ASUTOSH COLLEGE
Affiliated to Calcutta University

UNIVERSITY OF CALCUTTA • ADVANCEMENT OF LEARNING •

DEPARTMENT OF STATISTICS, ASUTOSH COLLEGE

GROSS ENROLLEMENT RATIO OF SCHOOL EDUCATION IN INDIA (HIGHER SECONDARY LEVEL)

-A STATISTICAL ANALYSIS

NAME: SOURAJ CHAKRABORTY

SEMESTER VI

COLLEGE ROLL- 922

C.U. ROLL- 203012-21-0109

C.U. REGISTRATION NO.- 012-1111-0836-20

STREAM: BSC STATISTICS HONS.

CONTENTS:

Topics	Page no
1. Executive summary	5
2. Introduction	6
3. Collected data sets	7-8
4. Methodology	9-15
5. Analysis and Results	16-24
6. Conclusion	25
7. References	26
8. Acknowledgement	27

EXECUTIVE SUMMARY:

In this project, I have attempted to study and analyse the GER of XI-XII students (male, female and total) in a state-wise comparison of sessions 2009-2010 and 2019-2020 to see whether it has increased/decreased over a decade. I've also taken year-wise total GER for 14 years to compute the trend.

At first I have compared state-wise GER by exploratory data analysis with the help of some statistical diagrams like bar diagrams, pie charts.

In the next part I have used regression method to study the underlying pattern of GER of whole India over time and analyse the trend. Then I have used time series modelling to fit a suitable trend line equation for the data and tried to predict the future GER of a recent session through the estimated trend line equation and compared it with the actual data to see if it varies or not.

INTRODUCTION:

Gross Enrolment Ratio (GER) is a statistical measure used in the education sector, and formerly by the UN in its Education Index, to determine the number of students enrolled in school at several different grade levels (like elementary, middle school and high school), and use it to show the ratio of the number of students who live in that country to those who qualify for the particular grade level. The United Nations Educational, Scientific and Cultural Organization (UNESCO), describes "Gross Enrolment Ratio" as the total enrolment within a country "in a specific level of education, regardless of age, expressed as a percentage of the population in the official age group corresponding to this level of education".

The motive of a nation is to increase GER to 100% (at least in school levels) over all the states so that every individual can be ensured the right to education both in urban and rural areas.

FORMULA FOR GER(higher secondary level):

$$GER = \frac{total\ no.of\ students\ enrolled}{total\ population\ of\ that\ official\ age\ group} \times 100$$

COLLECTED DATA SETS:

STATEWISE GROSS ENROLMENT RATIO of XI-XII students (2009-2010 and 2019-2020)							
Sl no	**state/UT**	**Boys (2009-2010)**	**Girls (2009-2010)**	**Total GER (2009-2010)**	**Boys (2019-20)**	**Girls (2019-20)**	**Total GER (2019-2020)**
1	Andhra Pradesh	47.47	40.36	44.00	48.9	55.7	52.2
2	Arunachal Pradesh	44.43	41.48	43.02	33.0	38.7	35.8
3	Assam	13.62	12.61	13.13	30.4	31.4	30.9
4	Bihar	17.38	13.33	15.47	30.6	31.0	30.8
5	Chhattisgarh	32.94	25.52	29.30	49.7	59.1	54.3
6	Goa	46.33	50.86	48.50	63.8	73.8	68.6
7	Gujarat	38.42	32.55	35.66	43.5	42.8	43.2
8	Haryana	59.67	59.84	59.75	60.6	63.1	61.8
9	Himachal Pradesh	68.58	70.06	69.29	80.3	87.0	83.4
10	Jammu & Kashmir	42.95	40.72	41.88	38.4	38.4	38.4
11	Jharkhand	7.39	5.63	6.54	40.5	41.1	40.8
12	Karnataka	41.63	43.38	42.48	48.0	56.5	52.1
13	Kerala	45.57	55.34	50.41	79.1	87.1	83.0
14	Madhya Pradesh	47.23	31.07	39.51	45.2	44.8	45.0
15	Maharashtra	59.63	50.62	55.34	67.2	66.7	67.0
16	Manipur	36.21	27.00	31.53	57.7	56.5	57.1
17	Meghalaya	10.10	12.45	11.25	38.9	47.9	43.4
18	Mizoram	38.39	37.74	38.06	51.2	56.3	53.7
19	Nagaland	18.26	16.74	17.53	32.6	35.5	34.0
20	Odisha	25.03	20.44	22.78	46.7	48.4	47.6
21	Punjab	41.96	44.56	43.15	69.5	73.5	71.3
22	Rajasthan	43.08	26.42	35.19	62.4	54.1	58.5
23	Sikkim	27.57	29.56	28.56	48.0	59.7	53.7
24	Tamil Nadu	44.81	54.69	49.62	66.3	80.6	73.2
25	Tripura	28.92	23.26	26.19	39.5	41.9	40.7
26	Uttar Pradesh	36.47	30.90	33.85	48.2	45.4	46.9
27	Uttarakhand	57.53	54.13	55.90	66.2	73.0	69.3
28	West Bengal	30.43	25.04	27.81	50.1	60.6	55.2
29	Andaman and Nicobar	43.14	50.34	46.51	38.4	52.8	45.1
30	Chandigarh	53.33	61.57	56.72	72.4	85.8	78.0
31	D&N Haveli	32.20	28.25	30.43	48.7	68.0	56.9
32	Daman & Diu	28.94	42.16	34.03	26.1	55.6	36.2
33	Delhi	59.14	58.33	58.76	69.0	77.5	72.8
34	Lakshadweep	79.02	77.97	78.49	66.0	69.0	67.6
35	Puducherry	51.76	65.68	58.40	63.1	76.1	69.3

YEARWISE TOTAL GROSS ENROLMENT RATIO				
YEAR	MALE	FEMALE	TOTAL	session no.
2004-2005	30.8	24.5	27.8	1
2005-2006	31.4	25.2	28.5	2
2006-2007	31.5	26.1	28.9	3
2007-2008	36.3	30.4	33.5	4
2008-2009	37.5	31.6	34.5	5
2009-2010	38.5	33.5	36.1	6
2010-2011	42.3	36.2	39.4	7
2011-2012	47.6	43.9	45.9	8
2012-2013	41.9	39.5	40.8	9
2013-2014	52.8	51.6	52.2	10
2014-2015	54.6	53.8	54.2	11
2015-2016	56	56.4	56.2	12

METHODOLOGY:

Comparison by Diagrammatic Representation:

The use of diagrams to illustrate statistical data is very essential. The greatest way for representing any numerical data obtained in statistics is through diagrammatic representations. "A picture is worth a thousand words," according to one famous quote. In comparison to tabular or textual representations, the diagrammatic display of data provides an immediate understanding of the true scenario to be defined by the data.

It efficiently converts the exceedingly complex ideas contained in numbers into a more concrete and readily understandable form. Although diagrams are less certain, they are far more efficient in displaying data than tables. There are numerous types of diagrams in common use. Similarly, the diagrammatic representation of data gives a lot of information regarding the numerical data.

- PIE CHART:
  A pie chart is another appropriate diagram used for exhibiting the relative sizes of the different parts of a whole. In this case, a circle is partitioned into several sectors by drawing angles at the centre, the area of each sector indicating the corresponding percentage. In fact, the area enclosed by the circle is regarded as 100. Since the total angle at the centre is 360o, the desired angle for some particular category will be 3.6 times the relevant percentage. The diagram, thus constructed, is termed as pie diagram.

- BAR DIAGRAM:
  Another mode of diagrammatic representation of data is the use of bar diagrams. These have more general applicability than line diagrams in the sense that they may be used for series varying either over time or over space. In this method bars of equal width are taken for the different items of the series, drawn over base line. The length or height of a bar representing the value of the variable concerned. It is preferable to take the bars horizontally for data varying over space and vertically in the case of a series varying over time. We can compare the different items of the series by visualizing bar diagram.

- DIVIDED BAR DIAGRAM:
  A divided bar diagram is a way of representing data in which the total magnitude is divided into different segments. In this diagram, first of all, we draw the simple bars for each class taking the total magnitude in that class then we divide that bar into segments of its various components.

- LINE DIAGRAM:
  This diagram is meant for representing chronological data. In fact, it exhibits the relationship of the variable (e.g. sales of coffee of a company, productions of a crop) may be specified for individual points of time or for different period of time. In constructing a line diagram, two axis of co-ordinates are taken, the horizontal one for time and the vertical one for variable. The scale for each axis is then selected and the data are plotted as different points on the plane, the plotting of variable values being done against points of time or mid-points of the time interval (for time period). The successive points are now joined by straight line segments and the chart so obtained is called a line diagram for the given data. Two or mutually related time series data having same unit of measurement can be represented using the same axis of co-ordinates, by drawing a number of line diagrams, one for each series. These different line diagrams are mutually distinguished by using distinct pattern of lines such as broken lines, dotted lines or multiple coloured lines. The resulting diagram is known as a Multiple Line Diagram. It is used for comparing two mutually related time series data e.g. if we want to compare the literacy rates for a number countries last 15 years, say, we may draw multiple line diagram.

FORMULA FOR GER(higher secondary level):

$$GER = \frac{total\ no.of\ students\ enrolled}{total\ population\ of\ that\ official\ age\ group} \times 100$$

□ **Mathematical Curve Fitting:**

It is an essential part of the concept of trend that the movement over fairly long periods is smooth. This is perhaps the best and most rational method of determining the trend. In this case, a suitable trend equation is selected and then the constants involved in the equation are estimated on the basis of the data in hand. After derived the estimated trend equation we can predict the future trend values.

Polynomial Trend Line:

Here we assume that our suitable trend equation is a polynomial in time element 't'. So here we consider a mathematical model:-

$$Y_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots.. + a_p t^p + e_t$$

where a_i's, i = 1(1)p are constants and e_t denotes a random error with

$E(e_t) = 0$ and $V(e_t) = \sigma^2$

Now we can estimate the constants a_i's by method of least squares and fit the polynomial model. In this method the constants are determined by minimizing,

$$S = \sum_t \left(y_t - a_0 - a_1 t - a_2 t^2 - \cdots - a_p t^p\right)^2$$

The normal equations are,

$$\frac{\partial S}{\partial a_j} = 0, \forall\, j = 0(1)p$$

$$\Rightarrow \sum_t t^j y_t = a_0 \sum_t t^j + a_1 \sum_t t^{j+1} + a_2 \sum_t t^{j+2} + \cdots + a_p \sum_t t^{j+p}, j = 0(1)p$$

By solving the normal equations we can obtain the estimates of the constants $a_0, a_1, a_2, \ldots, a_p$, and fit a trend equation on the time element 't'.

□ Significance Test:

Consider the regression model

$$Y_t = a_0 + a_1t_i + a_2t_i^2 + a_3t_i^3 + \ldots.. + a_pt_i^p + e_{ti}\,;\ i = 1(1)n$$

we assume that $e_{ti} \sim$ iid $N(0, \sigma^2)$

So here are p covariates t, t^2, t^3,...,t^p, the response variable Yt and let there are n observations of pair (t, Y_t). Now for simplification we define the covariates as

$t_1 = t$, $t_2 = t^2$, $t_3 = t^3$,...,$t_p = t^p$; response $Y_t = y$ and error $e_t = e$. Therefore the model becomes,

$$y_i = a_0 + a_1t_{1i} + a_2t_{2i} + a_3t_{3i} + \ldots.. + a_pt_{pi} + e_i\,;\ i = 1(1)n \text{ where } e_i \sim \text{iid } N(0, \sigma^2)$$

Now we want to test whether the covariate has any significant effect on response variable or not. So, our testing problem will be,

$H_0 : a_j = 0 \ \forall\ j = 1,2,\ldots,p$

vs.

$H_1 : a_j \neq 0$ for at least one j

Now, the unrestricted sum square error (SSE) will be,

$$SSE = \min_{a_i} \sum_{i=1}^{n} \left(y_i - a_0 - a_1t_{1i} - a_2t_{2i} - a_3t_{3i} + \cdots - a_pt_{pi}\right)^2$$

For the sake of simplicity, we rewrite the model as,

$$y_i = a_0 + a_1\acute{t}_{1i} + a_2\acute{t}_{2i} + a_3\acute{t}_{3i} + \ldots.. + a_p\acute{t}_{pi} + e_i$$

where, $\acute{t}_{ji} = t_{ji} - \bar{t}_j \ \forall\ j = 1,2,\ldots,p$

$$i.e., SSE = \sum_{i=1}^{n} \left(y_i - a_0 - a_1\acute{t}_{1i} - a_2\acute{t}_{2i} - a_3\acute{t}_{3i} - \cdots - a_p\acute{t}_{pi}\right)^2$$

$$\therefore \frac{\partial}{\partial a_0}\sum_{i=1}^{n}\left(\mathrm{y_i} - \mathrm{a_0} - \mathrm{a_1}\acute{\mathrm{t}}_{1i} - \mathrm{a_2}\acute{\mathrm{t}}_{2i} - \mathrm{a_3}\acute{\mathrm{t}}_{3i} - \cdots - \mathrm{a_p}\acute{\mathrm{t}}_{pi}\right)^2 = 0$$

$$\sum_{i=1}^{n}\left(\mathrm{y_i} - \mathrm{a_0} - \mathrm{a_1}\acute{\mathrm{t}}_{1i} - \mathrm{a_2}\acute{\mathrm{t}}_{2i} - \mathrm{a_3}\acute{\mathrm{t}}_{3i} - \cdots - \mathrm{a_p}\acute{\mathrm{t}}_{pi}\right) = 0$$

$$\sum_{i=1}^{n} y_i = na_0 + a_1\sum_{i=1}^{n}\acute{t}_{1i} + a_2\sum_{i=1}^{n}\acute{t}_{2i} + \cdots + a_p\sum_{i=1}^{n}\acute{t}_{pi}$$

$$Now, \sum_{i=1}^{n}\acute{t}_{ji} = \sum_{i=1}^{n}(t_{ji} - \bar{t}_j) = 0, \forall\, j = 1(1)p$$

$$\Rightarrow a_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\Rightarrow \hat{a}_0 = \bar{y}$$

$$\frac{\partial}{\partial a_j}\sum_{i=1}^{n}\left(\mathrm{y_i} - \mathrm{a_0} - \mathrm{a_1}\acute{\mathrm{t}}_{1i} - \mathrm{a_2}\acute{\mathrm{t}}_{2i} - \mathrm{a_3}\acute{\mathrm{t}}_{3i} - \cdots - \mathrm{a_p}\acute{\mathrm{t}}_{pi}\right)^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(\mathrm{y_i} - \mathrm{a_0} - \mathrm{a_1}\acute{\mathrm{t}}_{1i} + \mathrm{a_2}\acute{\mathrm{t}}_{2i} - \mathrm{a_3}\acute{\mathrm{t}}_{3i} - \cdots - \mathrm{a_p}\acute{\mathrm{t}}_{pi}\right)\acute{t}_{ji} = 0$$

We can solve the above normal equations and obtain the value of

$\hat{\mathrm{t}}_1, \hat{\mathrm{t}}_2, \hat{\mathrm{t}}_3, \ldots, \hat{\mathrm{t}}_\mathrm{p}$

$$i.e.\ SSE = \sum_{i=1}^{n}\left(\mathrm{y_i} - \hat{\mathrm{a}}_0 - \hat{\mathrm{a}}_1\mathrm{t}_{1i} - \hat{\mathrm{a}}_2\mathrm{t}_{2i} - \hat{\mathrm{a}}_3\mathrm{t}_{3i} - \cdots - \hat{\mathrm{a}}_\mathrm{p}\mathrm{t}_{pi}\right)^2\ with\ df = n - k - 1$$

$$Now, under\ H_0, a_j = 0\ \forall\, j\ and\ SSE_{H_0} = \min_{a_0}\sum_{i=1}^{n}(y_i - a_0)^2$$

Here, $\hat{\mathrm{a}}_0^{\mathrm{H0}} = \bar{\mathrm{y}}$ [SD is the least RMSD]

$$i.e.\ SSE_{H_0} = \sum_{i=1}^{n}(y_i - \bar{y})^2\ with\ df = n - 1$$

$$\therefore \frac{SSE_{H_0} - SSE}{\sigma^2} \sim \chi^2_{n-1-(n-k-1)} \equiv \chi^2_k\ \ and\ \ \frac{SSE}{\sigma^2} \sim \chi^2_{n-k-1}$$

Also it can be shown that SSE_{H0} – SSE is independent of SSE

$$F = \frac{\dfrac{\dfrac{SSE_{H_0} - SSE}{\sigma^2}}{k}}{\dfrac{\dfrac{SSE}{\sigma^2}}{n-k-1}} \sim F_{k,n-k-1}$$

So, we reject H_0 at level α if $F_{obs} > F_{\alpha;k,n-k-1}$

In case of p-value if our p-value i.e. $P(F>F_{obs}) < \alpha$ we reject our null hypothesis i.e. $a_j = 0\ \forall\ j = 1,2,\ldots,p$

So if our p-value < α we can say the covariates have significant effect on response and in that case our model is statistically significant.

R^2 – Coefficient of Determination and Adjusted R Square

In statistics the coefficient of determination denoted by R^2 or r^2 and pronounced "R squared", is the proportion of the variation in the dependent variable that is predicted from the independent variable(s).

Definition:-
A data set has n values marked $y_1, y_2, \ldots y_n$(denoted as y_i) each associated with a fitted(or modelled, or predicted) value $f_1, f_2, \ldots f_n$(denoted as f_i or sometimes $\hat{y}_i$)

Define the residuals as $e_i = y_i - f_i$

If $\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Then the variability of the dataset can be measured with two sum of squares formulas:

The total sum of squares (proportional to variance of the data):

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The sum of squares of residuals, also called residual sum of squares:

$$SS_{res} = \sum_{i=1}^{n} (y_i - f_i)^2 = \sum_{i=1}^{n} e_i^2$$

Then the most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression it measures how well the regression predictions approximate the real data points. The value of R^2 lies between 0 and 1. More the value of R^2 close to 1 indicates a better fit of model. R^2 of 1 indicates that the regression predictions perfectly fit the data.

Adjusted R^2

However each time we add a new predictor variable to the model the R- squared is guarantee to increase even if the predictor variable isn't useful.

The adjusted R-squared or adjusted R^2 is a modified version of R^2 that adjusts for the number of predictors in a regression model. It is calculated as

$$1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

R^2 : The coefficient of determination of the model
n : The number of observations
k : The number of predicted variables

Since R^2 always increases as you add more predictors to a model, adjusted R^2 can serve as a metric that tells you how useful a model is, adjusted for the number of predictors in a model. Therefore the adjusted R^2 tells us the percentage of variation explained by only the independent variables that actually affect the dependent variable. Same as R^2, the value of adjusted R^2 lies between 0 and 1. More the value of adjusted R^2 close to 1 indicates that all the predictor variables in model have better significant effects.

• Analysis and Results:-

At first, to study the state wise GER of India I've used diagrammatic representation of male, female and total students of the sessions 2009-2010 and 2019-2020 in details. Then I have used graphical representation of year wise total GER of India of 12 sessions and fitted a linear model to the time series data.

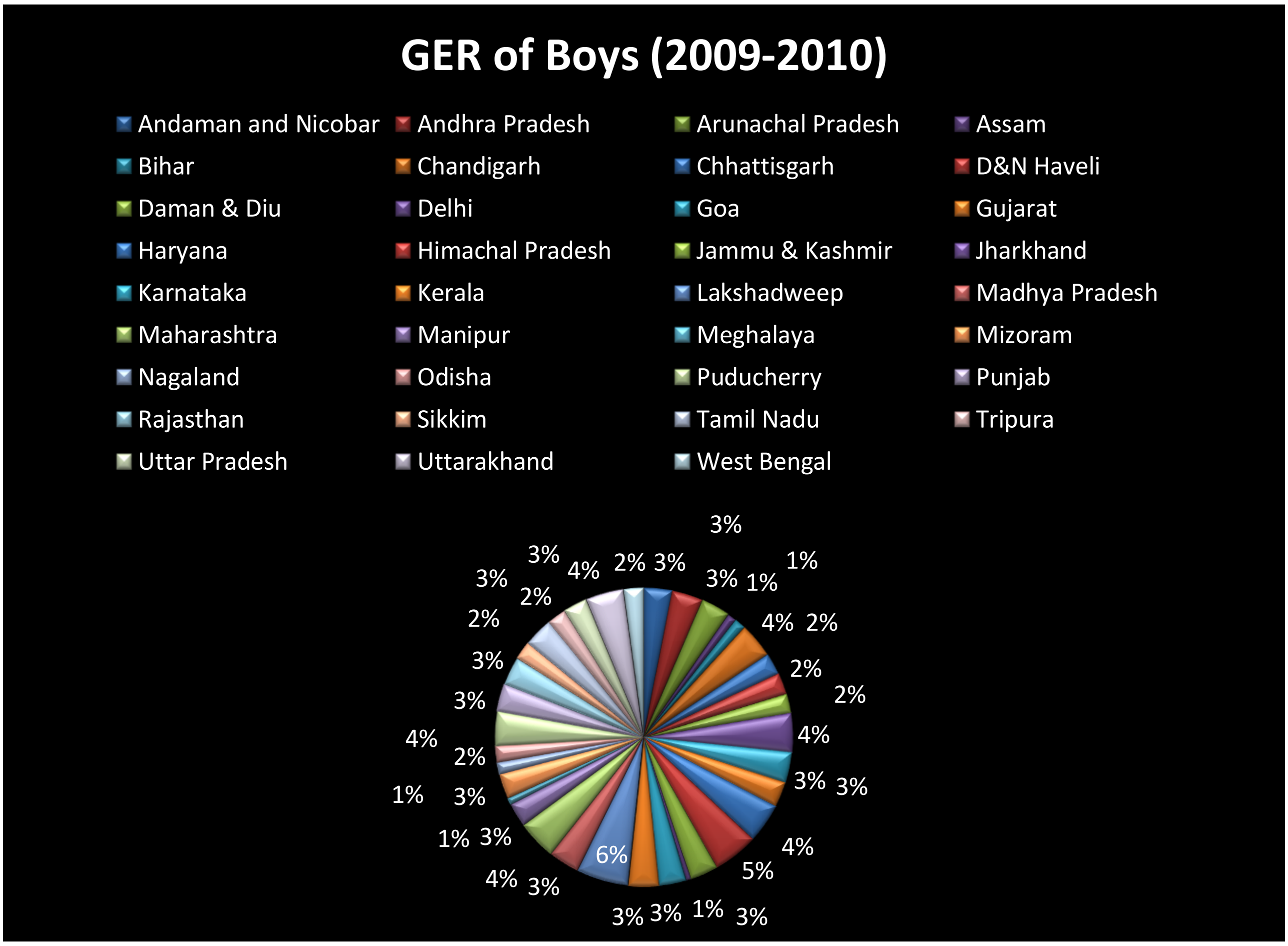
□ **Studying and Comparing Data Using Their Diagrammatic Representation:-**

For comparing the GER of different states if India, at first pie charts of GER of boys of 2009-2010 and 2019-2020 are drawn separately in Fig 1. The same is drawn for GER of girls for the aforesaid 2 sessions in Fig 2.

Next, multiple vertical bar diagrams are drawn to compare GER of 2 sessions for boys in Fig 3 and the same for girls in Fig 4.

Lastly, Multiple vertical bar diagram and multiple line diagrams are drawn for the total state wise GER for the 2 sessions in Fig 5 and Fig 6 respectively.

FIGURE 1: PIE CHARTS OF STATEWISE GER OF BOYS (2009-2010 AND 2019-2020)



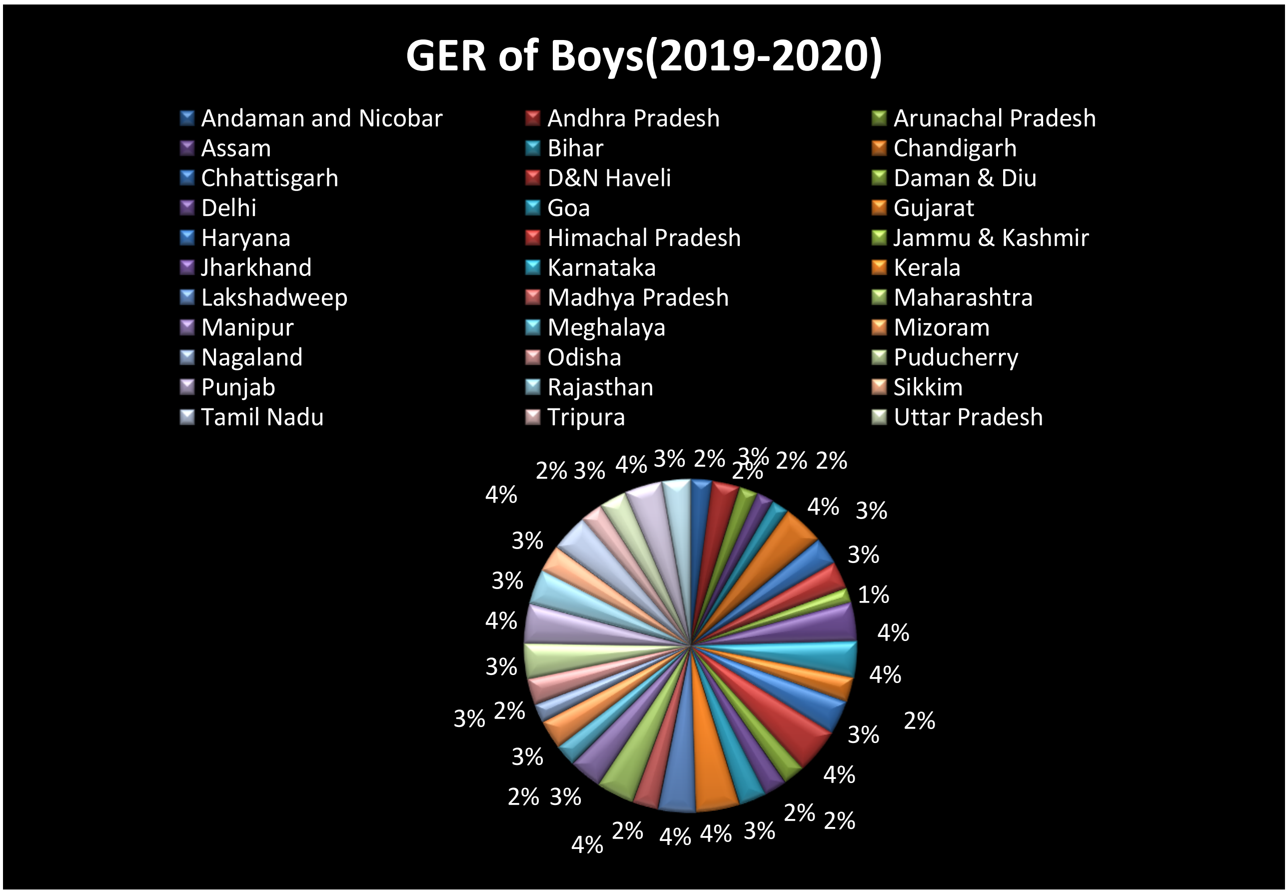
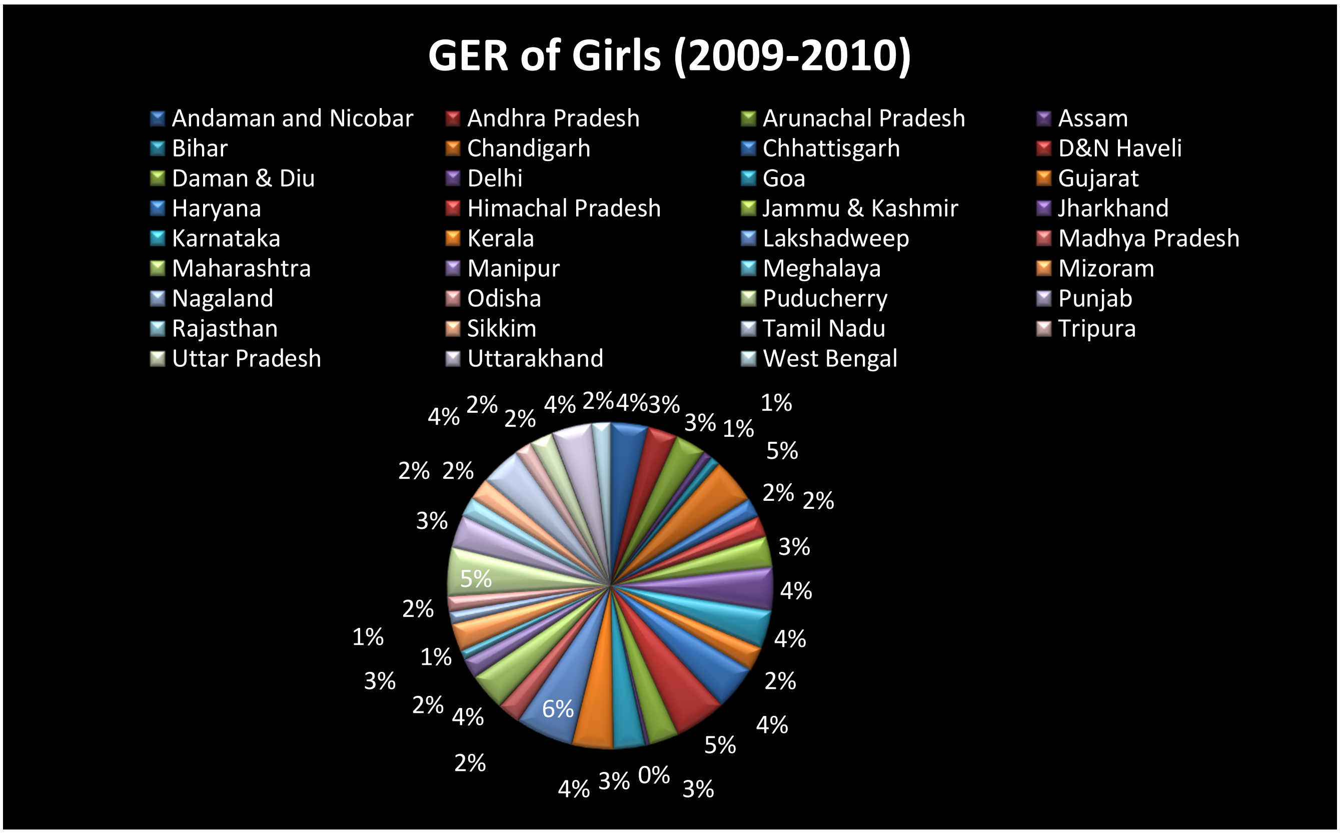


FIGURE 2: PIE CHARTS OF STATEWISE GER OF GIRLS (2009-2010 AND 2019-2020)



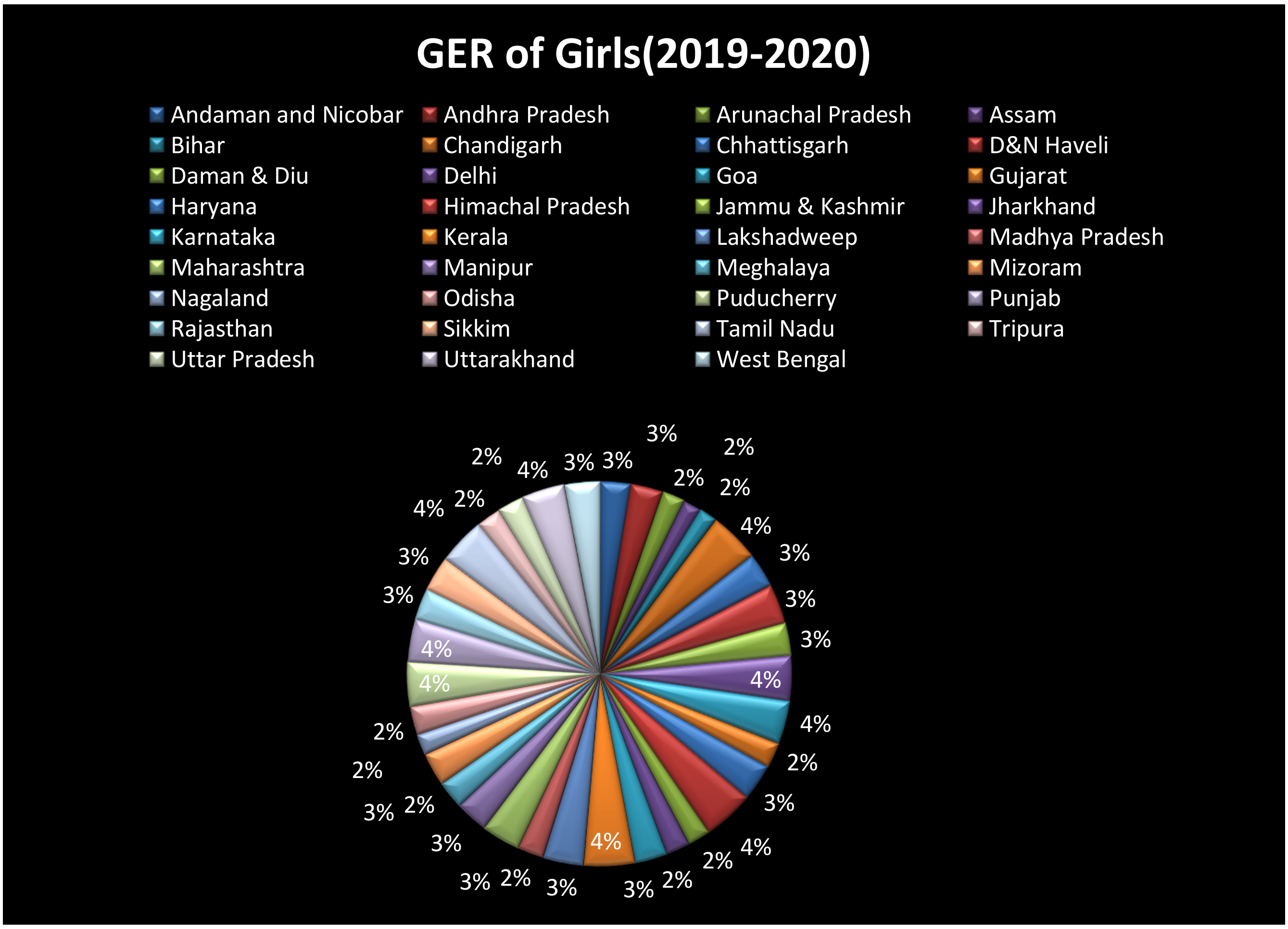


FIGURE 3: VERTICAL BAR DIAGRAM OF GER OF BOYS

(2009-2010 AND 2019-2020)

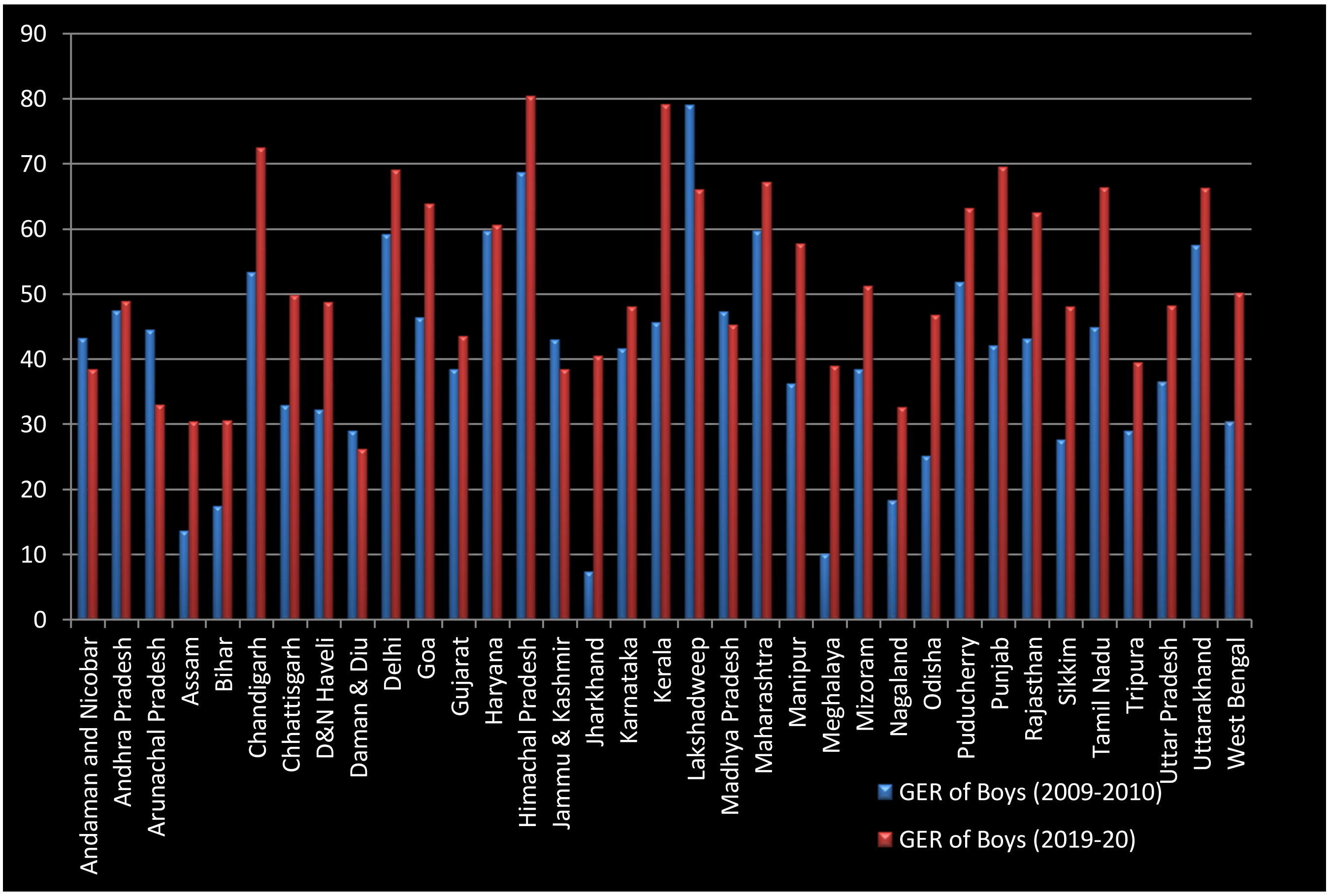


FIGURE 4: VERTICAL BAR DIAGRAM OF GER OF GIRLS (2009-2010 AND 2019-2020)

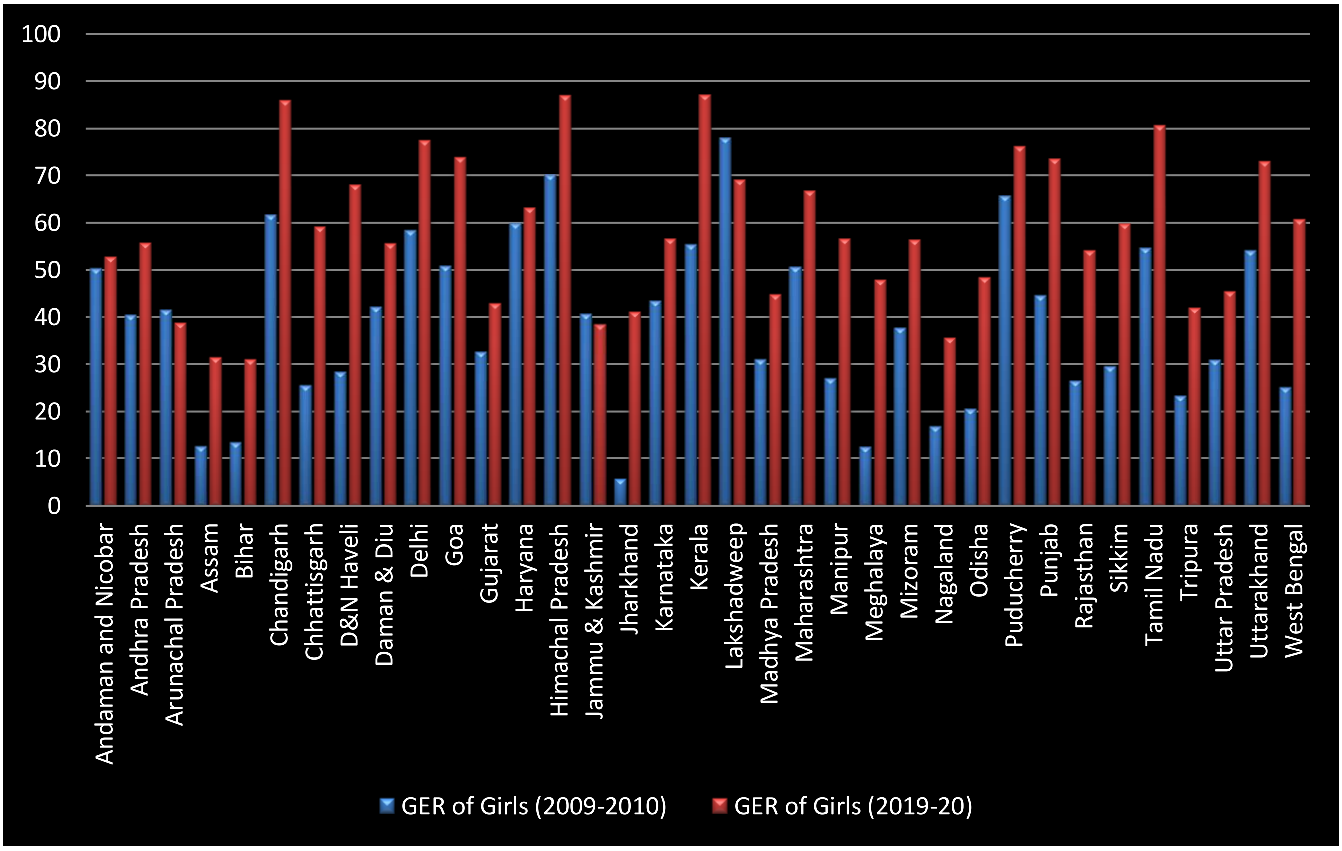


FIGURE 5: VERTICAL BAR DIAGRAM OF TOTAL GER OF XI-XII STUDENTS (2009-2010 AND 2019-2020)

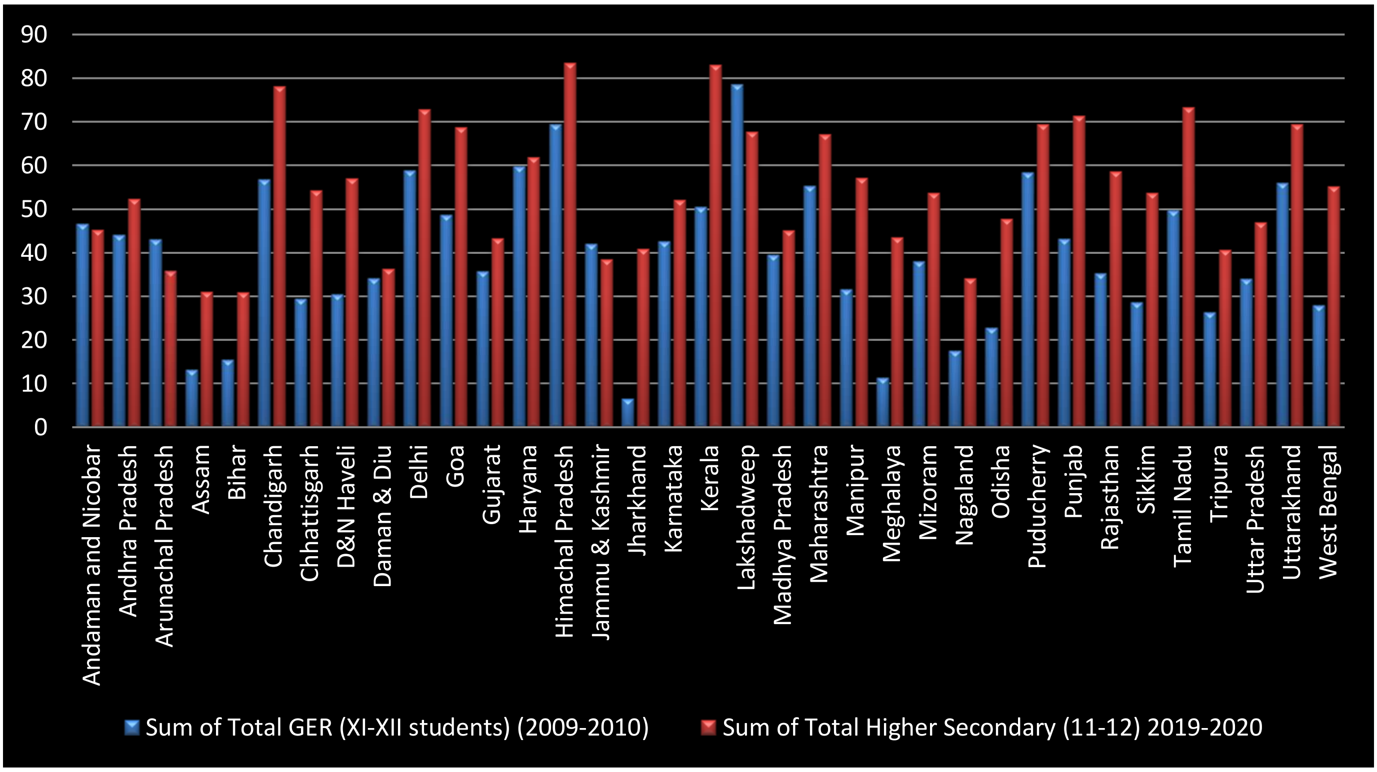
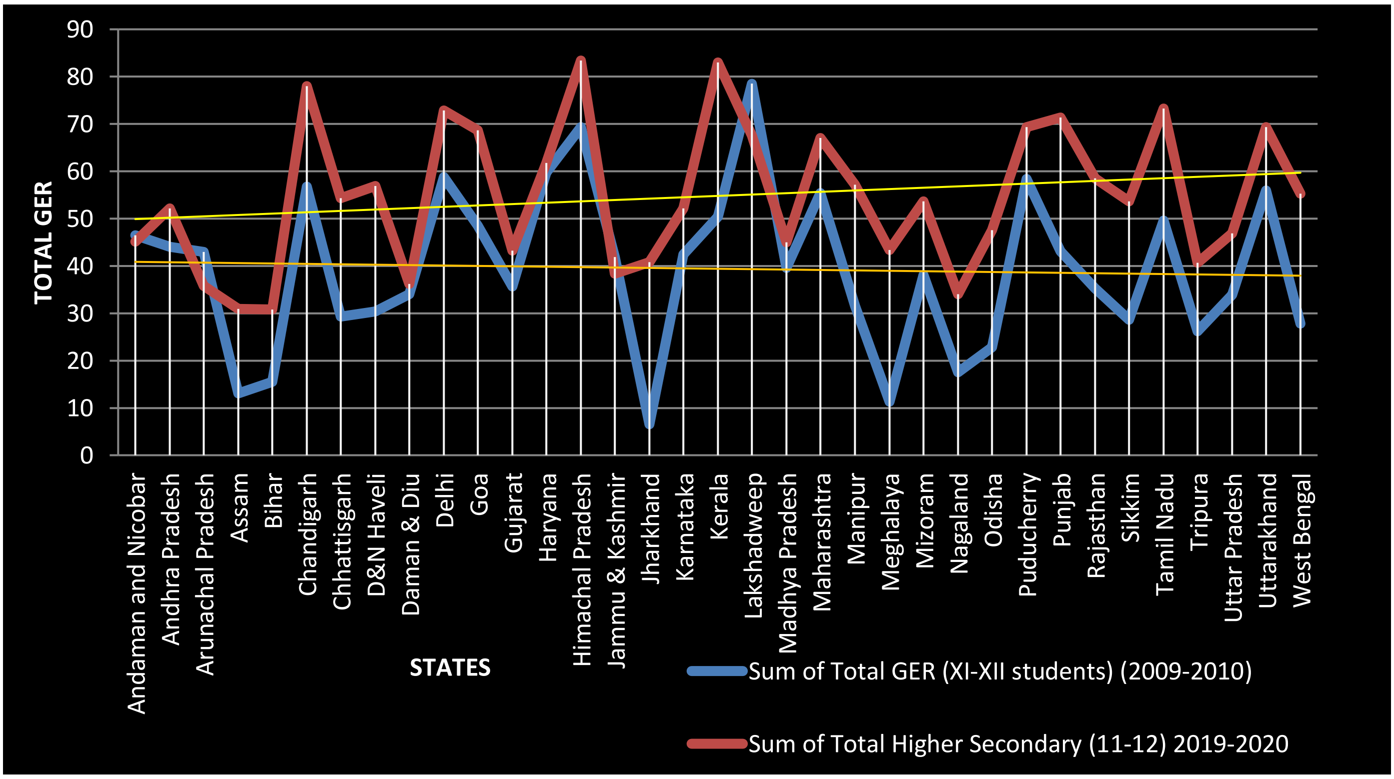


FIGURE 6: MULTIPLE LINE DIAGRAM OF TOTAL GER OF XI-XII STUDENTS (2009-2010 AND 2019-2020)



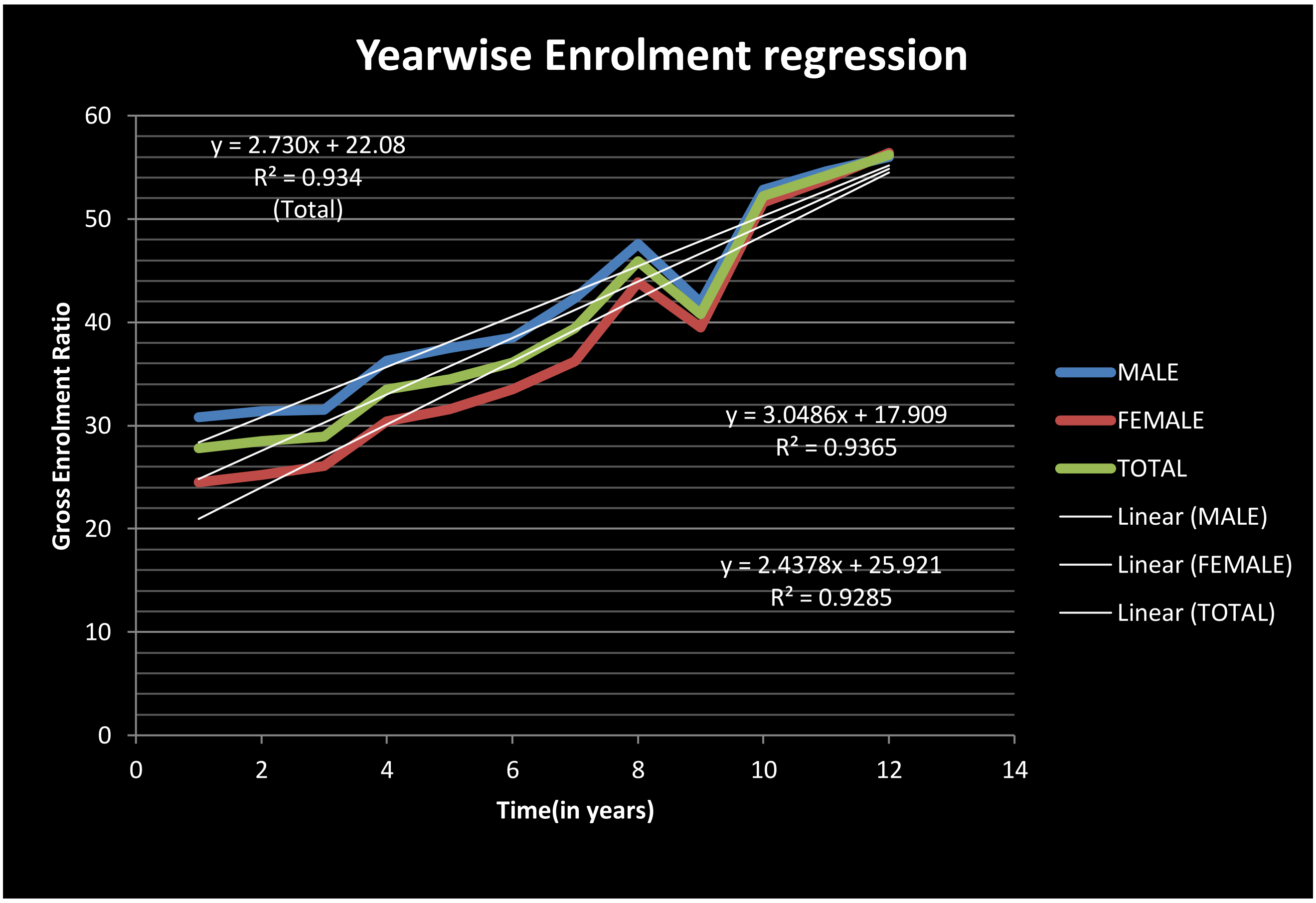
INTERPRETATION OF DATA FROM THE DIAGRAMS:

From Fig. 3, it can be seen that the GER of boys has increased in most states over a decade i.e. from 2009-2010 to 2019-2020. For instance it has increased significantly from 2009-2010 to 2019-2020 in Assam, Bihar, Chandigarh, Chhattisgarh, Jharkhand, Kerala, Manipur, Meghalaya, Odisha, Punjab, Tamil Nadu, West Bengal etc while it has decreased over the decade in Arunachal Pradesh, Andaman and Nicobar, Daman and Diu, Jammu and Kashmir and Madhya Pradesh.

From Fig. 4, it can be seen that the GER of girls has also increased in most of the states of India over the decade under study. For instance it has increased significantly in Chandigarh, Chhattisgarh, Jharkhand, Kerala, Meghalaya, Punjab, Tamil Nadu and West Bengal. It has only decreased in Arunachal Pradesh, Jammu and Kashmir and Lakshadweep.

From Fig. 5 and 6, it can be seen that the total GER has increased everywhere except Arunachal Pradesh, Lakshadweep, Andaman-Nicobar and Jammu-Kashmir.

FIGURE 7: GRAPHICAL REPRESENTATION OF YEARWISE GER (TOTAL, MALE AND FEMALE)



Here, the representation of time in years is as follows:

Session	X- Axis
2004-2005	1
2005-2006	2
2006-2007	3
2007-2008	4
2008-2009	5
2009-2010	6
2010-2011	7
2011-2012	8
2012-2013	9
2013-2014	10
2014-2015	11
2015-2016	12

□ **Fitting a Mathematical Curve:-**

In Fig 7 from the graphical visualization of our time series data it can be observed that the trend of my time series data i.e. is entirely increasing and there is not any sign of changing its direction in downward. So for polynomial trend line, I have successively fitted Linear model for my data. I have performed their ANOVA testing and made decision of accepting or rejecting basis on p values provided by the test. Here for the ANOVA testing is 1% level of significance is chosen. I have also observed the value of coefficient of determination i.e. the R^2 values and adjusted R^2 value in each case.

Linear Model:

The fitted linear model for total GER is:-

Y_t= 2.730t+ 22.08, with origin at 0 and 1 year unit of time=t.

ANOVA:

ANOVA					
	df	*SS*	*MS*	*F*	*p-value*
Regression	1	1065.819301	1065.819	142.8954	3.0291E-07
Residual	10	74.58736597	7.458737		
Total	11	1140.406667			

Since p-value=3.0291E-07<0.01, we accept the linear model fitting.

Regression Statistics	
Multiple R	0.96674496
R Square	0.934595817
Adjusted R Square	0.928055399
Standard Error	2.731068765
Observations	12

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	22.0878788	1.680856631	13.14085	1.2381E-07	18.34269684	25.83306074	18.34269684	25.8330607
X Variable 1	2.73006993	0.228383443	11.95389	3.0291E-07	2.221199909	3.238939951	2.221199909	3.23893995

CONCLUSION:

After analysing the time series data for GER of higher secondary school education in India over all the states for 12 sessions, 2004-2005,2005-2006……2015-2016, it can be seen from Fig 7 that the GER has consecutively increased over the years, with a slightly fall in 2012-2013 than the session before. Else for every session, the total GER has increased (combining male and female). So from the time series analysis, my data has an overall upward trend.

Also, from the linear model regression, the value of R^2 is 0.934596 which is quite close to 1 and thus it can be said that this statistical model can predict the data quite well (of course with some error).

Thus India has progressed in terms of higher secondary school education consecutively over the years but there obviously are known or unknown factors that still hinder the nation from achieving 100% GER.

**According to our fitted linear model and the regression equation, $Y_t = 2.730t + 22.08$, in order to find the GER of the session 2019-2020, we put the value of t=16 and get the estimated GER to be 65.76. But from the actual data obtained from a publication of RBI, (https://m.rbi.org.in/scripts/PublicationsView.aspx?id=20666) we see that the GER is 51.4. So there has been a considerable reduction in GER contrary to the upward trend of graph for the past sessions. Apart from other contributing factors, I think one of the main reasons is the ongoing COVID-19 pandemic situation which has led to a vast disruption in the education system among other things. The enrolment of students in higher secondary level has reduced because numerous students, especially in rural areas have dropped out of school in the lockdown phase and engaged in some kind of paid jobs mostly as a labourer to help their families amidst the crisis. This is a disastrous setback in the Indian school education system.

REFERENCES:

□ Sources of Theory:-

- Fundamental of Statistics (Volume One & Two) – A.M. Goon, M.K. Gupta, B. Dasgupta
- https://www.statisticshowto.com/adjusted-r2

□ Software Used:-

- Microsoft Excel 2013
- Microsoft Word 2013

□ Sources of Data:-

- https://m.rbi.org.in/scripts/PublicationsView.aspx?id=20666
- https://www.education.gov.in/sites/upload_files/mhrd/files/statistics-new/udise_201920.pdf
- https://www.education.gov.in/sites/upload_files/mhrd/files/statistics-new/ESG2016.pdf