EXPLORATORY DATA ANALYSIS

FEBRUARY 2022

SUBMITTED BY –

ISHANK VARSHNEY	ROLL NO.:15
ROHINI SINGH	ROLL NO. :36
VISHVENDRA SINGH	ROLL NO.:56

Contents

1 The Dataset

We have chosen the dataset - **Percentage of Schools with Computers from 2013-14 to 2015-16** from this link: https://data.gov.in/resources/percentage-schools-computers-2013-14-2015-16.

Description

The dataset consists of state wise, year wise, Percentage of schools having facility/availability of computers in that particular year. the It has two categorical and eleven numerical attributes, which are as follows:

• State and UT Name are categorical attributes that denote the name of the states and UTs respectively.

• Year is a categorical attribute that denotes the year in which the data was recorded.

• Other ten are continuous attributes that denotes the percentage of specific School Level of Indian Education System having computers.

The Levels and their description is as follows:

- Primary (class 1 to 5)
- Upper Primary(class 6 and 7)
- Secondary (class 8 to 10)
- Higher Secondary (class 11 and 12)

And the attributes based on these levels are:

- Primary_Only – Lower primary Schools of classes 1 to 5 (age 6 to 10)
- Primary_with_U_Primary – Primary Schools from classes 1 to 7 (age 6 to 12)
- Primary_with_U_Primary_Sec_HrSec – Schools from classes 1 to 12 (age 6 to 18)
- U_Primary_Only – Primary Schools from classes 6 and 7 (age 11 to 12)
- U_Primary_With_Sec_HrSec – Schools from classes from 6 to 12 (age 11 to 18)
- Primary_with_U_Primary_Sec – Schools from classes 1 to 7 (age 6 to 12)
- U_Primary_With_Sec – Schools from classes 6 to 10 (age 11 to 15)
- Sec_Only – Schools from classes 8 to 10 (age 13 to 15)
- Sec_with_HrSec. – Schools from classes from 8 to 12 (age 13 to 18)
- HrSec_Only - Schools from classes from 11 and 12 (age 17 and 18)

• All Schools is a continuous attribute that denotes the percentage of schools (i.e. all the levels of the school) which have facility of the computers in a particular state/UT.

Purpose of the Data

The data can be used to study and analyze percentage of Schools with computers of various States and UTs statistics. A few of them are as follows:

1. What is the percentage of schools with computers in each state in each specific year?

2. To study whether data of all schools has any influence due to data of any level of school or vice versa in each state.

3. Studying the states with the highest number of schools with computers. So that we get to know which state has good education facilities or schools of which state need improvements in this field.

4. Studying the Growth pattern and analyzing which state has constantly focused on the increase of computers in schools of their states.

5. Students of which class are exposed to the use of computers most?

6.To Study whether our dataset is right skewed or left skewed and lightly tailed or heavily tailed.

1.1 Reading and Cleaning the Dataset

```
library(ggplot2)
library(moments)
library(dplyr)
library(reshape2)
library(corrplot)
school <- read.csv("C:/Users/Dell/Downloads/schools_with_computer.csv")
head(school,5)
```

```
                  State_UT    year Primary_Only Primary_with_U_Primary Primary_with_U_Primary_Sec_HrSec U_Primary_Only U_Primary_With_Sec_HrSec
1 Andaman & Nicobar Islands 2013-14        30.40                 73.68                           89.74          0.00                  100.00
2 Andaman & Nicobar Islands 2014-15        30.94                 76.54                           92.11        100.00                   94.74
3 Andaman & Nicobar Islands 2015-16        28.44                 78.57                           92.50          0.00                   94.74
4           Andhra Pradesh 2013-14        12.73                 42.72                           86.99         45.45                   17.07
5           Andhra Pradesh 2014-15        10.29                 44.15                           88.46         50.00                   62.22
  Primary_with_U_Primary_Sec U_Primary_With_Sec Sec_Only Sec_with_HrSec. HrSec_Only All.Schools
1                      97.92              0.00     0.00          100.00       0.00       53.06
2                     100.00              0.00     0.00          100.00       0.00       57.25
3                     100.00              0.00     0.00          100.00       0.00       57.00
4                      68.18             73.23    60.00           33.33      19.32       29.57
5                      68.38             76.59    70.97           66.67      41.60       28.06
```

Since our dataset does not contains "NA" (not applicable) values, so there is no cleaning required.

2 Data Analysis

Let's first try to analyze the numeric attributes. We will study the mean and the standard deviation of the data, and then see how the values of these attributes are distributed. A boxplot is an ideal way to visualize the distribution of the data.

2.1 Mean and Standard Deviation

The arithmetic mean is one of the measures of central tendency, and, the standard deviation is used to measure the amount of variation of the data from its mean. Using the R programming language, the mean and standard deviation of the data can be found using the functions mean and sd respectively.
For the All.Schools attribute, the mean value is 40.008 percent, and the standard deviation is 27.65808.

```
cat("All.Schools")
cat("\nMean:", mean(school$All.Schools))
cat("\nSD:", sd(school$All.Schools))
```

All.Schools
Mean: 40.008
SD: 27.65808

2.2 Five Number Summary

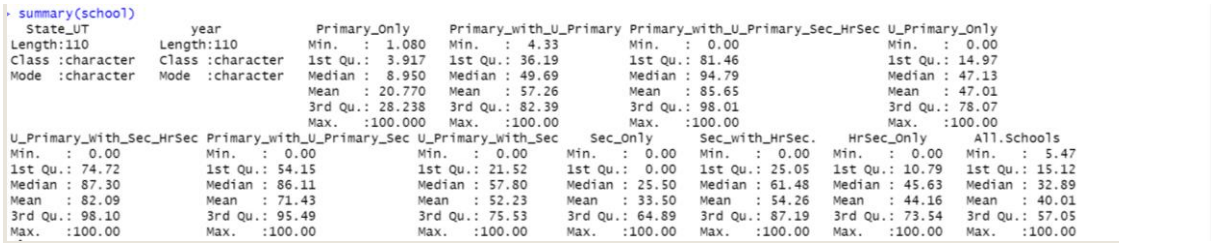
The mean is not a robust statistic, since it's easily influenced by outliers. We turn to a deeper analysis of the distribution of values, and make use of the five number summary of the data and plot the corresponding boxplots.

2.2.1 Boxplots

The five number summary can be generated by the function quantile or by function summary in R.

Note: The State_UT and Year attributes are categorical attribute since it denotes discrete values rather than a numerical quanitity. So, it does not make sense to calculate measures like mean, standard deviation, five number summary. Even if they show up in some visualization, we will ignore it since it has no meaning.

```
View(summary(school))
```



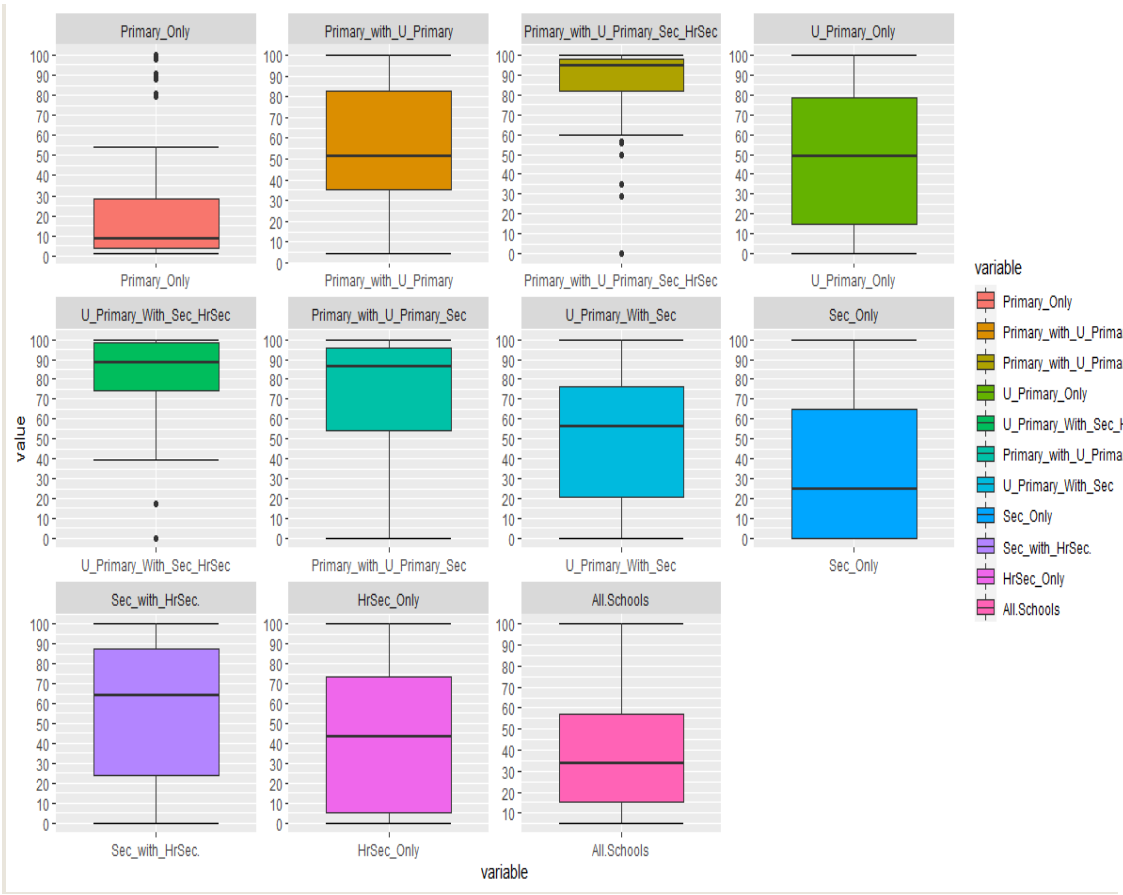
```
> summary(school)
  State_UT          year          Primary_Only     Primary_with_U_Primary Primary_with_U_Primary_Sec_HrSec U_Primary_Only
 Length:110         Length:110         Min.   :  1.080   Min.   :  4.33         Min.   :  0.00                 Min.   :  0.00
 Class :character   Class :character   1st Qu.:  3.917   1st Qu.: 36.19         1st Qu.: 81.46                 1st Qu.: 14.97
 Mode  :character   Mode  :character   Median :  8.950   Median : 49.69         Median : 94.79                 Median : 47.13
                                       Mean   : 20.770   Mean   : 57.26         Mean   : 85.65                 Mean   : 47.01
                                       3rd Qu.: 28.238   3rd Qu.: 82.39         3rd Qu.: 98.01                 3rd Qu.: 78.07
                                       Max.   :100.000   Max.   :100.00         Max.   :100.00                 Max.   :100.00
 U_Primary_With_Sec_HrSec Primary_with_U_Primary_Sec U_Primary_With_Sec    Sec_Only      Sec_with_HrSec.    HrSec_Only      All.Schools
 Min.   :  0.00           Min.   :  0.00             Min.   :  0.00     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  5.47
 1st Qu.: 74.72           1st Qu.: 54.15             1st Qu.: 21.52     1st Qu.:  0.00   1st Qu.: 25.05   1st Qu.: 10.79   1st Qu.: 15.12
 Median : 87.30           Median : 86.11             Median : 57.80     Median : 25.50   Median : 61.48   Median : 45.63   Median : 32.89
 Mean   : 82.09           Mean   : 71.43             Mean   : 52.23     Mean   : 33.50   Mean   : 54.26   Mean   : 44.16   Mean   : 40.01
 3rd Qu.: 98.10           3rd Qu.: 95.49             3rd Qu.: 75.53     3rd Qu.: 64.89   3rd Qu.: 87.19   3rd Qu.: 73.54   3rd Qu.: 57.05
 Max.   :100.00           Max.   :100.00             Max.   :100.00     Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :100.00
```

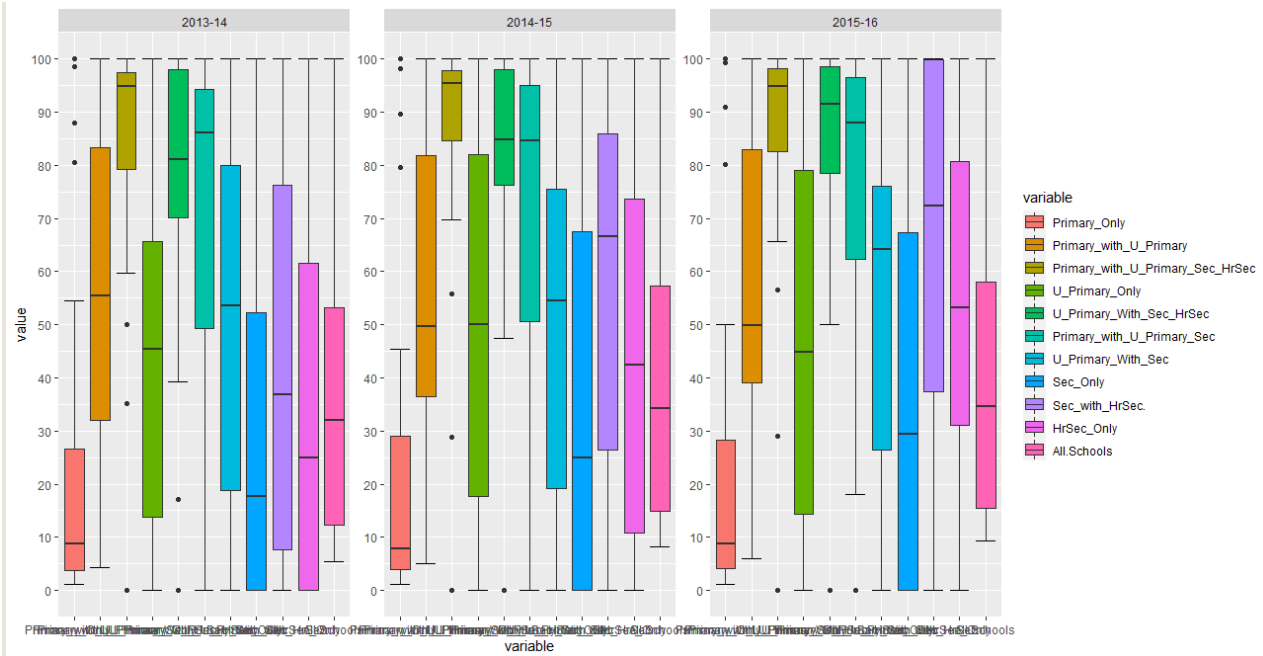
We can do the same by following piece of code:

```
cat("All.Schools\n")
print(quantile(school$All.Schools))
```

```
ggplot(data = melt(schoolStateUT), aes(x=variable, y=value))
+stat_boxplot(geom='errorbar') +geom_boxplot(aes(fill=variable)) +
scale_y_continuous(breaks = scales::pretty_breaks(n = 14)) + facet_wrap( ~
variable, scales="free")
```

```
ggplot(data = melt(schoolStateUT), aes(x=variable, y=value)) +
 stat_boxplot(geom='errorbar') +
 geom_boxplot(aes(fill=variable)) +
 scale_y_continuous(breaks = scales::pretty_breaks(n = 14)) +
facet_wrap(vars(year), scales="free")
```


Primary_Only
Primary_with_U_Primary
Primary_with_U_Primary_Sec_HrSec
U_Primary_Only
U_Primary_With_Sec_HrSec
Primary_with_U_Primary_Sec
U_Primary_With_Sec
Sec_Only
Sec_with_HrSec.
HrSec_Only
All.Schools
value
variable


2013-14
2014-15
2015-16
value
variable
Primary_Only
Primary_with_U_Primary
Primary_with_U_Primary_Sec_HrSec
U_Primary_Only
U_Primary_With_Sec_HrSec
Primary_with_U_Primary_Sec
U_Primary_With_Sec
Sec_Only
Sec_with_HrSec.
HrSec_Only
All.Schools

Interpretation

The five number summary is a set of five values that denote:

The data contains values of a lot of variables (Primary_Only, Primary_with_U_Primary ,Primary_with_U_Primary_Sec, U_Primary_Only, etc), so we are selecting a subset of the variables "All.Schools" for a clearer and more accurate picture.

1. Minimum: For the All.Schools attribute, the minimum value is5.47 percent. This means that 0% of the data falls below this value. The minimum corresponds to the horizontal line at the lower end of the bottom whisker in the boxplot.

2. First quartile: 25% of the All.Schools values fall below15.12 percent.The first quartile corresponds to the lower end of the box in the boxplot.

3. Second quartile (median): 50% of the All.Schools fall below 32.89 percent. The median corresponds to the horizontal line in the box in the boxplot.

4. Third quartile: 75% of the All.Schools fall below 57.05 percent. The third quartile corresponds to the upper end of the box in the boxplot.

5. Maximum: 100% of the Area values fall below 100 percent. The maximum corresponds to the horizontal line at the upper end of the upper whisker in the boxplot.

The range (maximum - minimum) and the Interquartile range (third quartile - first quartile) can be found using the R functions diff(range()) and IQR() respectively. They are as follows:

```
cat("---------- Range, IQR ----------\n")
cat("All.Schools")
cat("\nRange: ", diff(range(schoolStateUT$All.Schools)))
cat("\nIQR: ", IQR(schoolStateUT$All.Schools))
cat("\n---------------------------------\n")
```

```
---------- Range, IQR ----------

All.Schools
Range:  94.53
IQR:  42.09

-----------------------------------
```

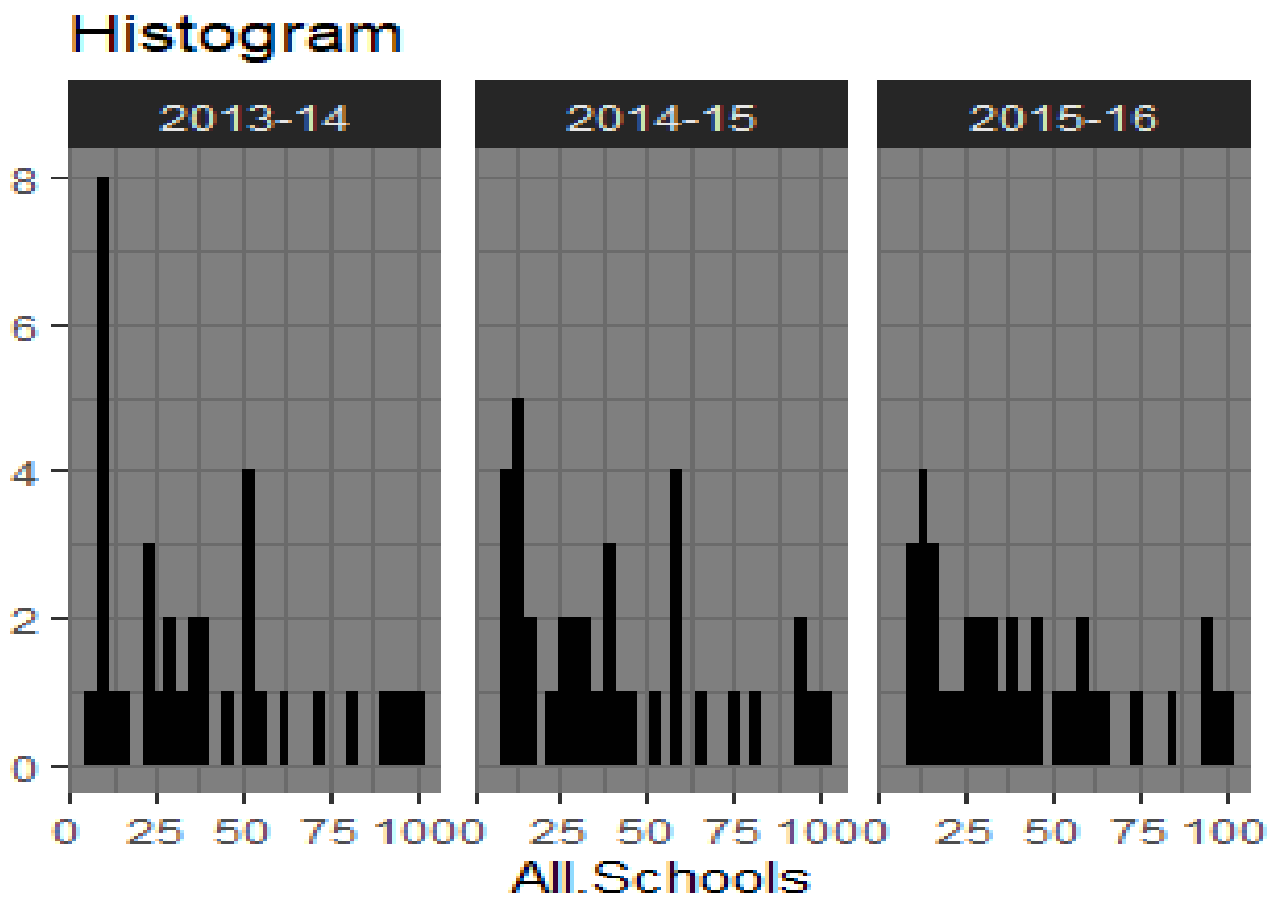
2.3 Distribution of the Schools(with computers) across the States

To get an overall idea of the state-wise percentage of all schools (with computers), we can study how many percent of all schools are available in each state in specific years which provide the facility of computers. This helps us to know, for example, how many states have a "low" number of schools with computers, how many states have a "medium" number of schools with computers, and how many states have a "high" number of schools with computers, depending on how we define "low", "medium", and "high". Here we are using the attribute "All.Schools" for analysis.

2.3.1 Histogram

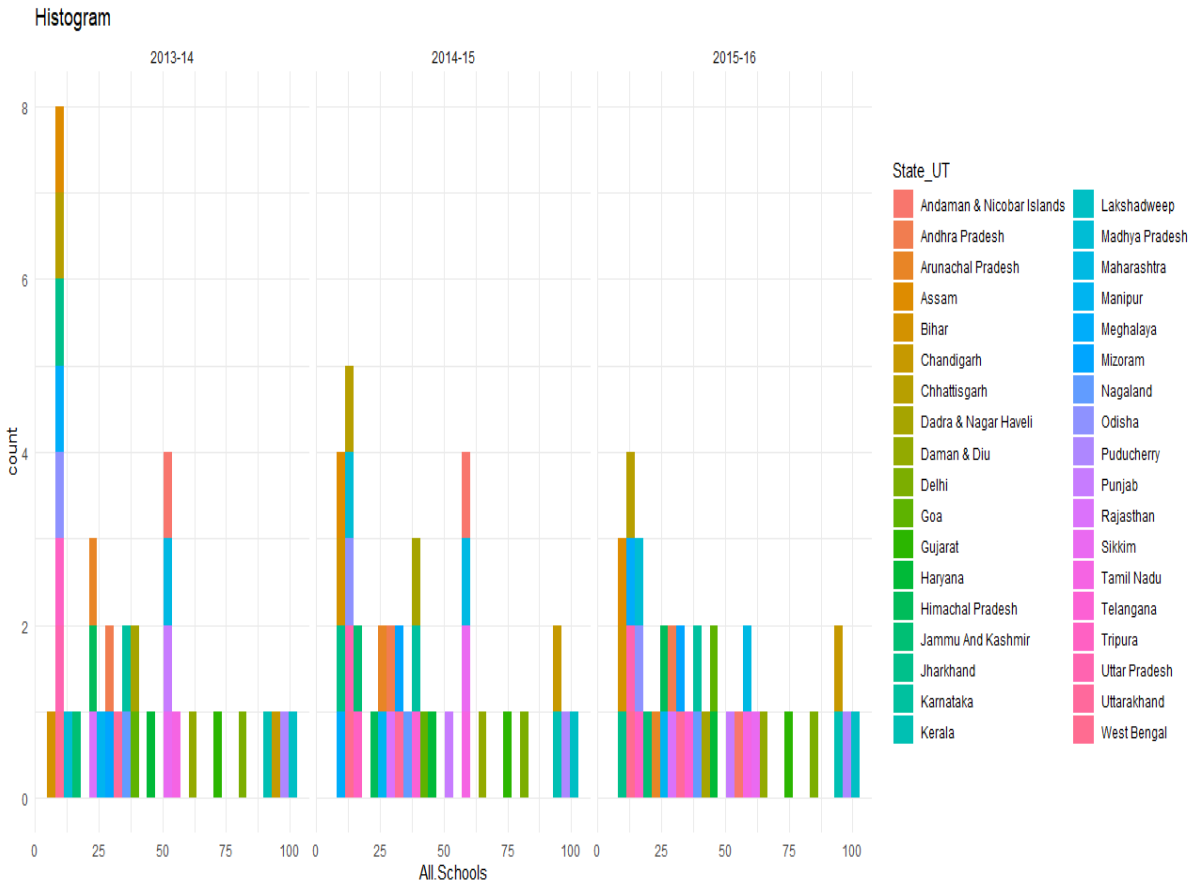
A histogram is an ideal way to plot the frequency counts of states that availability of computers within each interval of percentage values. We divide the percentage values into "bins" and count how many states fall into each bin.

```
ggplot(schoolStateUT) +
aes(x = All.Schools) +
geom_histogram(bins = 30L, fill = "#010101") +
labs(y = "", title = "Histogram") +
theme_dark() +
facet_wrap(vars(year))
```



In this graph, we are examining that how many states fall under the particular interval of percentage(i.e. actually showing percent of all schools having computer facility)

```
ggplot(schoolStateUT) +
aes(x = All.Schools, fill = State_UT) +
geom_histogram(bins = 30L) +
labs(title = "Histogram") +
scale_fill_hue() +
theme_minimal() +
facet_grid(vars(), vars(year))
```



In this graph, we are analysing the count as well as the state that fall under the particular percentage.
We define "low" as percentage of schools between 0 to 33%, "medium" as percentage of schools between 33% to 67% and "high" as percentage of schools above 67%.

From these histogram, we can see that the data is skewed right in all the years. Most of the states (18 states in 2013-14,18 states in 2014-15,19 states in 2015-2016) have "low" number of schools as compared to the other states. A lot of states have "medium" number of schools with facility of computers, and, only few state have significantly large percentage of schools with computers in schools in all of India.

2.4 Correlation between All.Schools and Primary_only

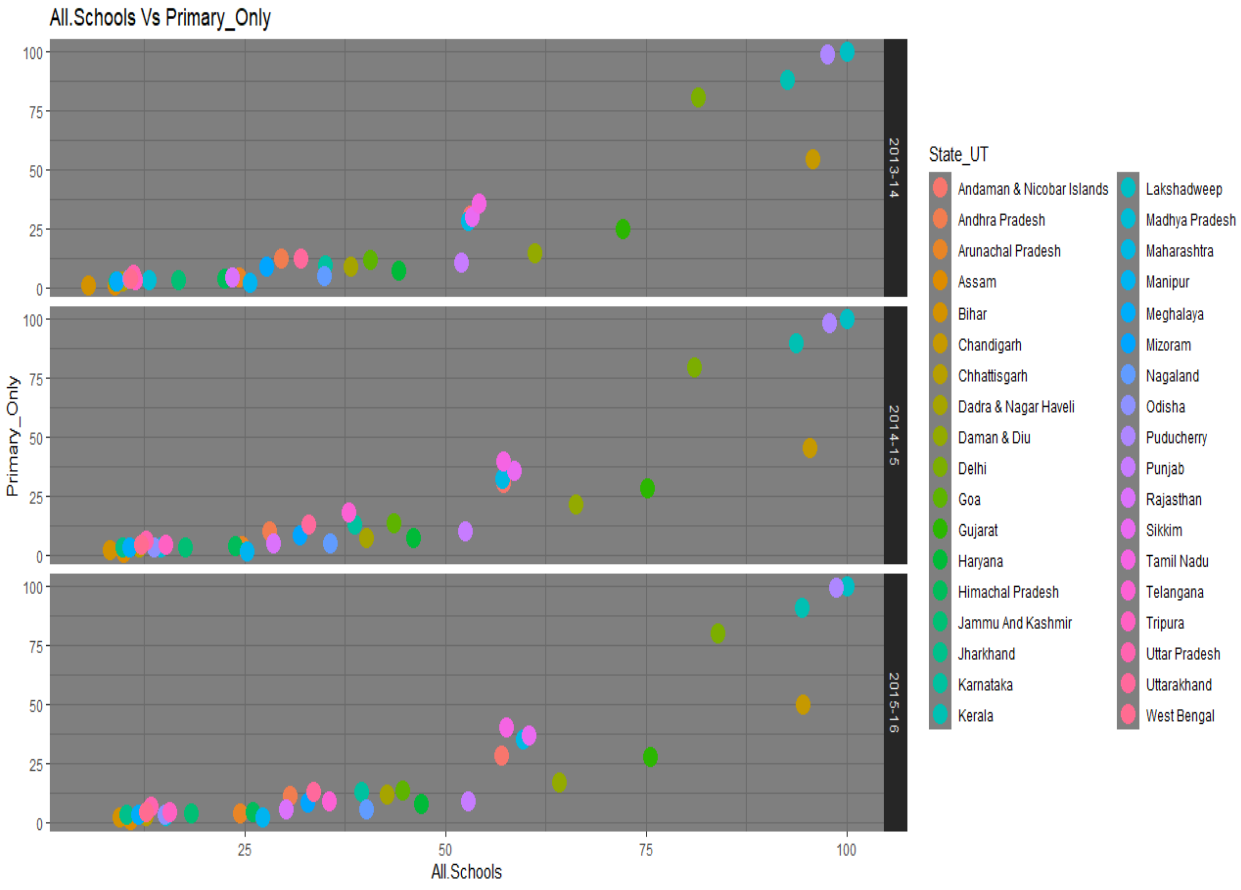
Correlation measures the degree to which two variables are related to each other. Here, we study whether attributes "All.Schools and Primary_only" have any sort of correlation. Also , we can study correlation of attribute "All.Schools" with other numeric attributes in the same way.

2.4.1 Scatter Plot

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

The scatter plot of All.Schools and Primary_only clearly shows us the positive exponential correlation between All.Schools and Primary_only as an upward rising trend. As the Percentage of primary_only schools increases, the corresponding percentage of all schools of that particular state also increases.

```
ggplot(schoolStateUT) +
aes(x = All.Schools, y = Primary_Only, colour = State_UT) +
geom_point(size = 5L) +
labs(title = "All.Schools Vs Primary_Only") +
scale_color_hue() +
theme_dark() +
facet_grid(vars(year), vars())
```

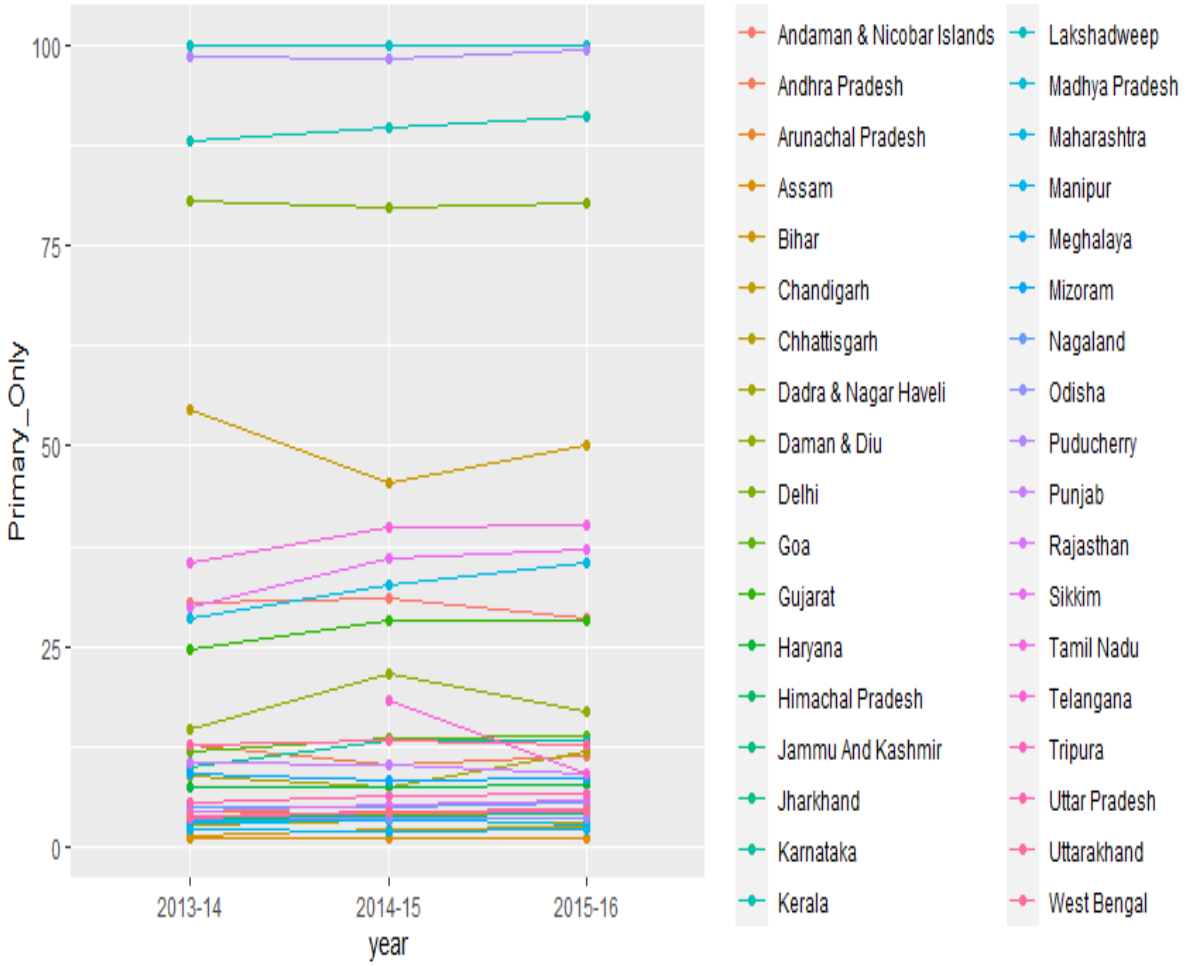


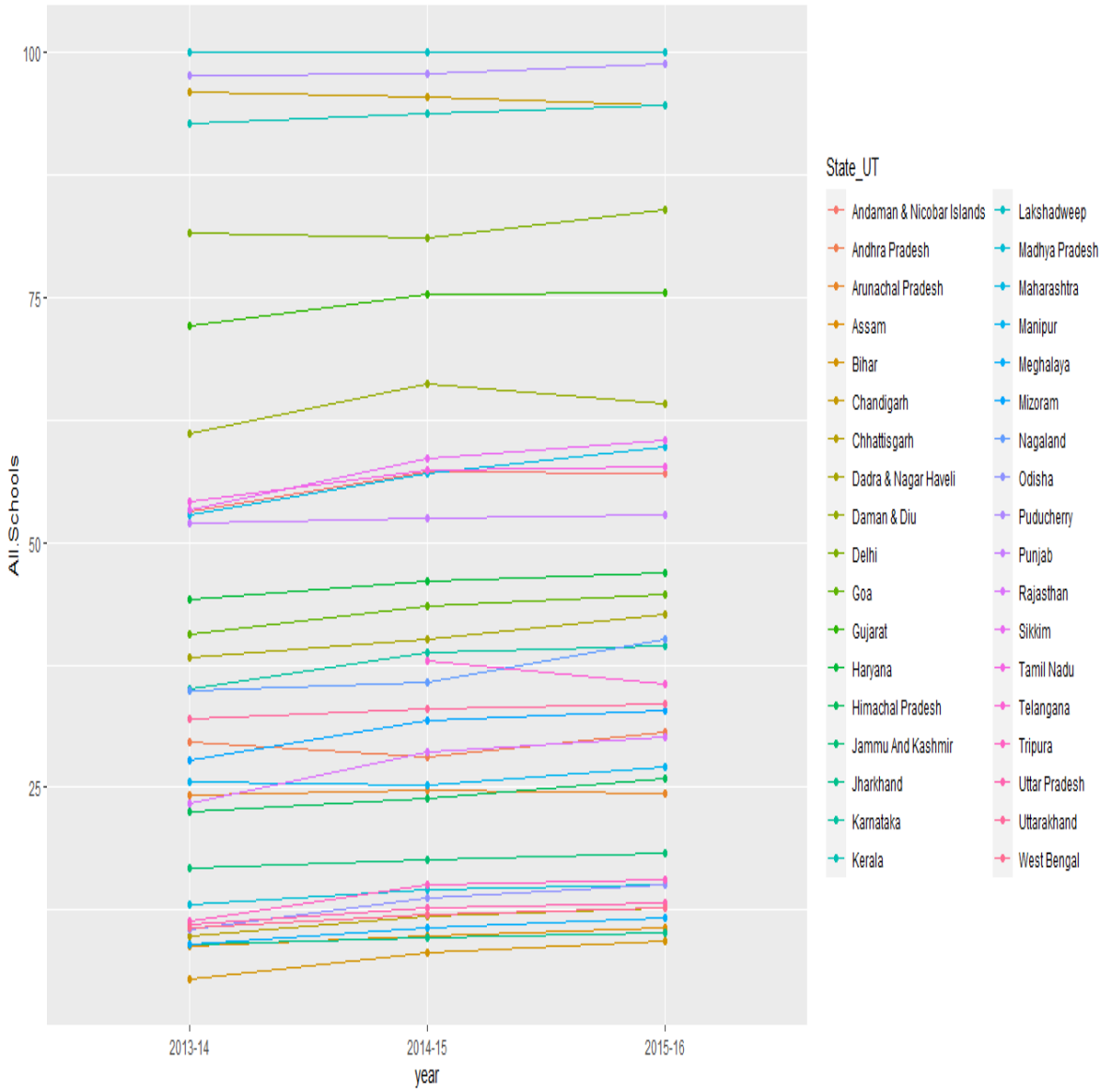
2.4.2 Line Plot

Another way to visualize the correlation is to plot a line chart for the percentage of All Schools of all states over the years 2013-2016.
This could be done in a single chart, but due to the vastly different scales of values and numerous states, we plot separate charts.

```
ggplot(data =schoolStateUT, aes(x=year, y=All.Schools,colour = State_UT,
group=State_UT)) +
 geom_line() +
 geom_point()

ggplot(data =schoolStateUT, aes(x=year, y=Primary_Only,colour = State_UT,
group=State_UT)) +
 geom_line() +
 geom_point()
```





Again, we can see the similar trends between their values in the two charts. For example, in the years, clearly there was a increase in primary only schools of most of the states. If we see all the schools in those years, we can see the same increase in most of those states.

2.5 State-wise Top contributor of schools with computers

We can study which are the states that have maximum schools with computers in a specific year. The visualization of this can be accomplished using a pie chart or a bar chart. For this purpose, we have filtered those rows which contains the year '2013-14'. So finally, we have the all the data in the year 2013-14 for each state. We can do the same for other years.

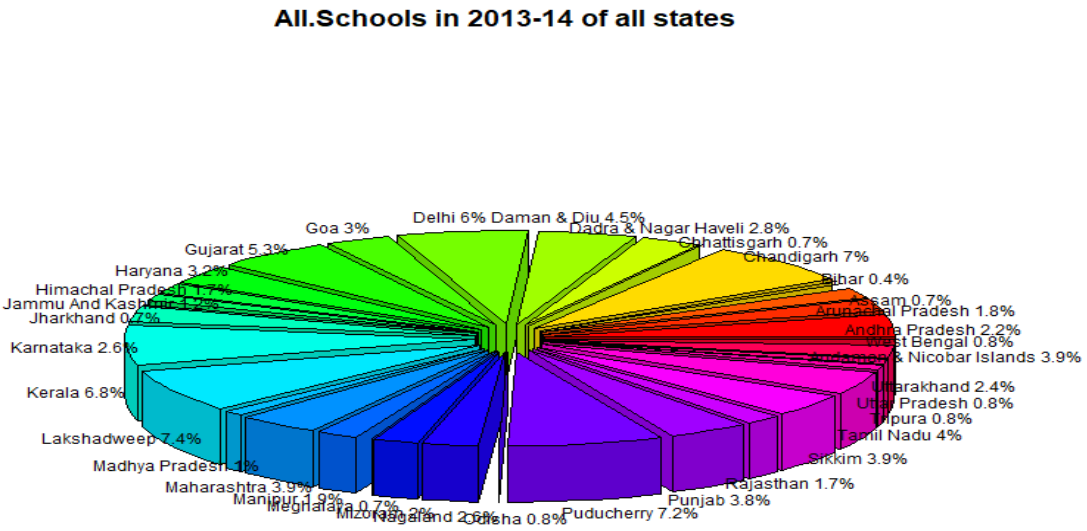
2.5.1 Pie Chart

We now plot the pie chart which shows the percentage of computers in all the schools of in the year 2013-14 in different states.

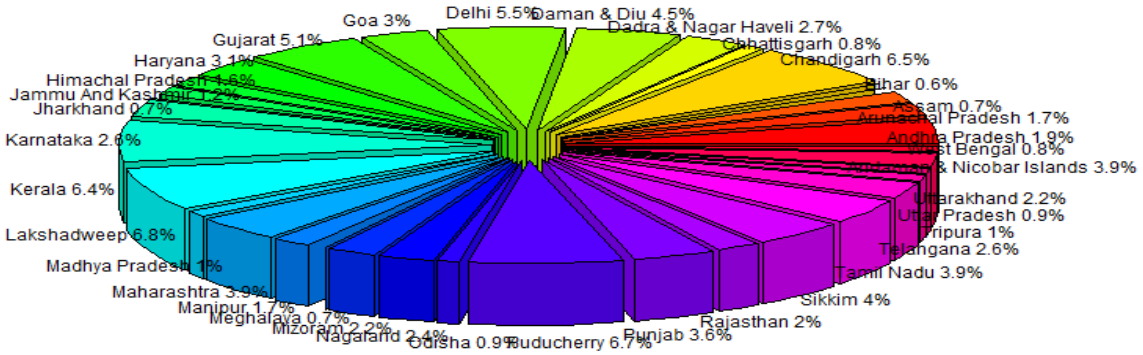
```
library(plotrix)
subset1<-schoolStateUT[schoolStateUT$year=="2013-14",]
pct <- round(subset1$All.Schools/sum(subset1$All.Schools)*100,1)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(subset1$All.Schools,labels =lbls,  explode = 0.1, main = "3D Pie Chart in R ")
```

```
subset2<-schoolStateUT[schoolStateUT$year=="2014-15",]
pct <- round(subset2$All.Schools/sum(subset2$All.Schools)*100,1)
lbls <- paste(subset2$State_UT,pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(subset2$All.Schools,labels =lbls,  explode = 0.1,labelcex =0.75,  main =
"All.Schools in 2014-15 of all states ")
```

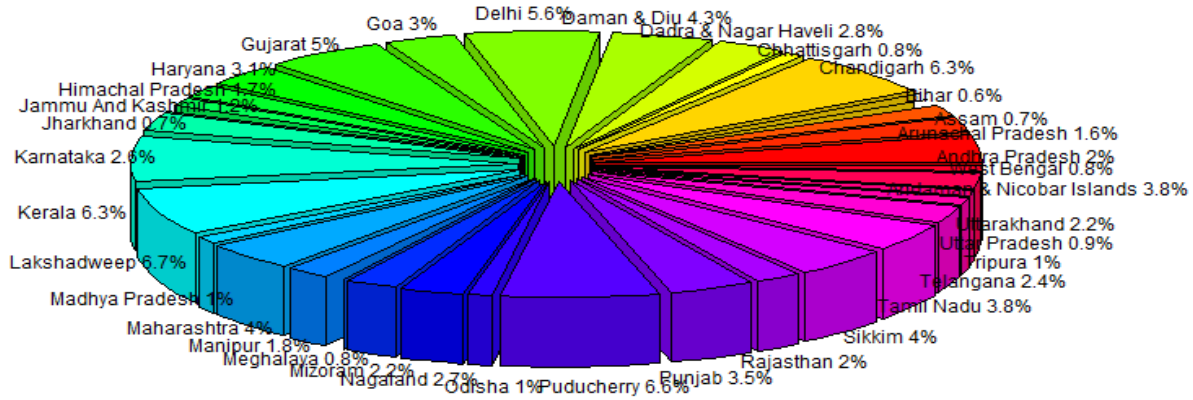
```
subset3<-schoolStateUT[schoolStateUT$year=="2015-16",]
pct <- round(subset3$All.Schools/sum(subset3$All.Schools)*100,1)
lbls <- paste(subset3$State_UT,pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(subset3$All.Schools,labels =lbls,  explode = 0.1,labelcex =0.75,  main =
"All.Schools in 2015-16 of all states ")
```



All.Schools in 2014-15 of all states



All.Schools in 2015-16 of all states

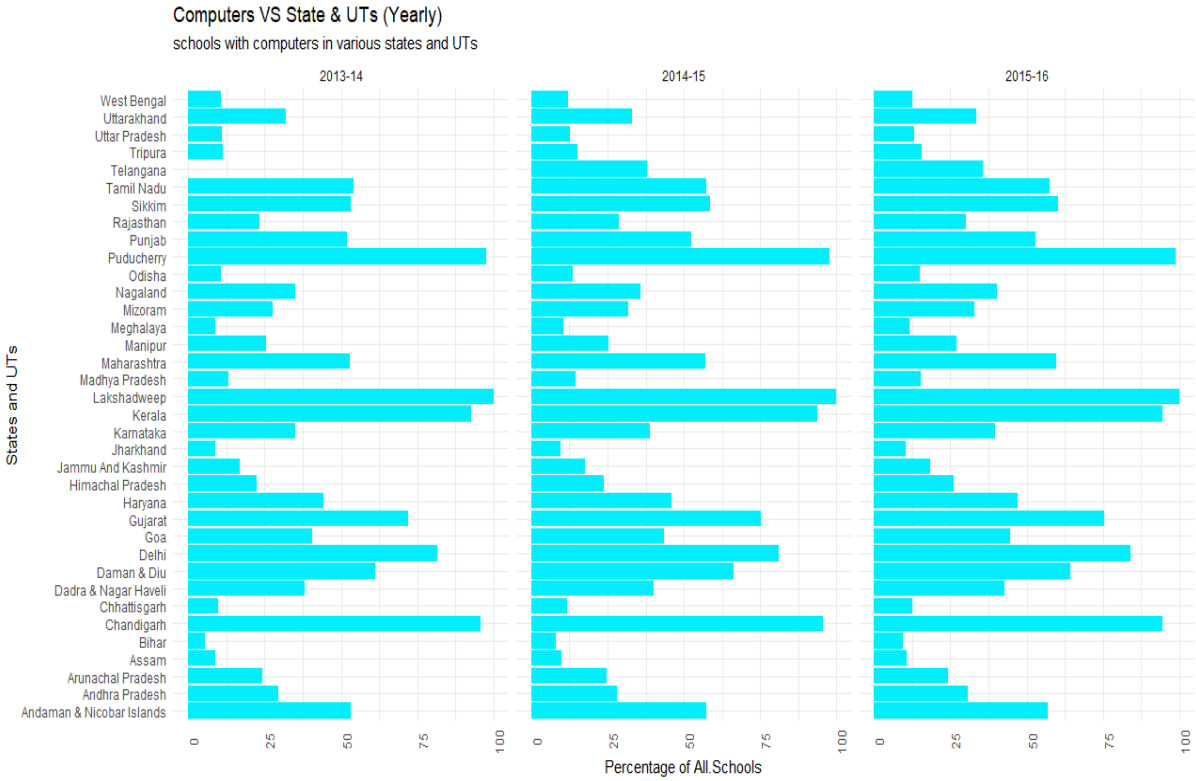


These pie charts show the contribution of all the states in all schools in the year 2013-14,2014-15 and 2015-16 respectively. It is evident from the pie chart that Lakshadweep(7.4%,6.8%,6.7%) is the on the top in all these years, followed by Puducherry(7.2%,6.7%,6.6%) and Chandigarh(7%,6.5%,6.3%) . Most of the states have average contribution in schools in providing computer facility.

2.5.2 Bar Chart

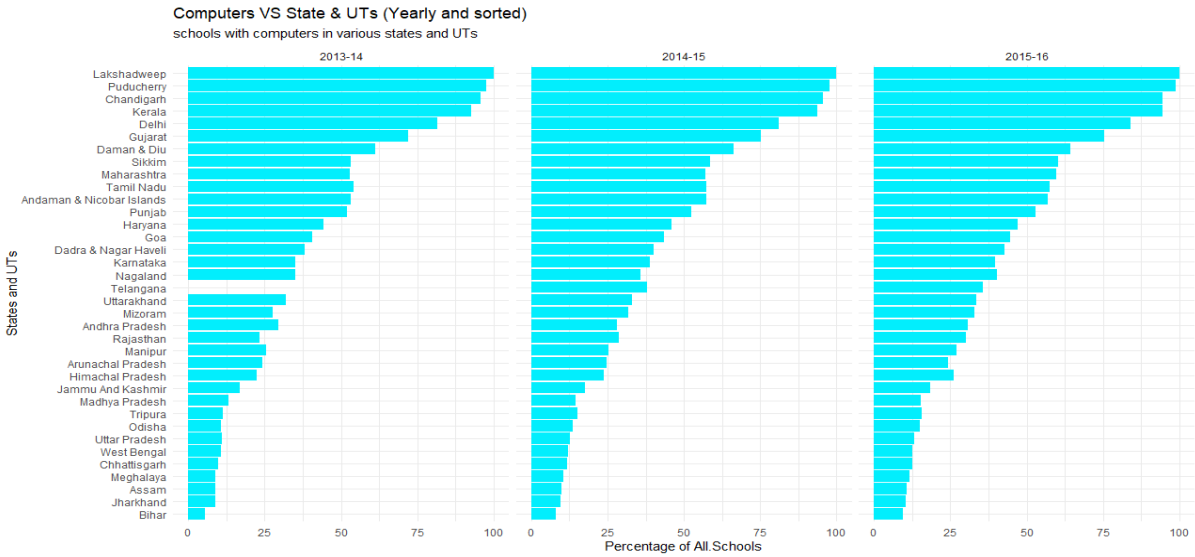
We can also plot a bar chart to show the percentage of schools in the year 2013-14, 2014-15 and 2015-16 in different states.

```
ggplot(schoolStateUT) +
 aes(x = State_UT, weight = All.Schools) +
 geom_bar(fill = "#02eefd") +
 labs(x = "States and UTs", y = "Percentage of All.Schools", title = "Computers VS State & UTs (Yearly)", subtitle = "schools with computers in various states and UTs") +
 coord_flip() +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 90)) +
 facet_grid(vars(), vars(year))
```



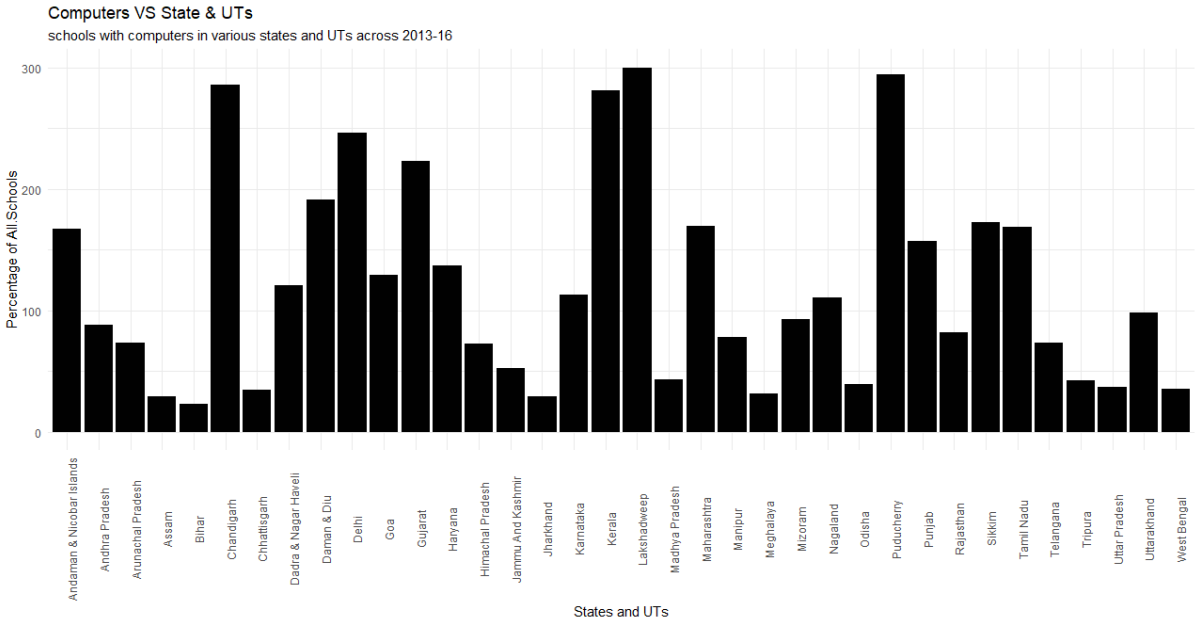
Sorting the bars of above graph will give more clearer picture.

```
ggplot(schoolStateUT, aes(reorder(State_UT, All.Schools), All.Schools)) +
 geom_col(fill = "#02eefd") +
 labs(x = "States and UTs", y = "Percentage of All.Schools", title = "Computers VS State & UTs (Yearly and sorted)", subtitle = "schools with computers in various states and UTs") +
 coord_flip() +
 theme_minimal() +
 facet_grid(vars(), vars(year))
```

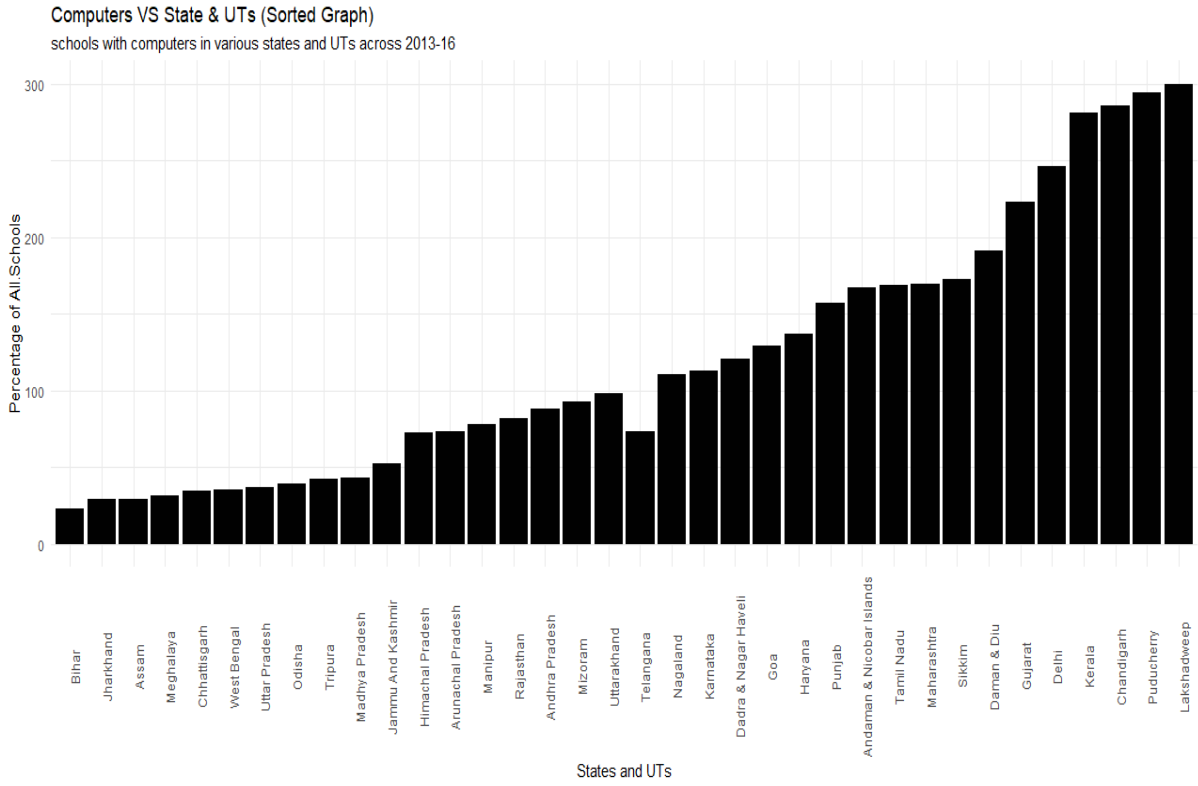


Again, we see that Lakshadweep is the on the top in all these years, followed by Puducherry and Chandigarh. Also, most of the states have average contribution in having schools (providing computer facility). The additional information that we get in bar chart is the percentage value for all the states which was missing in the case of Pie chart.

```
ggplot(schoolStateUT) +
 aes(x = State_UT, weight = All.Schools) +
 geom_bar(fill = "#010101") +
 labs(x = "States and UTs", y = "Percentage of All.Schools", title = "Computers VS State & UTs", subtitle = "schools with computers in various states and UTs across 2013-16") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 90))
```



```
ggplot(schoolStateUT, aes(reorder(State_UT, All.Schools), All.Schools)) +
  geom_col(fill = "#010101") +
  labs(x = "States and UTs", y = "Percentage of All.Schools", title = "Computers VS State & UTs (Sorted Graph)", subtitle = "schools with computers in various states and UTs across 2013-16") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90))
```



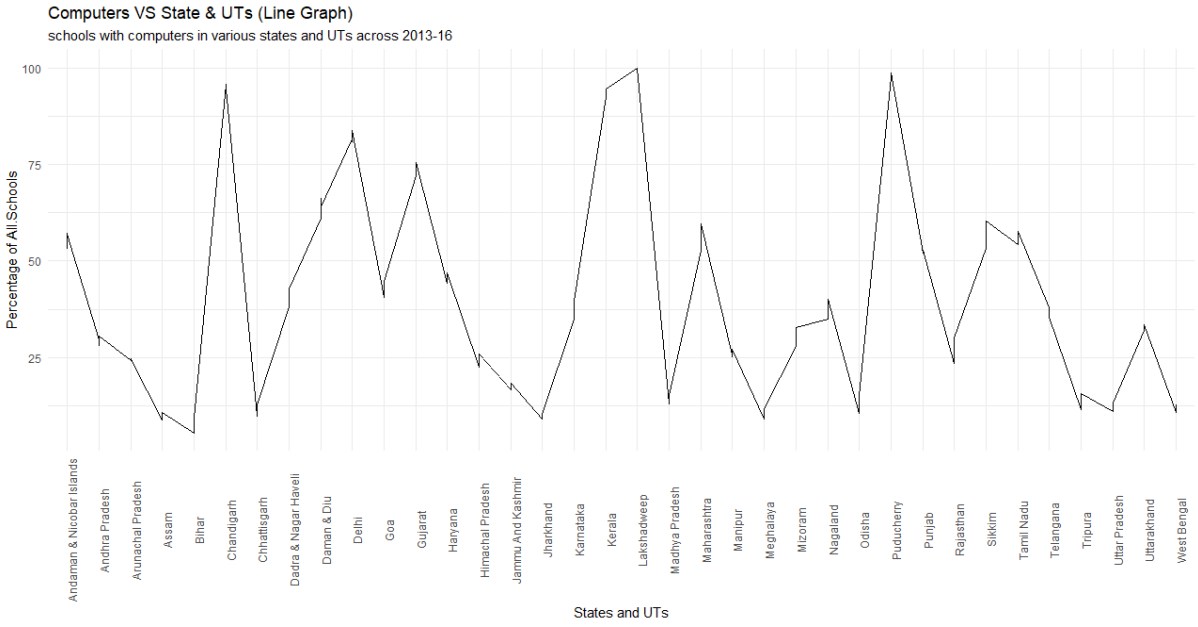
We can get the same inference from the combined graph of all the years i.e. from 2013-16 that Lakshadweep ,followed by Puducherry and Chandigarh are on the top. And we can also get an inference that 24 states with percentage less than 50 needs improvements in schools for educational facilities regarding computers.

2.6 Increase in percentage of All Schools of states during the growing years

2.6.1 Line Chart

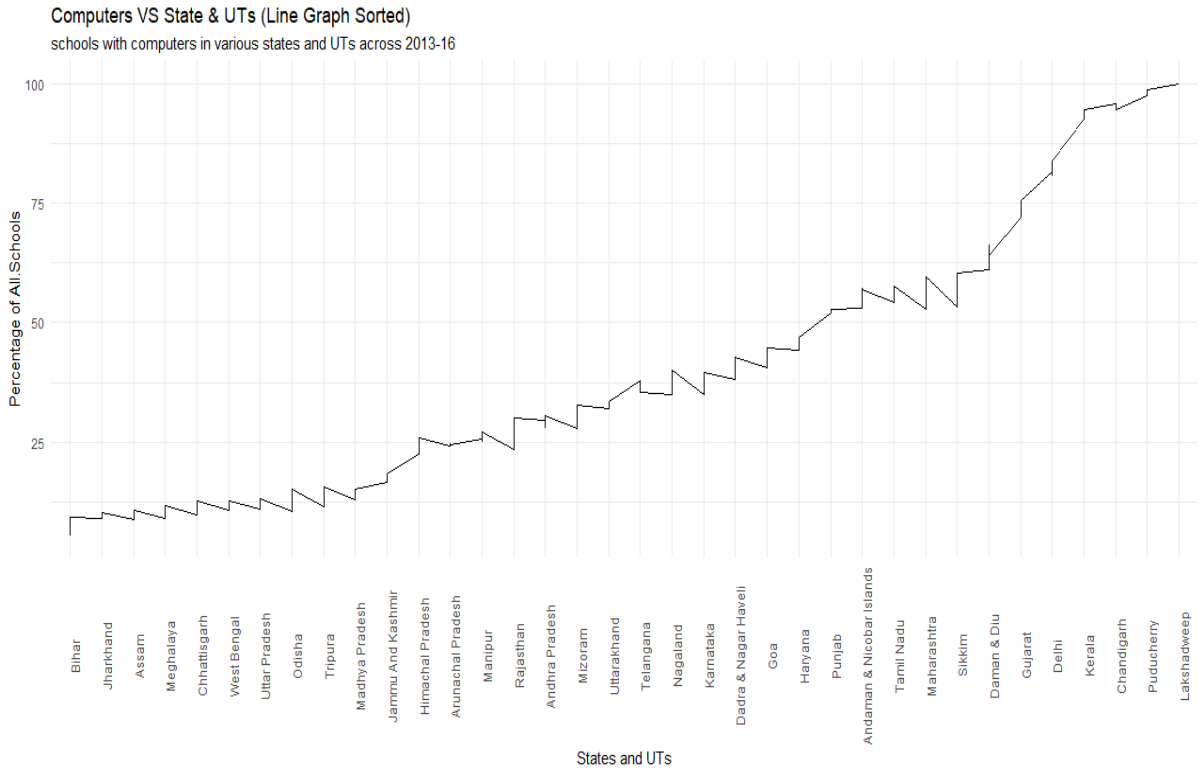
We can perform the same using a line chart.

```
ggplot(schoolStateUT, aes(x =State_UT , y =All.Schools, group=1)) +
 geom_line() +
 labs(x = "States and UTs", y = "Percentage of All.Schools", title = "Computers VS State & UTs (Line Graph)", subtitle = "schools with computers in various states and UTs across 2013-16") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 90))
```

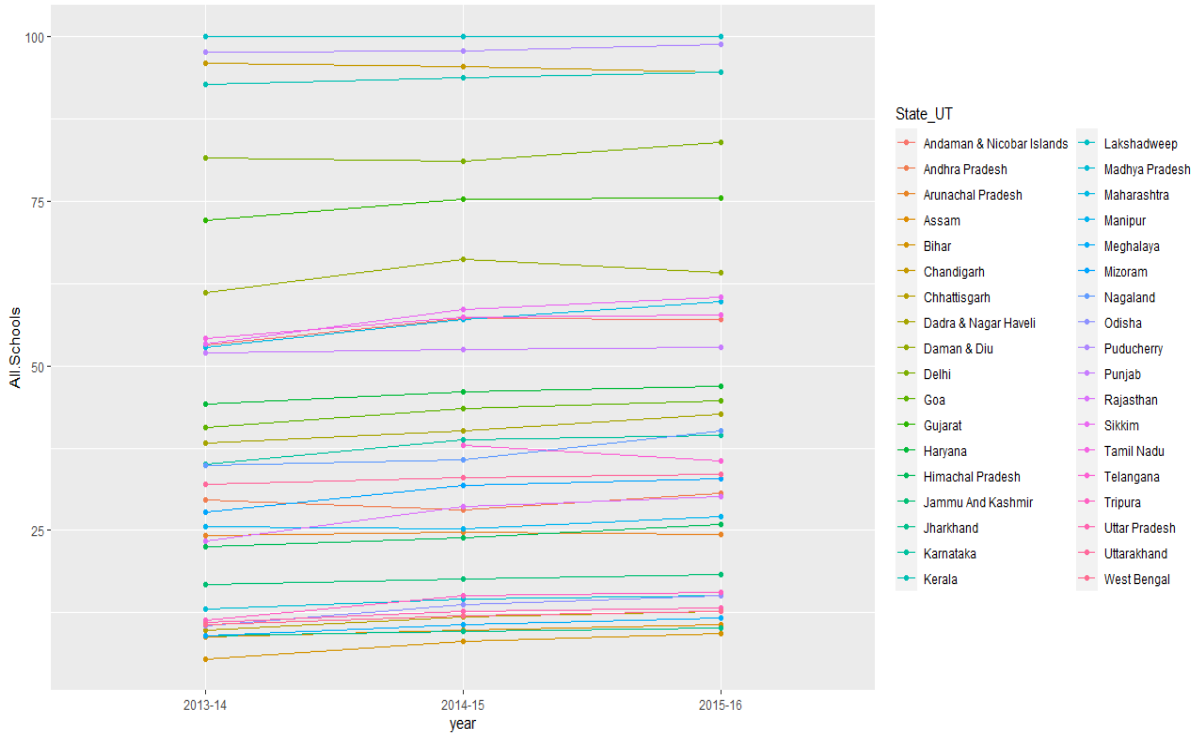


Sorting above graph will give more clearer picture.

```
ggplot(schoolStateUT, aes(reorder(State_UT, All.Schools), All.Schools, group=1)) +
 geom_line() +
 labs(x = "States and UTs", y = "Percentage of All.Schools", title = "Computers VS State & UTs (Line Graph Sorted)", subtitle = "schools with computers in various states and UTs across 2013-16") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 90))
```



We can see the increase or decrease in the slope of the line showing the percentage increment of schools with computers. To see more clearly, Let's see another line graph which we have plotted earlier also.

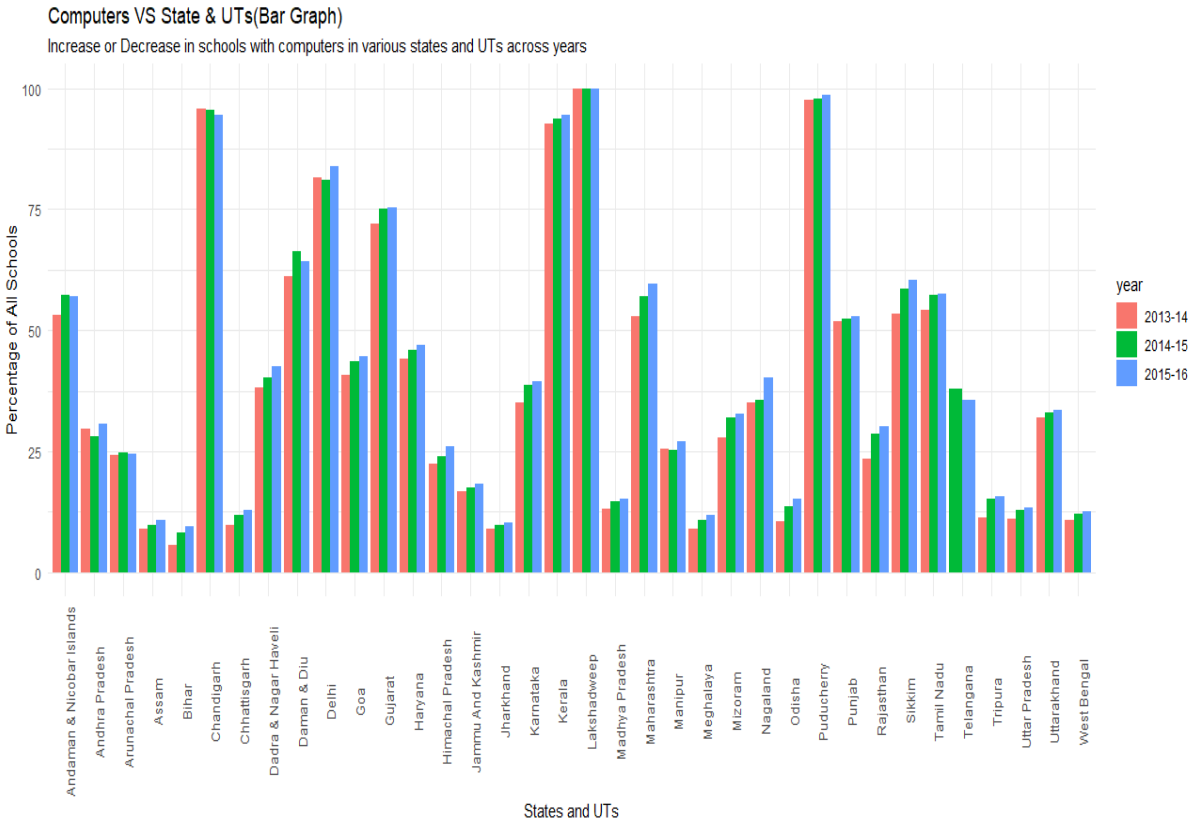


Again, from the above plots, it can be seen that the line graph of most states is increasing. This is a clear indication that the Percentage of schools of most states increased by growing years (in general).

2.6.2 Bar Chart

For more Clearer picture of above observation, bar chart is the best alternative .

```
ggplot(data=schoolStateUT, aes(x=State_UT, y=All.Schools, fill=year)) +
 geom_bar(stat="identity", position=position_dodge())+
 labs(x = "States and UTs", y = "Percentage of All Schools", title = "Computers VS State & UTs(Bar Graph)", subtitle = "Increase or Decrease in schools with computers in various states and UTs across years") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 90))
```



Now, we can more clearly see there is not only increase in percentage of schools but also decrease.For example, from 2013-14 to 2014-15 Andhra Pradesh, Chandigarh, Delhi, and Manipur have shown decrement. And from 2014-15 to 2015-16 Andaman & Nicobar, Arunachal Pradesh, Chandigarh, Daman & Diu and Telangana have shown decrement.

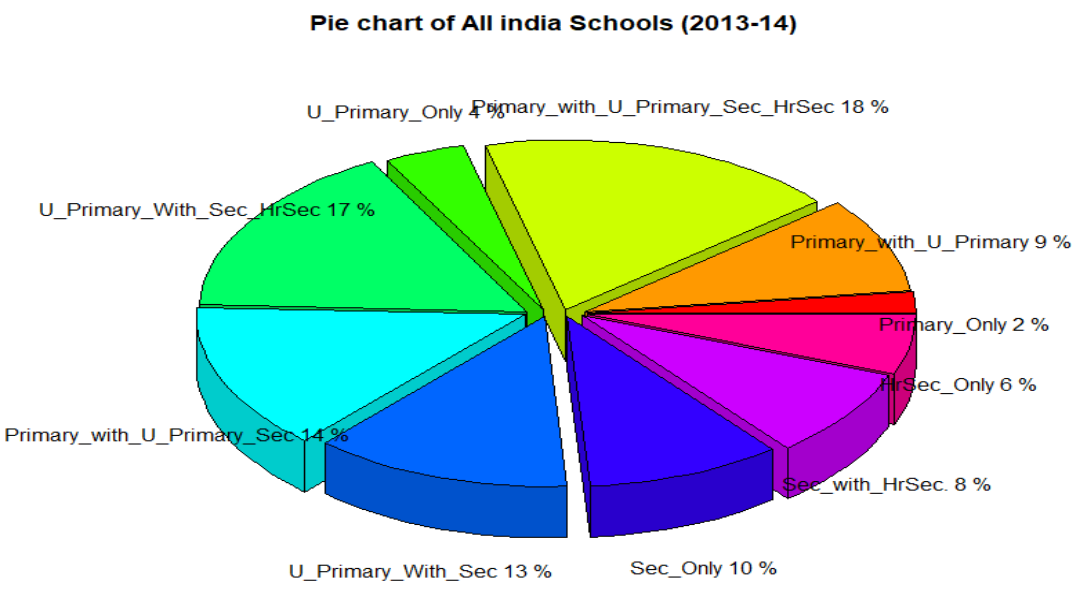
2.7 Contribution of levels of Schools in Overall india

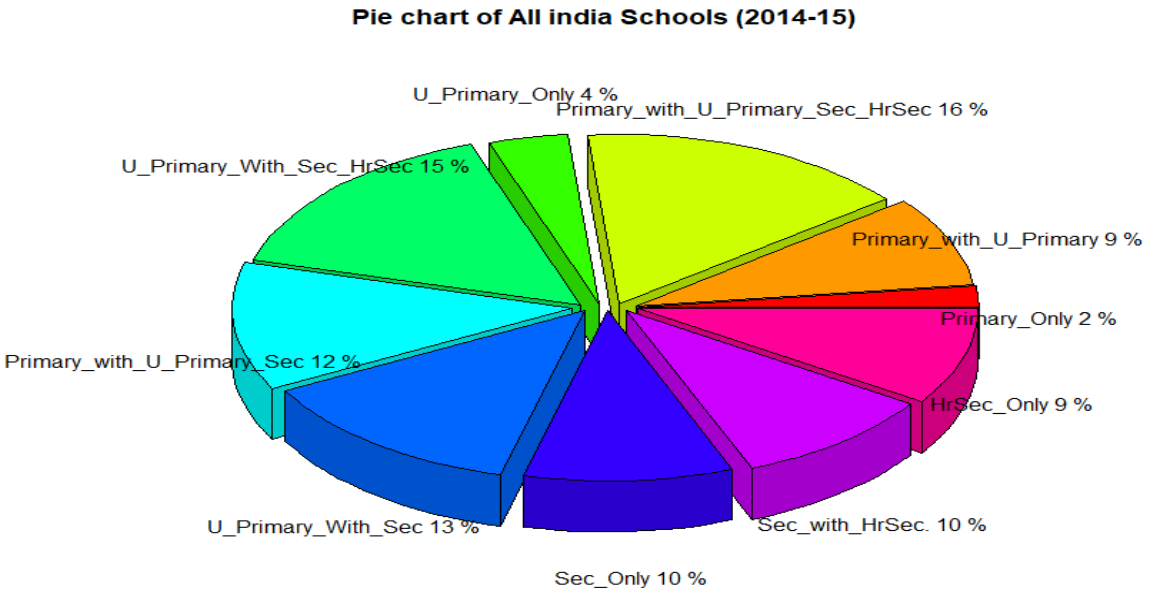
Here we want to analyze the specific School Level of Indian Education System in percentage not particulary in different states but in All India. Also we will observe which level of Schools contributes the most and least for this.

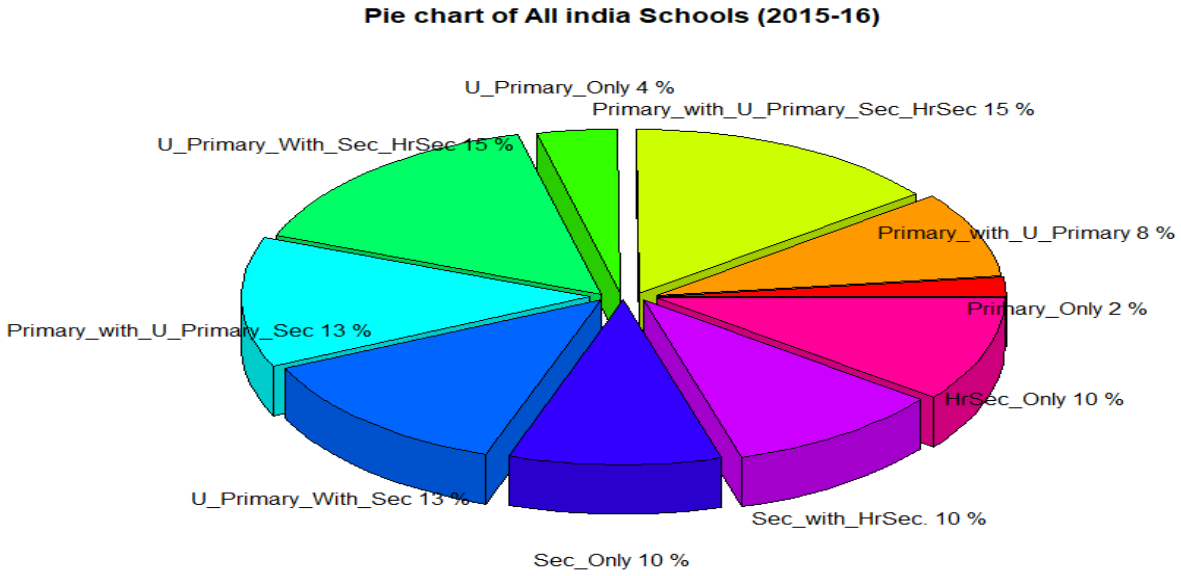
```
AllIndia_Schools <- school[-c(1:107),]
slices = AllIndia_Schools[1,3:12]
lbls=names(AllIndia_Schools[0, 3:12])
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep=" ") # add % to labels
pie3D(x=t(slices),labels=lbls,main="Pie chart of All india Schools (2013-
14)",explode=0.1,theta=pi/3,radius=1,labelcex=1,)
```

```
slices = AllIndia_Schools[2,3:12]
lbls=names(AllIndia_Schools[0, 3:12])
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep=" ") # add % to labels
pie3D(x=t(slices),labels=lbls,main="Pie chart of All india Schools (2014-
15)",explode=0.1,theta=pi/3,radius=1,labelcex=1,)
```

```
slices = AllIndia_Schools[3,3:12]
lbls=names(AllIndia_Schools[0, 3:12])
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep=" ") # add % to labels
pie3D(x=t(slices),labels=lbls,main="Pie chart of All india Schools (2015-
16)",explode=0.1,theta=pi/3,radius=1,labelcex=1,)
```





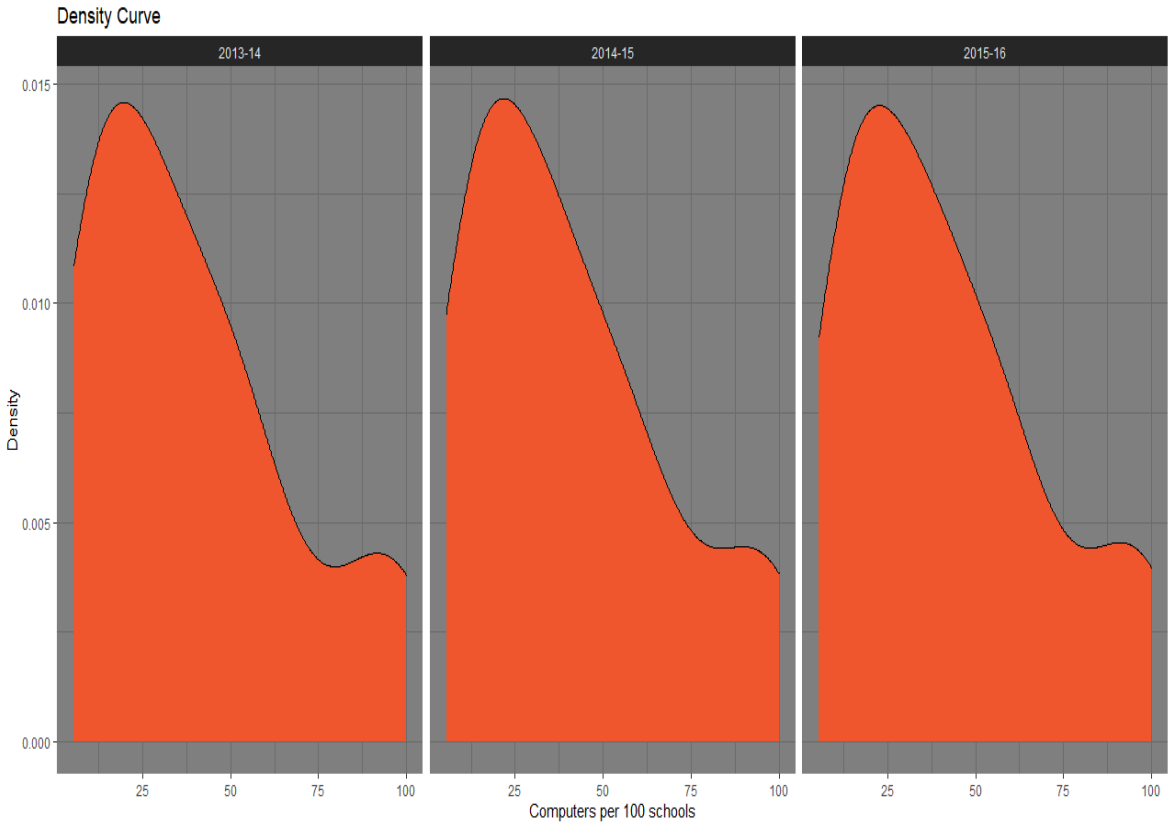


From the above pie-charts, we can observe that ***Primary_with_U_Primary_Sec_HrSec*** contributes the most and ***Primary_Only*** contributes the least during all the years in Schools of overall india.

Also, we can see percentage of schools remain the same for most of the school level but might have decreased in ***Primary_with_U_Primary_Sec_HrSec, U_Primary_with_Sec_HrSec, Primary_with_U_Primary_Sec*** from year 2013-14 to 2014-15 and have increased in ***HrSec_Only, Sec_with_HrSec*** from year 2014-15 to 2015-16. And they have decreased in ***Primary_with_U_Primary_Sec_HrSec, Primary_with_U_Primary*** *and have increased in* ***Primary_with_U_Primary_Sec, HrSec_Only*** from year 2014-15 to 2015-16.

2.8 Density Curve

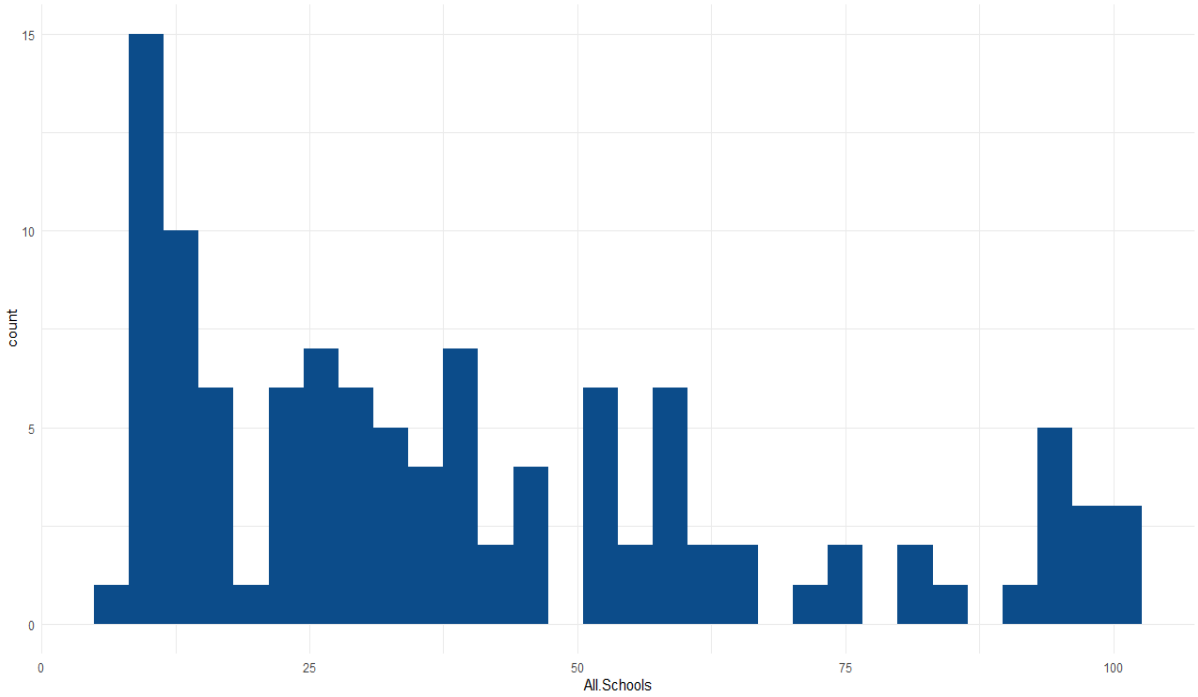
```
ggplot(schoolStateUT) +
aes(x = All.Schools) +
geom_density(adjust = 1L, fill = "#ef562d") +
labs(x = "Computers per 100 schools", y = "Density", title = "Density Curve") +
theme_dark() +
facet_wrap(vars(year))
```



Since the tail is at right side, so the density curve is positive and right skewed suggesting the same about our dataset.

2.9 Skewness and Kurtosis

```
allSchool <- school["All.Schools"]
ggplot(allSchool) +
 aes(x = All.Schools) +
 geom_histogram(bins = 30L, fill = "#0c4c8a") +
theme_minimal()
```




```
print(skewness(allSchool))
All.Schools
0.808019

print(kurtosis(allSchool))
All.Schools
2.592466
```

- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
- As our skewness value is 0.808019 ,it suggests our dataset is moderately skewed and positive. A positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail.
- Kurtosis is a measure of the combined sizes of the two tails. It measures the amount of probability in the tails. The value is often compared to the kurtosis of the normal distribution, which is equal to 3. If the kurtosis is greater than 3, then the dataset has heavier tails than a normal distribution (more in the tails). If the kurtosis is less than 3, then the dataset has lighter tails than a normal distribution (less in the tails).
- As our kurtosis value is 2.592466, it suggests our dataset has lighter tails(platykurtic) and almost near to normal data distribution.