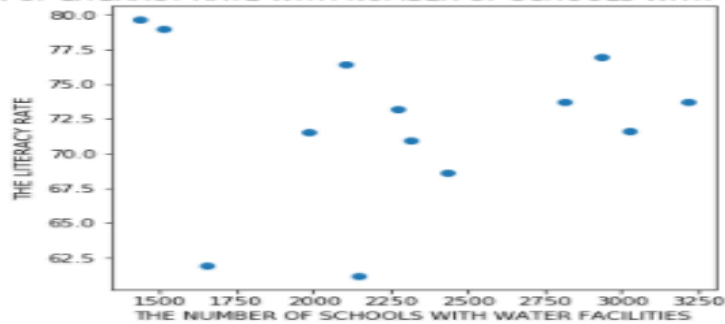
More number of schools result in a greater number of people enrolling their wards to these schools. This could also influence other people who haven't realised the importance of education to enroll their children to these schools which is bound to have affordable expenses due to their large number.

Hence there's a proportional increase in growth rate along with the increase in literacy.

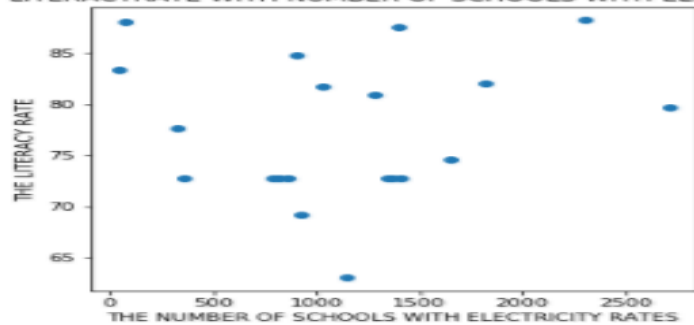
## 8.>WE CANNOT ASSUME THAT BETTER EQUIPPED SCHOOLS GUARANTEE BETTER LITERACY RATE!!!



VARIATION OF LITERACY RATE WITH NUMBER OF SCHOOLS WITH WATER FACILITIES



VARIATION OF LITERACY RATE WITH NUMBER OF SCHOOLS WITH ELECTRICITY FACILITIES



This is an interesting observation we come across in our dataset. Our paradigm of a 'well equipped' school includes better educational facilities, supply of electricity and drinking water, playgrounds etc. But according to the

data we have assessed, these well-equipped schools hardly play a role in the increase of literacy rate.

It is a convention that better the quality of schools more people are encouraged to enroll their wards. **But our dataset contradicts that very fact as the states with extremely well facilitated schools have a literacy rate on the lower end.** This gives us a significant insight into the process of data analysis i.e., it is not always possible to estimate an accurate measure of quantities in the dataset and draw possible conclusions based on them because there's always a good probability of going wrong when contradictions such as these occur. Hence assuming that better equipped schools guarantee better literacy rates would be an example for bad visualization.

## CONCLUSION

As we have seen from the above analysis, it is possible to understand the crucial characteristics that play a major role in the educational sector of the country. We have also seen contradictory data which requires another thorough sampling of the data lest we go wrong in our conclusions.

We can see that there's a lot of scope for improvement to increase the literacy percentage. For example, imparting free education or also considerably reducing the expenses will surely increase the admission of students in the rural sector. The Right to Free and Compulsory Education Act implemented in 2009 has definitely brought an improvement clearly depicted in this dataset which is from 2015-16.

The introduction of mid-day meal schemes in government schools has encouraged parents in the rural sector to send their children to school. Though there's a need for tweaking the Indian education system, we can surely see a positive scope for boosting the growth rate. With the vast social media influence today, it is feasible to instill the importance of education in everyone's mind and hope for a better future for the progeny as well as the nation.