# R Programming

## WDI Data with R

The World Bank is a trove of information, and it makes a lot of its data available over the Web. For more sophisticated analysis, the public can download data from the World Bank's Data Catalog or access it through an API. The most popular dataset is the World Development Indicators (WDI). WDI contains, according to the World Bank, "the most current and accurate global development data available, and includes national, regional and global estimates." WDI comes in two downloadable forms: Microsoft Excel and commaseparated values (CSV) files. (Because Microsoft Excel files aren't suitable for programmatic analysis, we deal with the CSV files here.) . The main aim of the data is plot the correlation based on following

## India - Enrolment in primary and Secondary Education

In this first of all I was extracted the India's education related data from WDI world bank data(.csv format) using Education **Indicators**.

Later I did the following steps

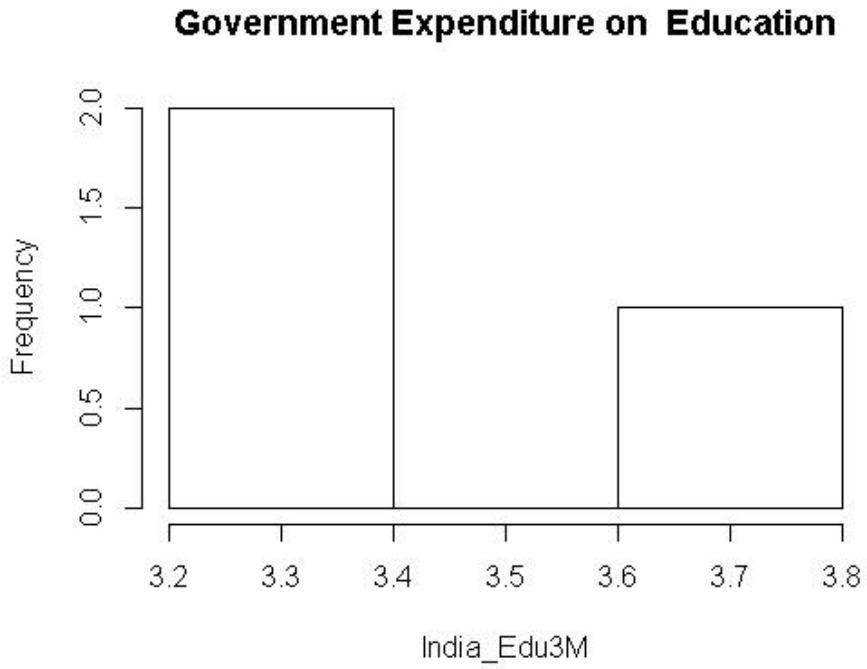
1. To know the total no of Rows and Columns data frame.

In this I found Variables is in rows 1421, Years is in columns.

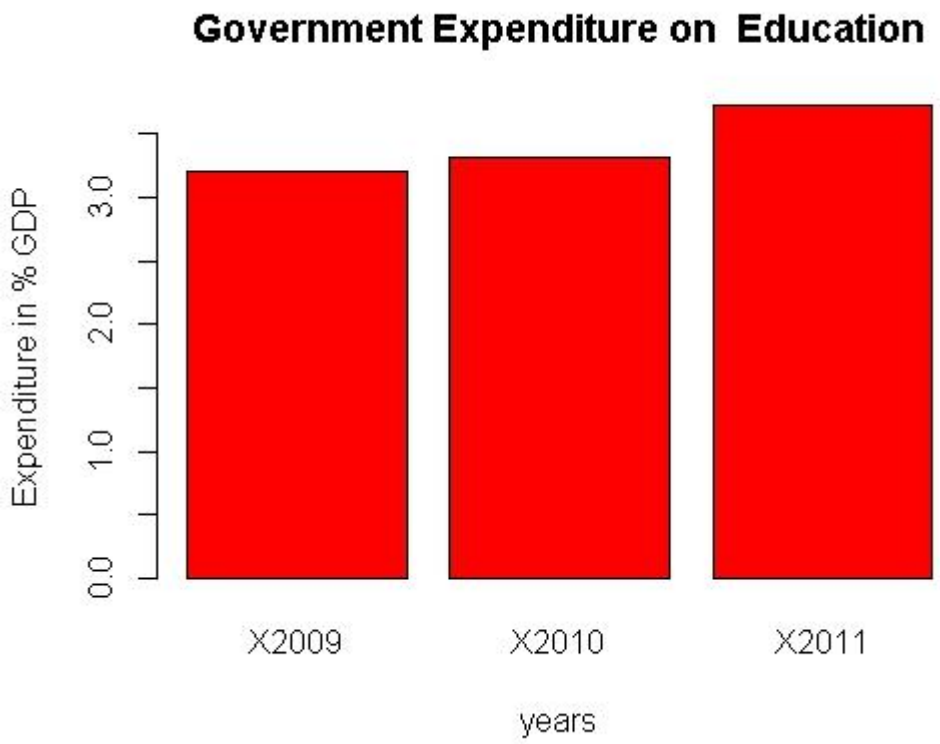
2. Selecting all Financial variables (in Row) and no of years of values (Columns)
3. Extracting India's Gross enrolment ratio, pre-primary, both sexes (%) variables year 1960 to 2014
4. Government expenditure on education as % of GDP (%)
   a. Gross enrolment ratio, both sexes (%)
   b. Extracting Data from year 2009 to 2011
   c. Here I was defined mean, standard deviation
   d. summary

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 3.210 | 3.265 | 3.320 | 3.417 | 3.520 | 3.720 |

   e. Virtualization - Plotting Graghics : Histograms




f. Government expenditure on education in the years 2009,2010,2011




year X2009 X2010 X2011

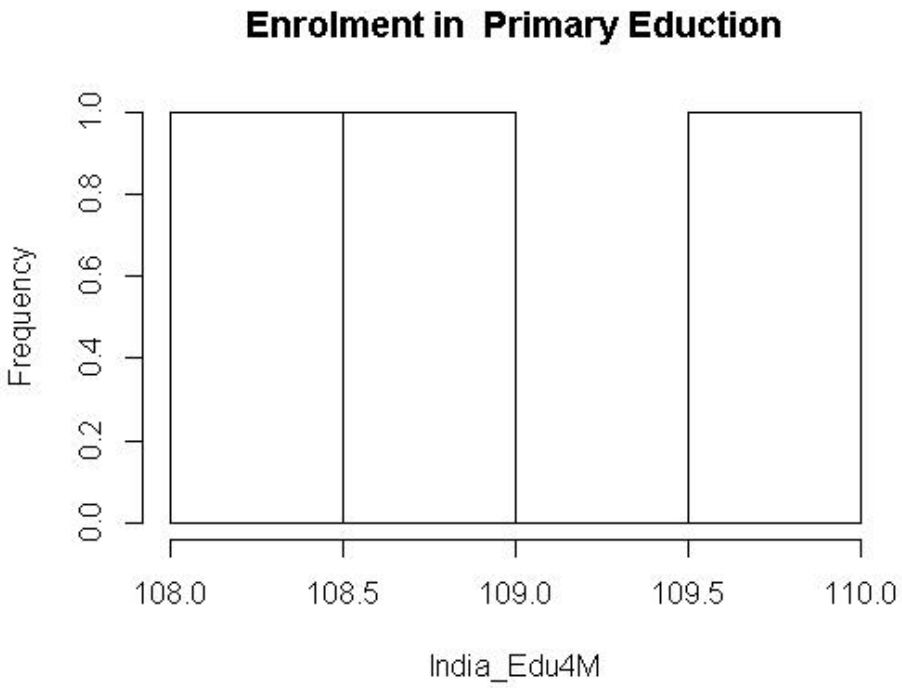
Enrol 110 109 108

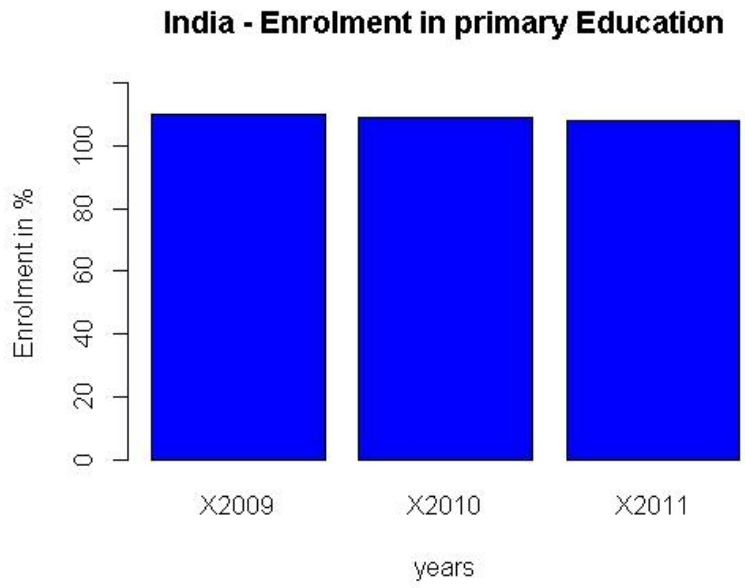
5. India - Enrolment in Primary Education
   a. Gross enrolment ratio, primary, both sexes (%)
   b. Extracting Data from year 2009 to 2011
   c. Summarizing

      Min. 1st Qu. Median Mean 3rd Qu. Max.

      108.0 108.5 109.0 109.0 109.5 110.0

   d. Virtualization - Plotting Graphics






6. Gross enrolment ratio, secondary, both sexes
   a. Extracting Data from year 2009 to 2011

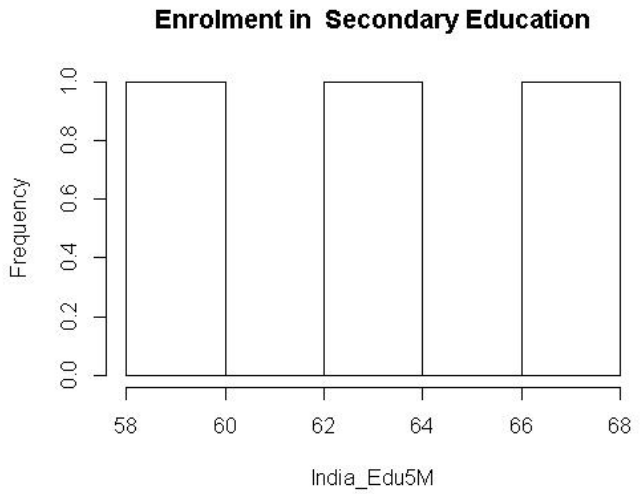
      Year X2009 X2010 X2011

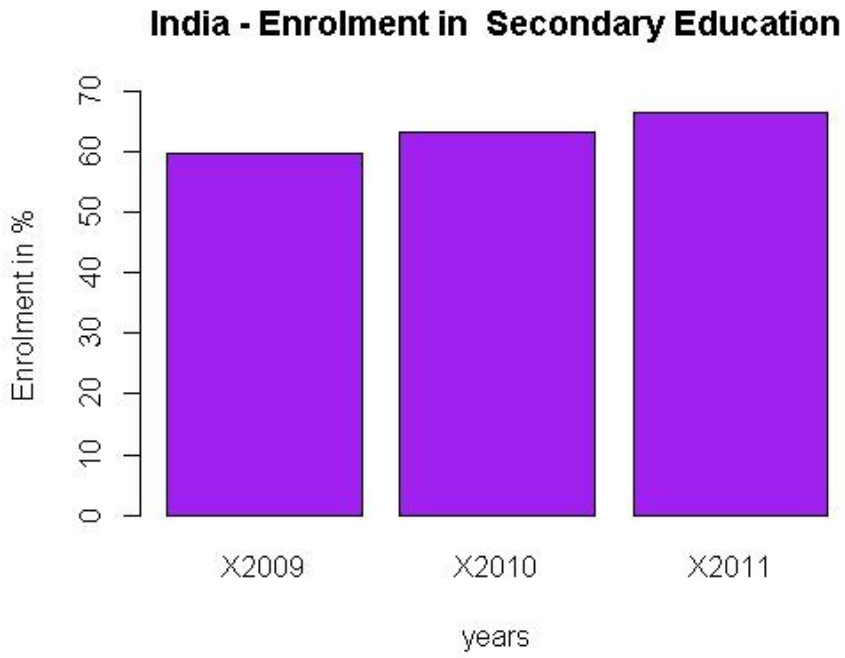
      Enrol 59.8 63.3 66.4

   b. Summarizing

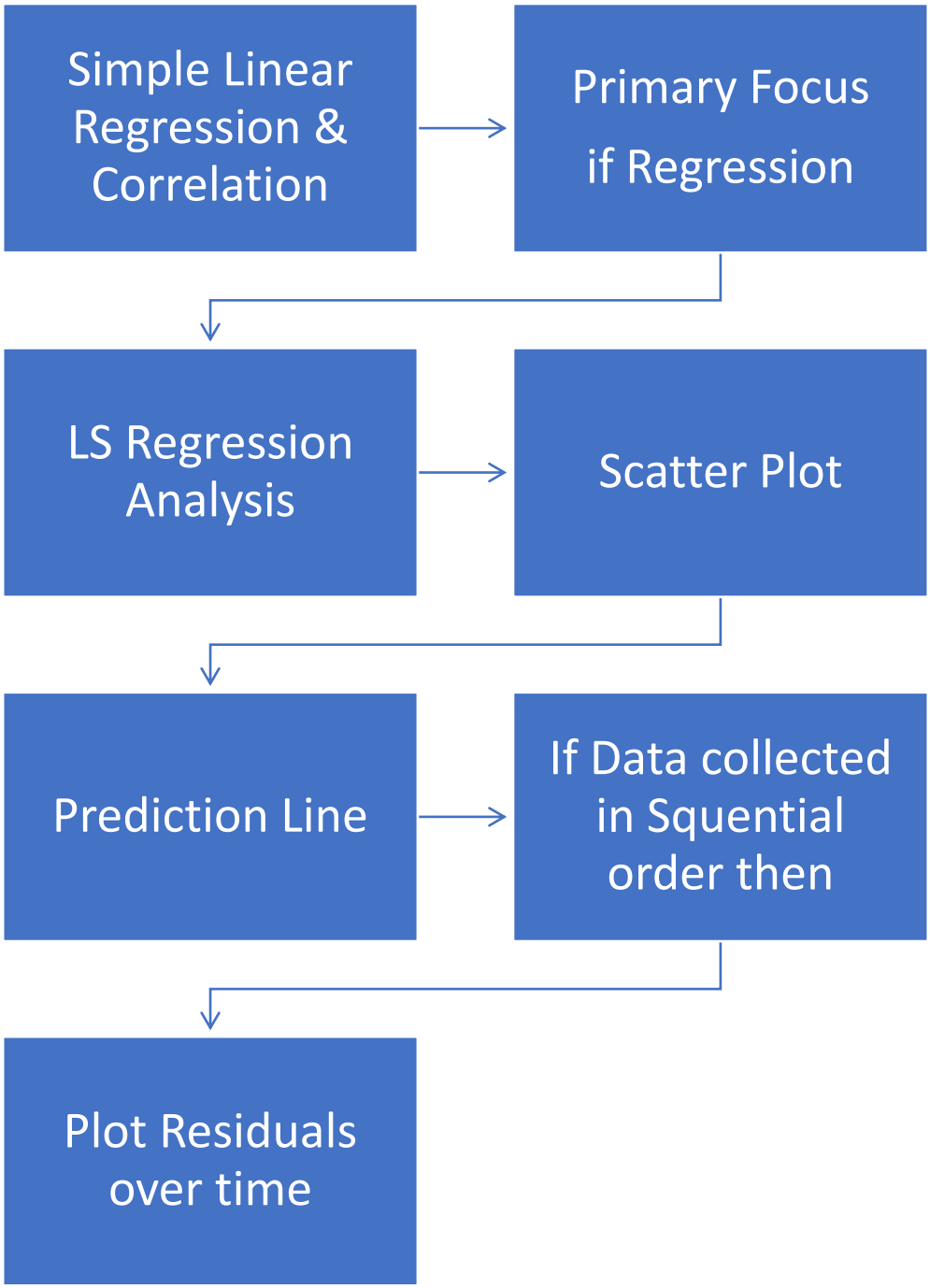
      Min. 1st Qu. Median Mean 3rd Qu. Max.
      59.80 61.55 63.30 63.17 64.85 66.40

   c. Virtualization - Plotting Graphics

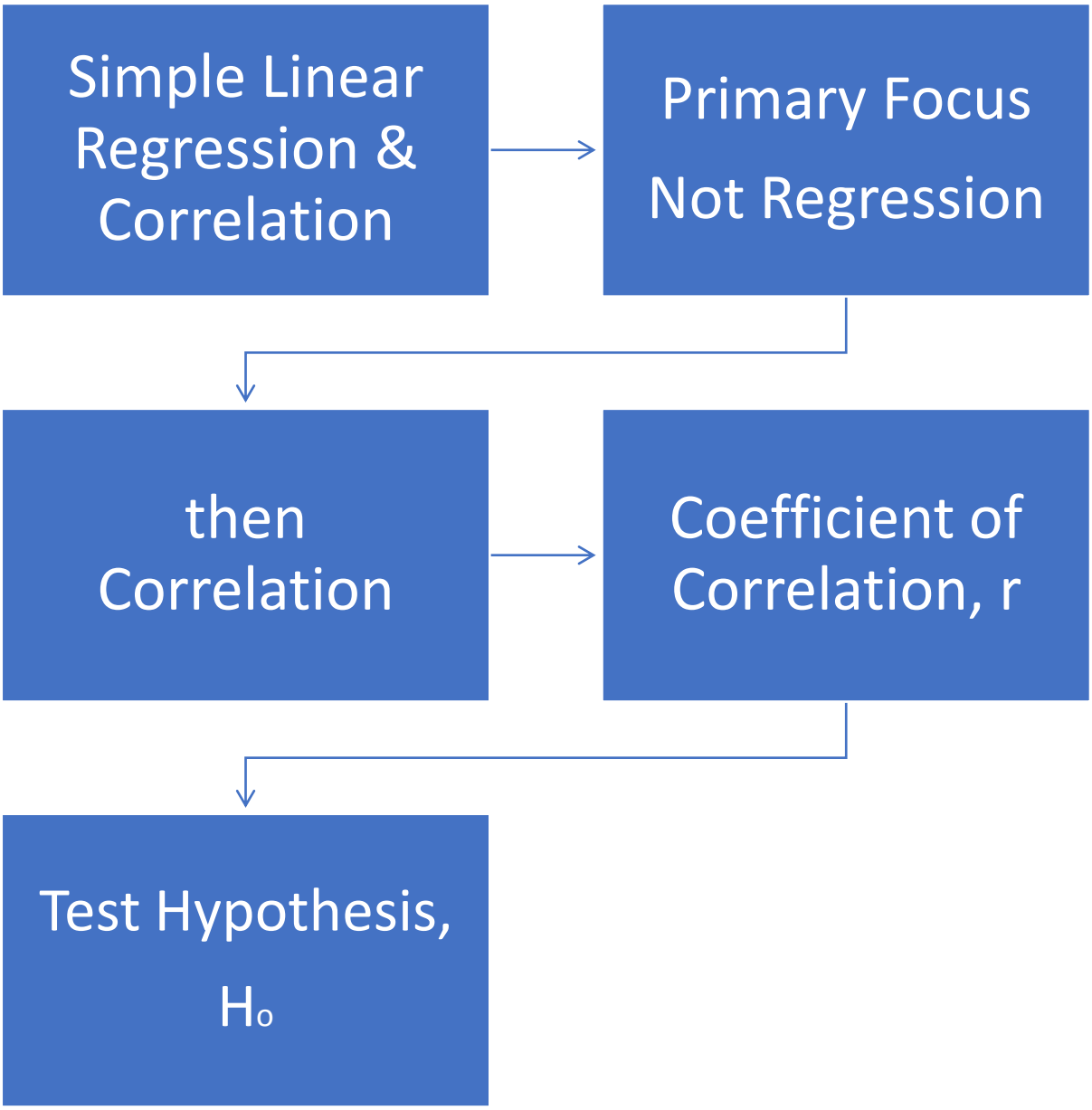

Enrolment in Secondary Education
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
58
60
62
64
66
68
India_Edu5M


India - Enrolment in Secondary Education
Enrolment in %
0
10
20
30
40
50
60
70
X2009
X2010
X2011
years

**Flow Method of Simple Linear Regression Model**



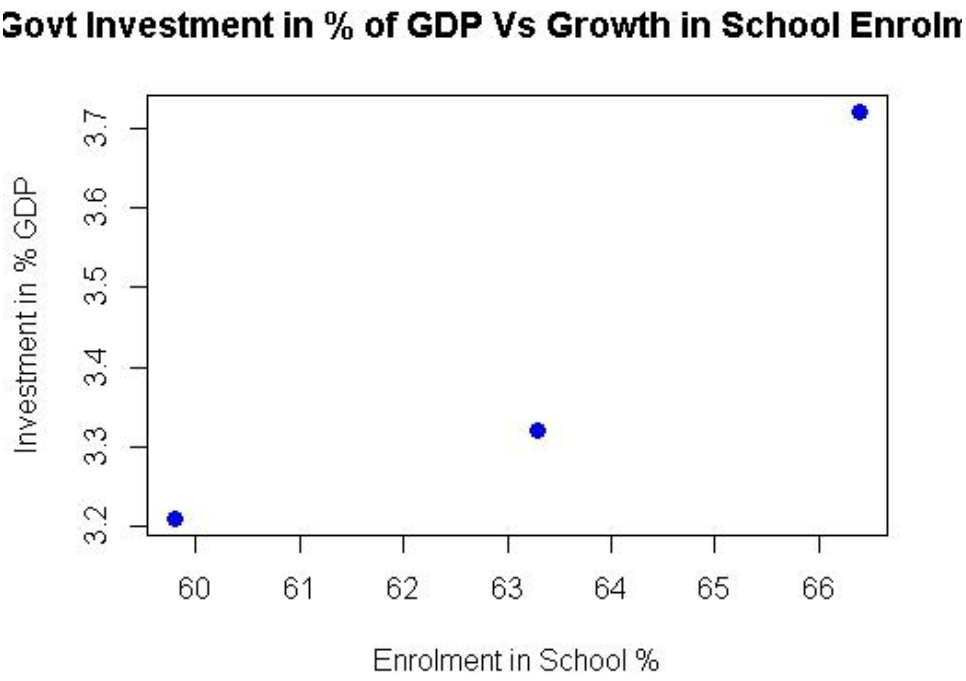
**Flow of Correlation Map**



7. Simple Regression and correlation Analysis
   a. Merging data frame –Secondary enrolment & Govt Spending on Education
   b. Covariance's and correlations
   c. Displays detailed results for the fitted model

```
summary(relation)

Call:

lm(formula = x ~ y)


Residuals:

    1       2       3

-0.9801  1.2496 -0.2695


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   23.711     14.530   1.632    0.350

y             11.548      4.244   2.721    0.224


Residual standard error: 1.611 on 1 degrees of freedom

Multiple R-squared:  0.881,           Adjusted R-squared:  0.762

F-statistic: 7.404 on 1 and 1 DF,  p-value: 0.2242


#   Lists the predicted values in a fitted model

fitted(relation)

      1        2        3

60.78010 62.05037 66.66953


#   Lists the residual values in a fitted model

residuals(relation)

        1         2         3

-0.9801018  1.2496298 -0.2695280
```

d. plot the Chart




Conclusion:

The type of the regression model is considered for the analysis is scatter plot method. I consider the relevant range of the independent variable of X in making the predication. The relevant range is smallest to largest values of X but I can interpolate within the range of X values. The scatter plot shows that the enrolment was increased by government expenditure in the education- government spending on Education directly proposal to school enrolment.

From the Pr(>|t|) column, I see that the regression coefficient (23.71) is significantly different from zero ($p < 0.001$) and indicates that there's an expected increase 23.71enrolment in school for every 1 % of GDP spend on the Education. The multiple R-squared (0.881) indicates that the model accounts for 88.1 per-cent of the variance in enrolment. The multiple R-squared is also the correlation between the actual and predicted value (that is, R ^2 =rr ). The residual standard error (1.611.) can be thought of as the average error in predicting enrolment from government spending using this model.

The F statistic tests whether the predictor variables taken together, predict the response variable above chance levels. Because there's only one predictor variable in simple regression, in this example the F test is equivalent to the t-test for the regression coefficient for government spending. For demonstration purposes, I've printed out the actual, predicted, and residual values. Evidently, the largest residuals occur for low and

high spending, which can also be seen in the plot. The plot suggests that we might be able to improve on the prediction by using a line with one bend. For example, a model of the form Y  = + + β β β 0 1 2 2 X X may provide a better fit to the data. Further analysis on the Polynomial regression allows us to predict a response variable from an explanatory variable, where the form of the relationship is an nth degree polynomial.

Therefore, according to the secondary school enrolment data set, I could conclude that every money spending on education increase the Indian student enrolment in school. But whereas, the primary data set shows that decrease enrolment in the school. However, I am could not conclude that spending on the primary schooling for children, because the we got most recent data values in the three year period only –year 2009, 2010, and 2011. I presume that this number of the data may not give correct conclusion of the primary school enrolment by children. For that we may require further analysis on quantitative and qualitative to get conclusion.

**Program No – 2** Indian's Agricultural Growth Data Analysing

Similarly, India's agricultural data set also analysed using similar R-programming's above coding that has presented above for Educational Enrolment in India analysis.

I am not attached the code details here, but we enclosed the two charts for review.

Analysing the below two charts, I observed that in the chart -1 the agricultural land area in sq-m and chart-2 agricultural status data in chart-2 do not match with chart – 1. In chart -2 shows that land area is constant over the period of the years – no increment in the agri-land area whereas the chart -1 shows that various every years. Therefore data has been collected may not correct and need review with concern source for correcting data values.

**Chart1**

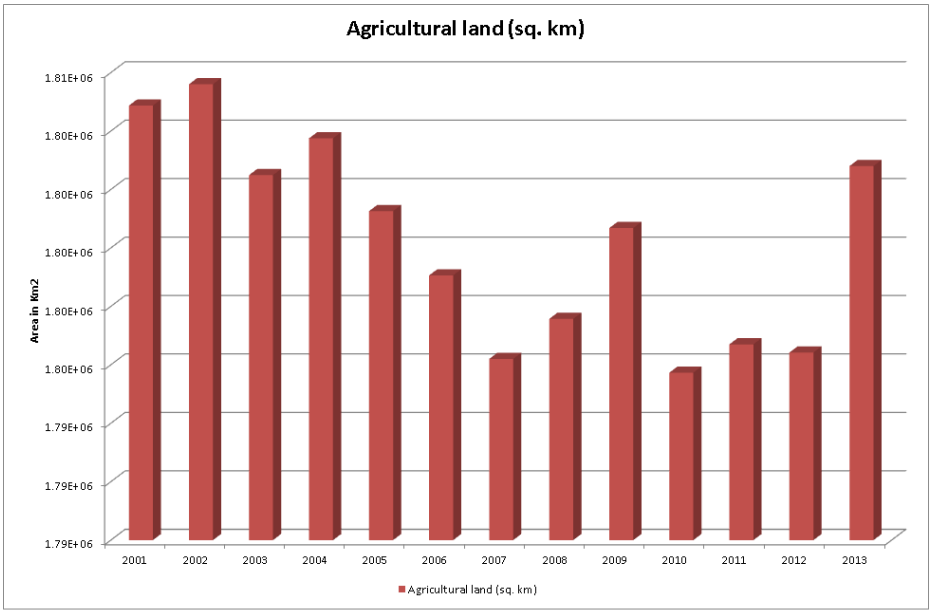
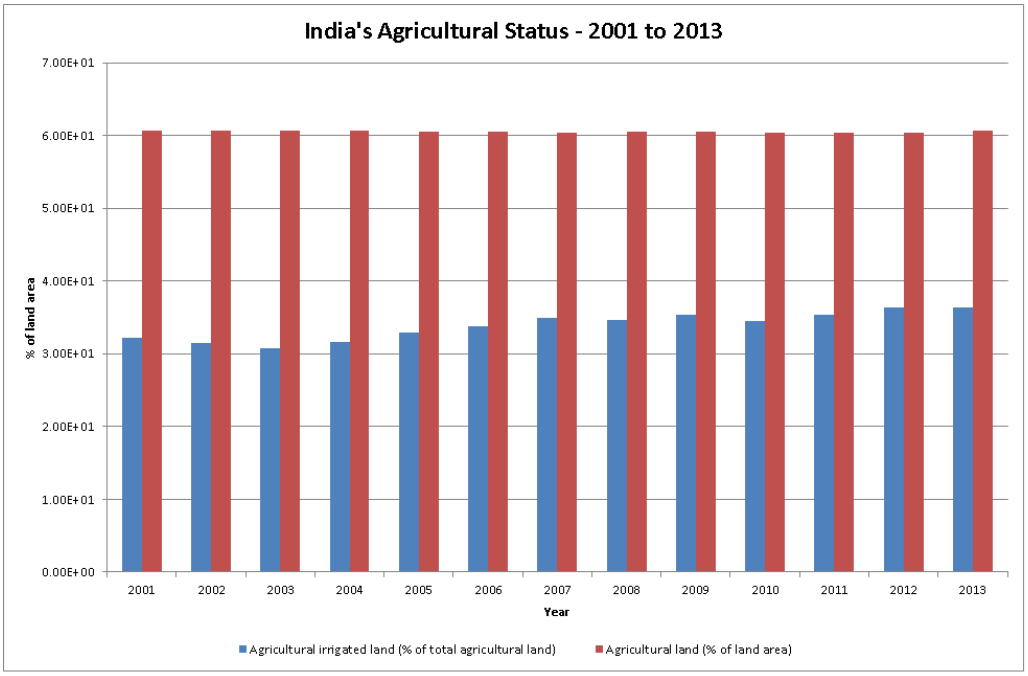



**Chart2**




## The Python Statistics Ecosystem

## WDI Data with Python

The main aim of the data is plot the correlation based on following

## Spearman's Pair wise or Rank-Order Correlation

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, ( $\rho$, also signified by $r_s$) measures the strength of association between two ranked variables.

The Spearman correlation coefficient, $r_s$, can take values from +1 to -1. A $r_s$ of +1 indicates a perfect association of ranks, a $r_s$ of zero indicates no association between ranks and a $r_s$ of -1 indicates a perfect negative association of ranks. The closer $r_s$ is to zero, the weaker the association between the ranks.

**Assumptions of the Test**

We need two variables that are either ordinal, interval or ratio. Although we would normally hope to use a Pearson product-moment correlation on interval or ratio data, the Spearman correlation can be used when the assumptions of the Pearson correlation are markedly violated. A second assumption is that there is a monotonic relationship between your variables. In our data assumption is Indicator Code.

A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases.

**Definition of Spearman's rank-order correlation**:

There are two methods to calculate Spearman's rank-order correlation depending on whether: (1) your data does not have tied ranks or (2) your data has tied ranks. The formula for when there are no tied ranks is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i$ = difference in paired ranks and $n$ = number of cases. The formula to use when there are tied ranks is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where $i$ = paired score.

**Programming Design**:

A program for calculating World Development Indicators correlations using Python. The program collects the top 30 most measured indicators, calculates the Spearman pairwise correlations, and shows the results graphically.

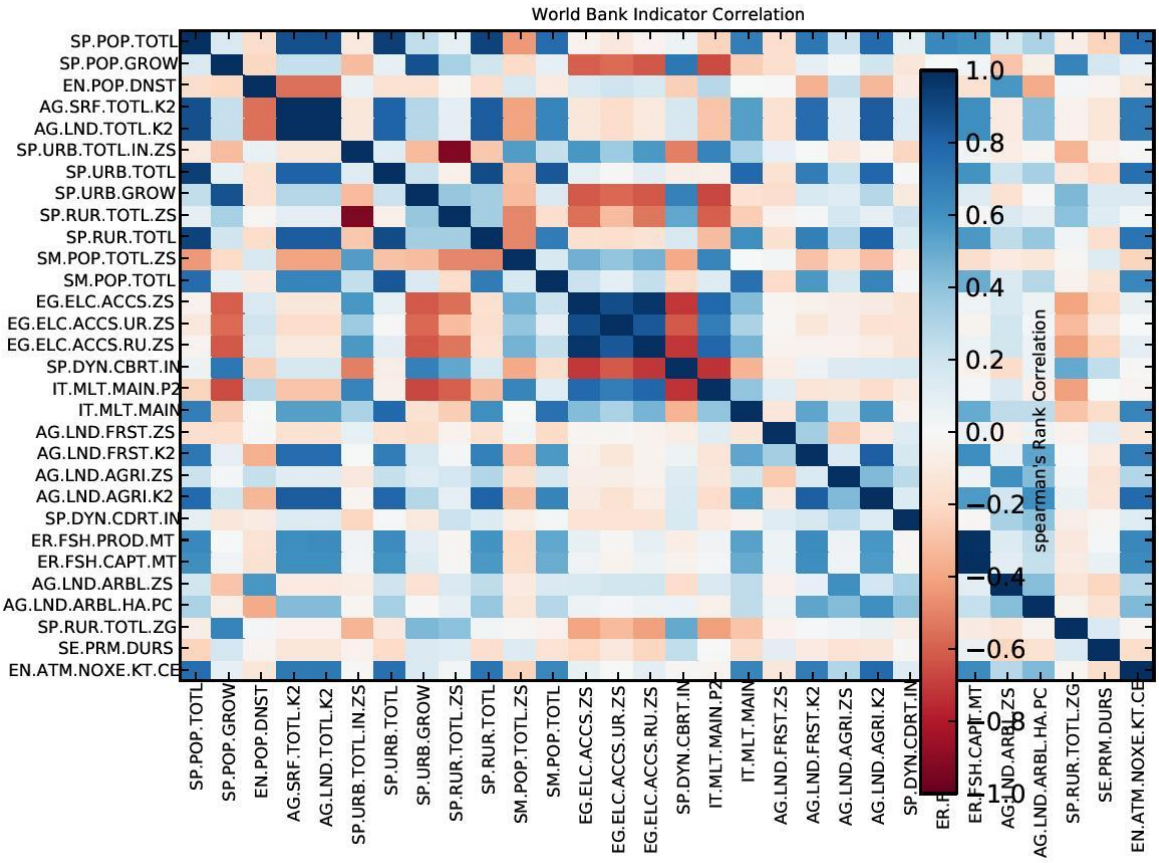



Figure 1: World Development Indicators correlations matrix with Python created from the program.

**Conclusion:**

In this we are not extracting data from web because it is very easy task. Already so many algorithms are designed by researchers for extracting WDI data. So, We are directly gives the input(downloaded from web in .csv format) to Python Programming

for reading .csv files. In this task generating correlation is not easy because the data has so many columns, rows and it is semi structured data. In this semi structured data we will make the spear's man correlation from all countries (Indicates in colors format) using with Indicator correlation. The diagonal has perfect correlation—as it should, because we're examining the same indicators. In addition to that, we do see that there are indicators that correlate with each other—some positively, even strongly so, and some negatively or very negatively.

**Thanks ☺**