1. **Methodology**

This section is divided into two subsections: Sects. 3.1 and 3.2. Section 3.1 describes the ICDA dataset, Section 3.2 for Proposed ICDA approach.

**3.1** **ICDA dataset**

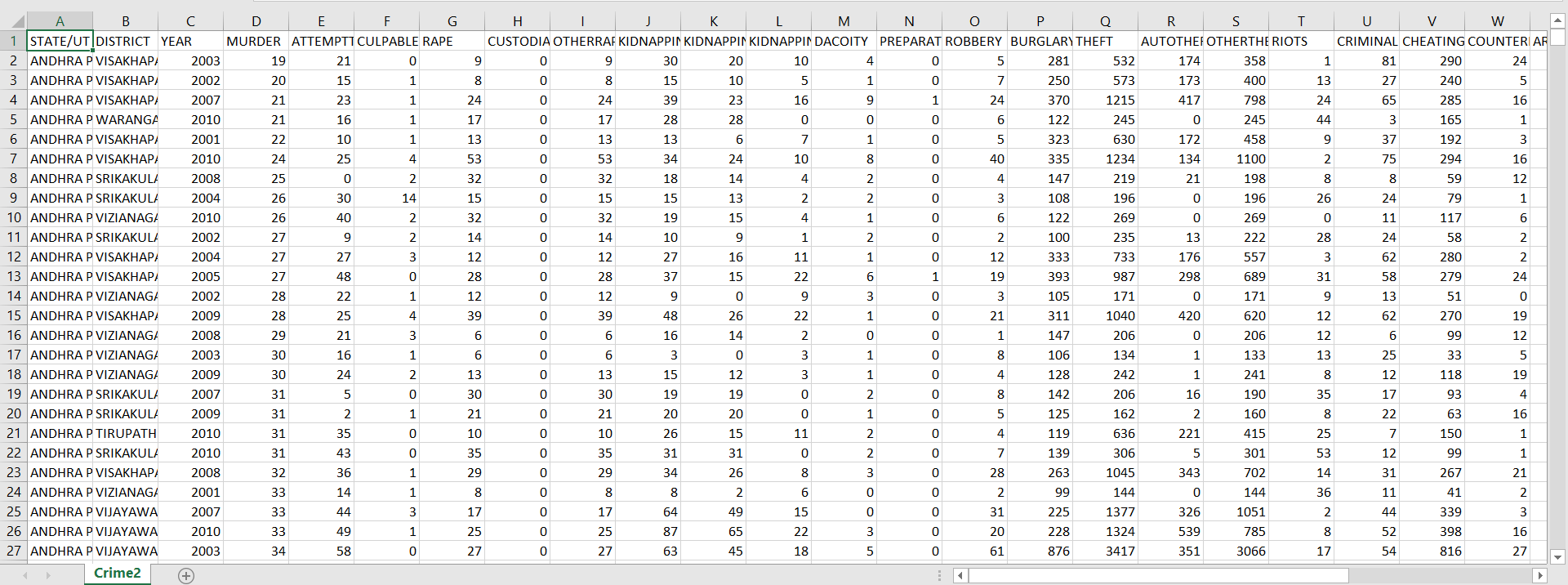
The raw data set is taken from data.gov.in. The data set contains 9017 records and 33 columns (variables), each record is distinct on the basis of STATE/UT name, DISTRICT name and Year. Other 30 variables are: 28 Variables representing specific crime count in the corresponding area and year:

MURDER,ATTEMPT TO MURDER,CULPABLE HOMICIDE NOT AMOUNTING TO MURDER,RAPE,CUSTODIAL RAPE,OTHER RAPE,KIDNAPPING & ABDUCTION,KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS,KIDNAPPING AND ABDUCTION OF OTHERS,DACOITY,PREPARATION AND ASSEMBLY FOR DACOITY,ROBBERY,BURGLARY,THEFT,AUTO THEFT,OTHER THEFT,RIOTS,CRIMINAL BREACH OF TRUST,CHEATING,COUNTERFIETING,ARSON,HURT/GREVIOUS HURT,DOWRY DEATHS,ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY,INSULT TO MODESTY OF WOMEN,CRUELTY BY HUSBAND OR HIS RELATIVES,IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES,CAUSING DEATH BY NEGLIGENCE

OTHER IPC CRIMES: Count of crimes other than the above mentioned 28 crimes.

TOTAL IPC CRIMES: Count of total crimes which are crime under IPC.

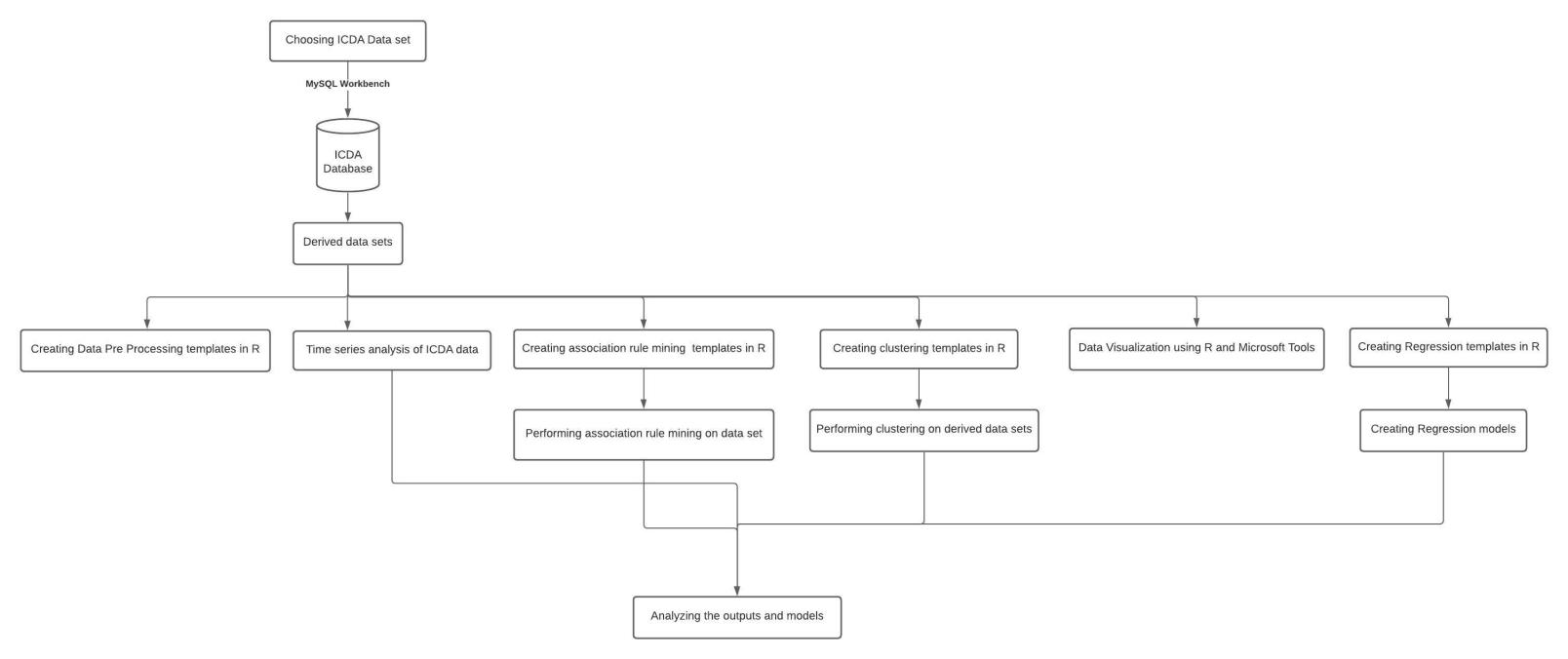
Fig.1 shows the ICDA sample dataset.



**Fig.1** ICDA sample dataset

**3.2 Proposed** **ICDA approach**

This section tells us the work flow of our ICDA approach of data analysis using R and other tools. Figure fig 3.2 is a flow chart which explains the work flow of the ICDA approach. We chose appropriate data set from web sources and then initialized the creation of ICDA database by importing the chosen ICDA data set in ICDA data base created using MySQL workbench. Then we used SQL queries to get the desired derived data sets as per requirement. We performed time series analysis and forecasting on ICDA data set. for We used R for data visualization of derived data sets. In order to save time we reduced the redundant R coding we created R templates for data pre processing, association rule mining(apriori and eclat), clustering (K means and hierarchical) and regression (SLR, MLR, PLR, SVR, Decision Tree Regression, Random Forest Regression). Then we used these templates to create different regression models and perform clustering and association rule mining on derived data sets. We created different number of models using different data sets and changing dependent variables. Similarly, we performed clustering on different data sets.



**Fig 3.2** Work flow for ICDA approach

1. **Experiments and results**

As per the chosen raw data set we have performed data pre-processing to get clean, useful and necessary derived and transformed data sets for applying machine learning algorithm and perform data visualization using the data sets. There are many machine learning algorithms and data visualization techniques, but we used only those algorithms and techniques which will be able to produce efficient models which in future can be used to predict possible specific and total ipc crime count provided desired inputs and for this we have used different available regression models; we have used k means and hierarchical clustering algorithms which can produce outputs to study similarity among areas in same and different year on the basis of specific and total ipc crime counts; we have performed time series analysis and forecasting on the ICDA data set; we have used transformed transaction data set to perform apriori and eclat association rule mining algorithms to produce sets of crimes which were likely to happen together frequently in India in 2001-2012. Our work can be used as a template and can be used for similar data sets and on a larger data set to produce models and outputs which are based on recent data.

This section discusses the above mentioned ICDA experimentation and its results in detail. This section is divided into five subsections: Sects. 4.1-4.5 as mentioned below.

* 1. **ICDA data pre-processing**

Initially we have raw data set which requires data preprocessing before using the data for visualization or machine learning. Data pre-processing is an important step in the process of machine learning as the quality of a machine learning model is completely dependent upon the data provided to the algorithm. Throughout our work we carried out following data preprocessing:

## Taking care of missing values using R:

We replaced the **NA/na** values with the mean of all the column values of the corresponding column.

1. **Dummy coding the categorical variables:**

Since, regression algorithm needs numerical data to work with and label encoding will be a wrong approach. Therefore, we used dummy.data.frame() method from the dummies library to create dummy variables for each state and district, which leads to creation of 35 state dummy columns and 808 district dummy columns.

**R Code (dummyVariable.R):**



1. **Data Transformation before implementing association rule mining on the chosen data set:**

Before applying any association rule mining algorithm to the data set, we have to first convert the data set into a transaction data set.

For this:

1. We calculated the median of each numerical data column (represents the median of the count of specific crimes and total IPC crime in the time period of 2001-2012) i.e. median for all the 30 variables.
2. Then we used SQL queries to replace the cell value with less than median value with 0 and greater one with 1:

dataToSparseMatrix.sql

1. Then we changed the datatype of each column to text from int using MySQL Workbench data import wizard and then replaced the 1s with the corresponding crime name and zeroes with no value:

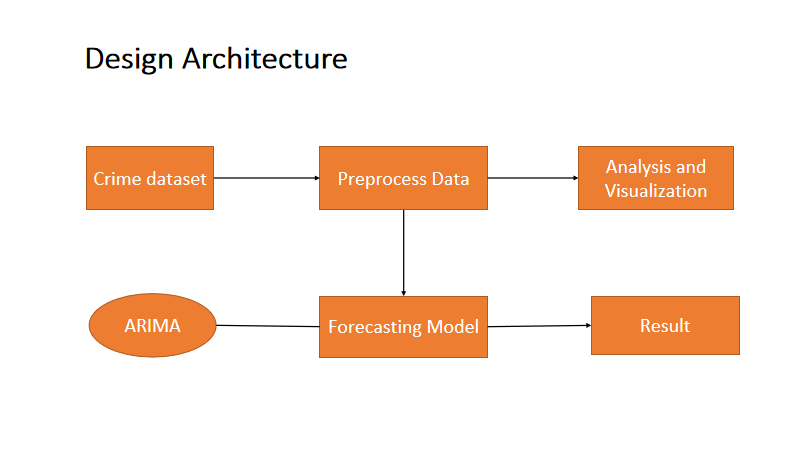
SparseDataToTransaction.sql

* 1. **ICDA time series analysis and forecasting implementation**

‘Time Series Analysis and Forecasting’ states that any information periodically recorded with time can be used for forecasting a future event related to the information. In India where criminal activities take place more frequently. By applying modern technology forecasting techniques to these cities crime data, future crime rates can be forecasted. This project analyses crime data and gives various visualizations for easy understanding of the results. It also uses past 11 years’ crime data from Kaggle to forecast future crime rate.

This crime analysis helps the government, police and residents of the cities in various ways. This information could help communities in different ways, say, alerting the neighbourhood watch or patrol departments during the time of high probability for a crime or suggesting students or business travellers to plan their stay a bit safer.

For Time Series Forecasting in this project forecasting methods like ‘ARIMA’ (Auto Regressive Integrated Moving Average) has been used. Fig. 3 shows design architecture of ICDA time series analysis and forecasting with the help of a flow chart.



**Fig. 3** Design architecture of ICDA time series analysis and forecasting

Data sets considered for this project are crime data information of India. Kaggle gives information about various crimes in different States at India. In this project India with different Year and Total IPC Crime are selected.

{Total IPC Crime = Murder + Attempt to murder + Dowry +…}

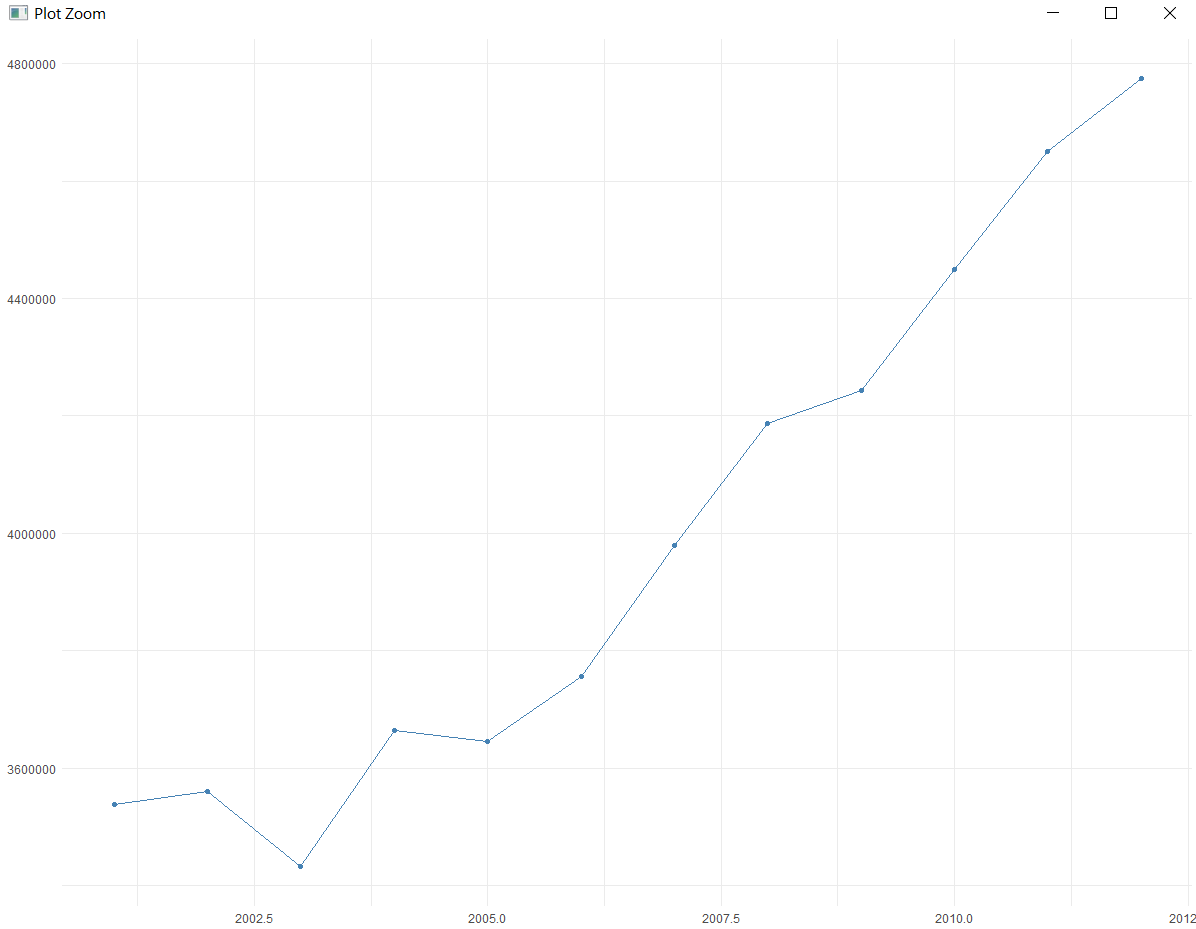
India crime data from 2001 to 2012 is taken from Kaggle.

Data Pre-processing is the important stage in any analytics/machine learning project. After extracting the required data, it is a crucial step to get the important attributes from the data set. This project analyses and takes data from 2001 to forecast future crime. So, first step of pre-processing is to extract data from 2001 to 2012.

**Data Exploration:**

**How has Crime Evolved over time in the India?**

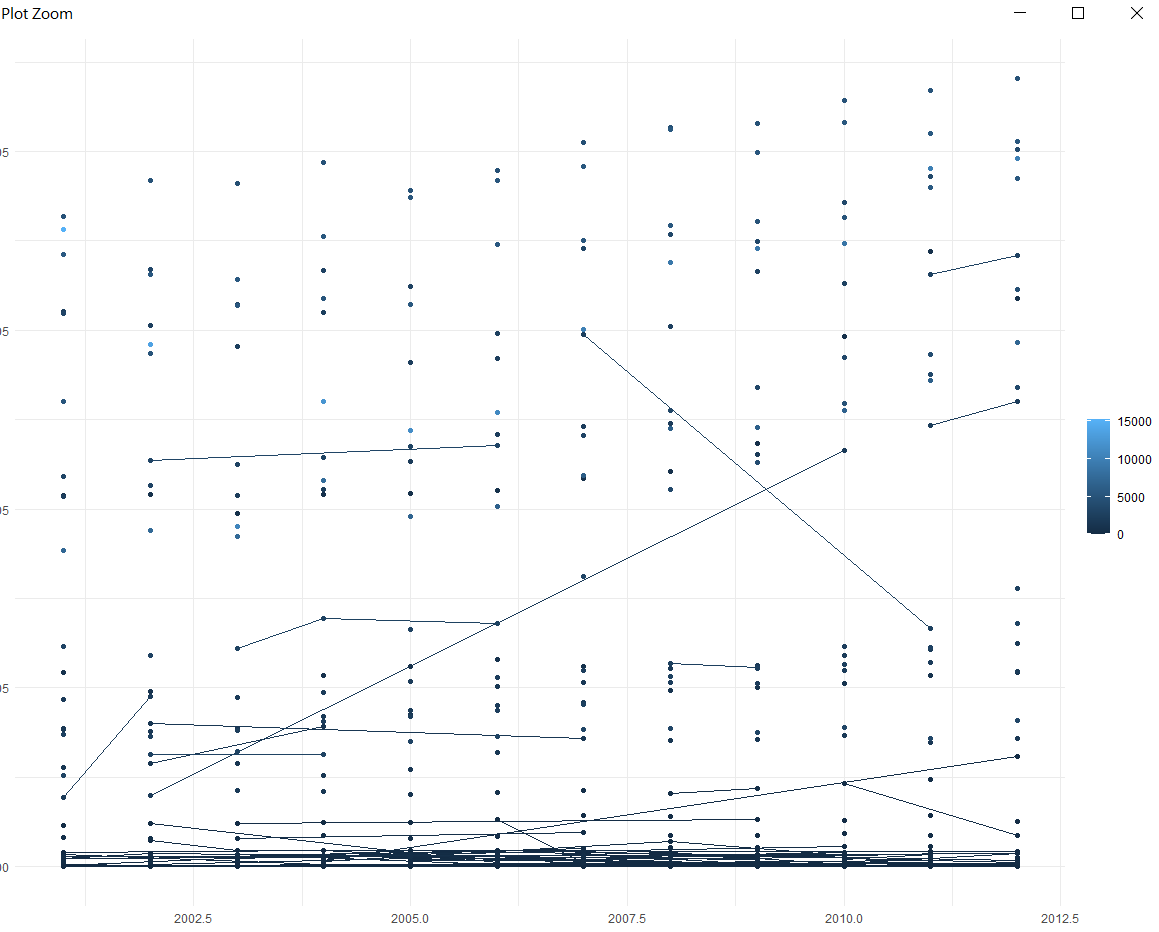
To answer the question we plot the number of crime per year from 2001 to 2012. The graph (Fig. 4) shows that Crime in the India has been increased year after year with continuous Incline.



**Fig. 4** Number of total IPC crimes versus year

To understand the complexity in numbers, trend in the data and to observe the entire Murder data through a single visualization. Murder crime count in Year over the period of 11 years for the India is represented as map in given fig. 5.

Darker colour shades of blue represents lower crime count due to Murder in each Year and lighter colour shades represents higher crime count due to Murder. X axis contains Year 2001-2012 representing Year from 2001 to 2012. These maps gives clear understanding to Police about the crime history of a India in a single glance.



**Fig. 5** Number of murder cases to year map plot

Likewise Murder we can plot different crime rate due to different crime in year like Dowry, kidnapping, Rape, etc.

**Conclusion:**

1. Time Series Analysis and Forecasting is performed with several visualizations and statistical models in this project.
2. According to forecasting results after for the year 2002 crimes are slightly decreasing for Some Year then It Increasing as year Increases.
3. This forecasting results can help police to take necessary precautions according to the crime rate.
   1. **ICDA clustering algorithm implementations**

**Note:** While applying clustering on the data with 29 variables, we used PCA (Principal Component Analysis) feature extraction. So before moving forward let us briefly discuss about PCA:

Principal Component analysis is a feature extraction method which is uses the mathematical concept of eigen values and eigen vectors to perform linear transformation on a data having n dimensions to produce p number of principal components which explains the maximum of the variability of the original data. The mapping between principal components and original data is done such that the value of sum of least square of differences between the recovered data by using the mapping and the original data is minimum.

PCA is completely independent of target variable rather than linear discriminant analysis and that’s why we found it of our use as we need to visualize the clusters which will be fruitful only if we have two variables explaining 100 % variability of the data rather than applying clustering with 29 variables and getting the output of clusplot() function as a plot between two components not explaining 100 % variability.

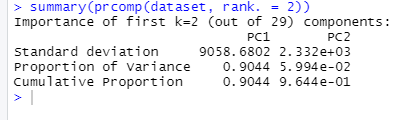
This section is divide into further subsections and subsections of subsections as per different clustering algorithm used and different data set used.

**4.3.1 Applying clustering algorithm to:**

**Data set:** 01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv

**4.3.1.1 Applying k means clustering:**

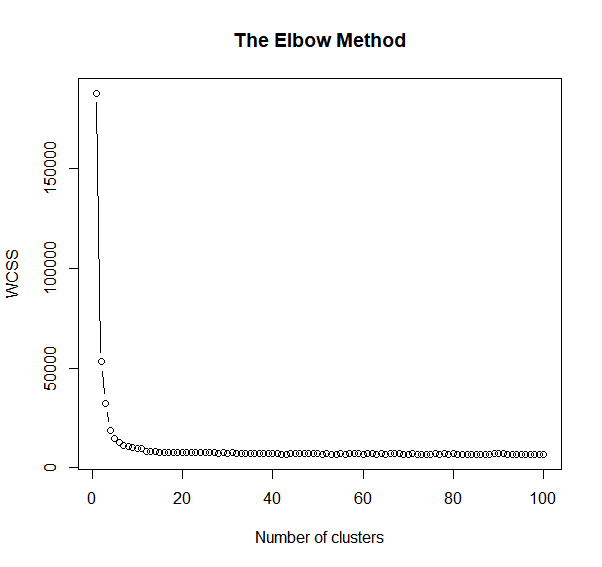
**Output:**

****

**Conclusion:**

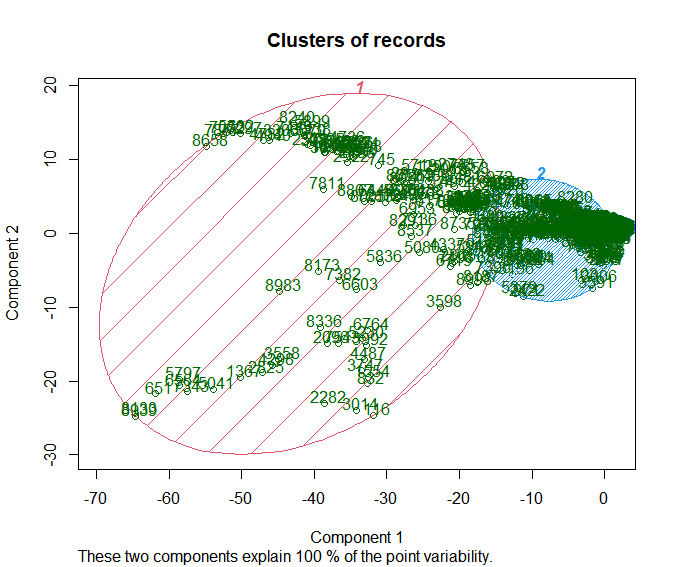
So, we have principal component, PC1 explaining 90.44 % of variability in dataset and PC2 explaining 5.994 % of variability of the dataset.

**Output:**



**Conclusion:**

From the above plot, we have our K = 2 => 2 clusters.

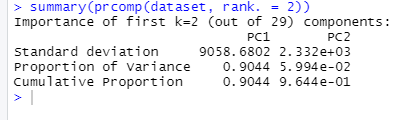
**Output:**

**Conclusion:**

1. We can observe applying pca was successful as the two components explains 100%point variability.
2. We can observe that on the basis of similarity in the number of cases registered against different crimes corresponding to 9012 district, year pairs can be segmented into two groups.
3. We can observe that cluster 2 has the majority if records belonging to it region: PC1 ranging from -20 to 0 and PC2 ranging from -10 to 10.
4. We can observe that cluster 1 has minority of records belonging to region PC1 ranging from -70 to around -20 and PC2 ranging from -30 to around +20.
5. The relation between principal components and original variables can be analyzed using ggbiplot.
6. If we consider region,year pair as an instance of a region in past time then having the data of regions or instances of regions sharing some similarity can add a lot up to crime management .

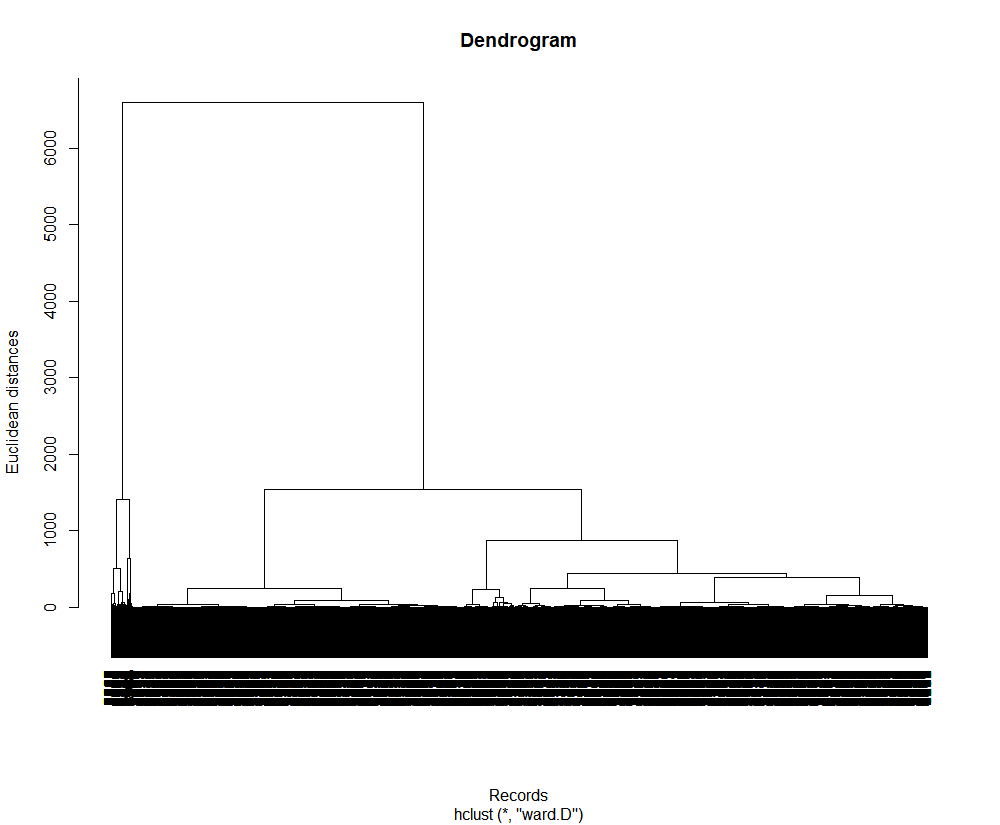
**4.3.1.2 Applying Hierarchical Clustering on the data set:**

**Output:**

****

**Conclusion:**

So, we have principal component, PC1 explaining 90.44 % of variability in dataset and PC2 explaining 5.994 % of variability of the dataset.

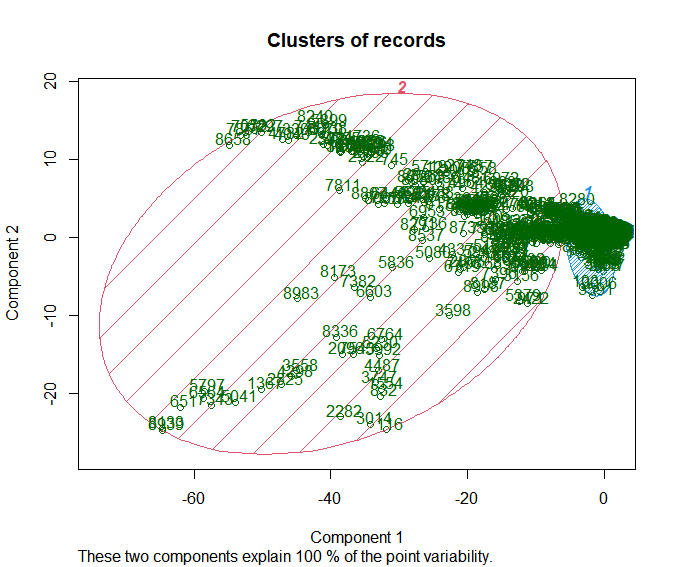
**Output:**

**Conclusion:**

From the above plot we can observe that the highest vertical line which is not cut by any hypothetical horizontal level line is intersecting at two points if cut with hypothetical horizontal line.

Therefore, 2 clusters/2 centres will be the optimal number of clusters/centres.

**Output:**



**Conclusion:**

1. We can observe applying pca was successful as the two components explains 100%point variability.
2. We can observe that on the basis of similarity in the number of cases registered against different crimes corresponding to 9012 district,year pairs can be segmented into two groups.
3. We can observe there is difference in clusters obtained by hierarchical clustering and k means clustering.
4. We can observe that cluster 1 has the majority if records belonging to it region: PC1 ranging from -7 to 0 (approx.) and PC2 ranging from -7 to +7 (approx.).
5. We can observe that cluster 2 has minority of records belonging to region PC1 ranging from -70 to around -20 and PC2 ranging from -13 to +17 (approx.).
6. The relation between principal components and original variables can be analyzed using ggbiplot.
7. If we consider region,year pair as an instance of a region in past time then having the data of regions or instances of regions sharing some similarity can add a lot up to crime management .

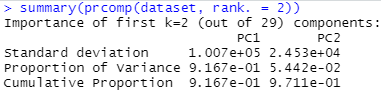
**4.3.2 Applying clustering to:**

**Data set:** DataGroupedBySTATE\_DISTRICT\_Names.csv

**Note:** The data set is obtained by sql queries to 01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv data set.

**4.3.2.1 Applying K means clustering:**

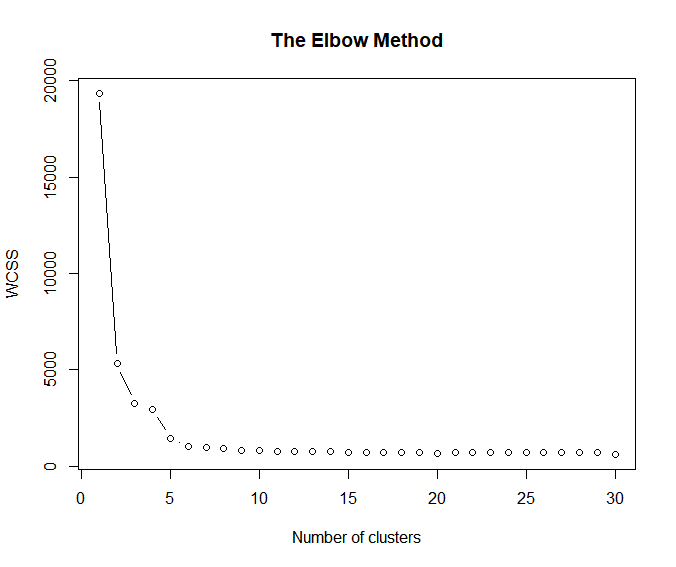
**Output:**

****

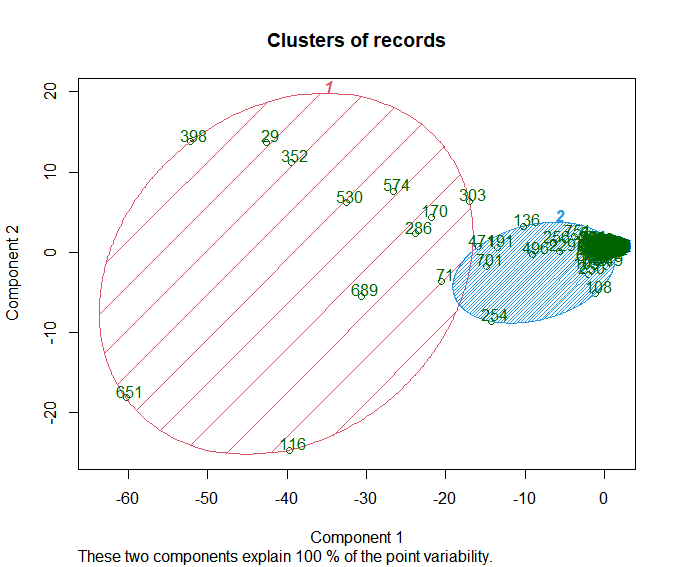
**Conclusion:**

PC1 explains 91.67 % of variability of data.

PC2 explains 5.442 % of variability of data.



From the elbow method plot we can see that optimal number of clusters will be 2.

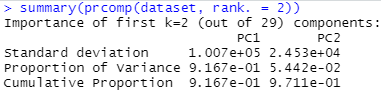


**Conclusion:**

1. As the data set we have chosen contains 862 records each corresponding to a pair of state and district and the pairs are unique, so what we obtained by applying clustering on this data set is that the records are segmented into two groups on the basis of similarity in the total number of different crimes committed in the particular state, district pair in the time period of 2001 to 2012.
2. As we can see from the clusplot plot that we have successfully obtained two separable clusters with cluster 2 having majority of the records.
3. Cluster 1 has 13 records and cluster 2 has 849 records.
4. Having the data of state, district pair sharing some similarity can add a lot to crime management.
5. The relation between principal components and original variables can be analyzed using ggbiplot.

**4.3.2.2 Applying hierarchical clustering:**

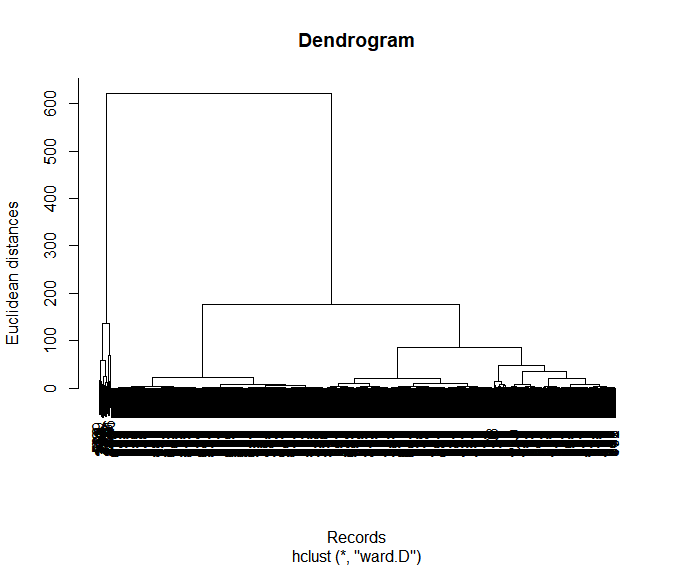
**Output:**

****

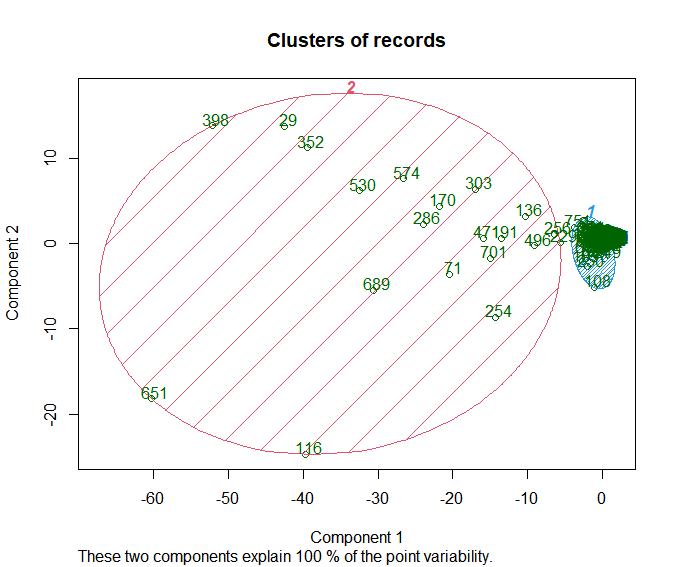
**Conclusion:**

PC1 explains 91.67 % of variability of data.

PC2 explains 5.442 % of variability of data.



From the dendrogram, we can observe that the optimal number of clusters/centres will be 2.



**Conclusion:**

1. As the data set we have chosen contains 862 records each corresponding to a pair of state and district and the pairs are unique, so what we obtained by applying clustering on this data set is that the records are segmented into two groups on the basis of similarity in the total number of different crimes committed in the particular state, district pair in the time period of 2001 to 2012.
2. As we can see from the clusplot plot that we have successfully obtained two separable clusters with cluster 2 having majority of the records.
3. Having the data of state, district pair sharing some similarity can add a lot to crime management.
4. The relation between principal components and original variables can be analyzed using ggbiplot.
5. We can observe hierarchical clustering obtained clusters differs from that of k means clustering, with more points closer to the center of there clusters in cluster 1.

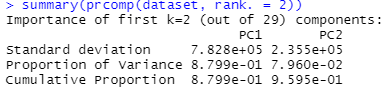
**4.3.3 Applying clustering to:**

**Data set:** DataGroupedBySTATENames.csv

**Note:** The data set is obtained by sql queries to 01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv data set.

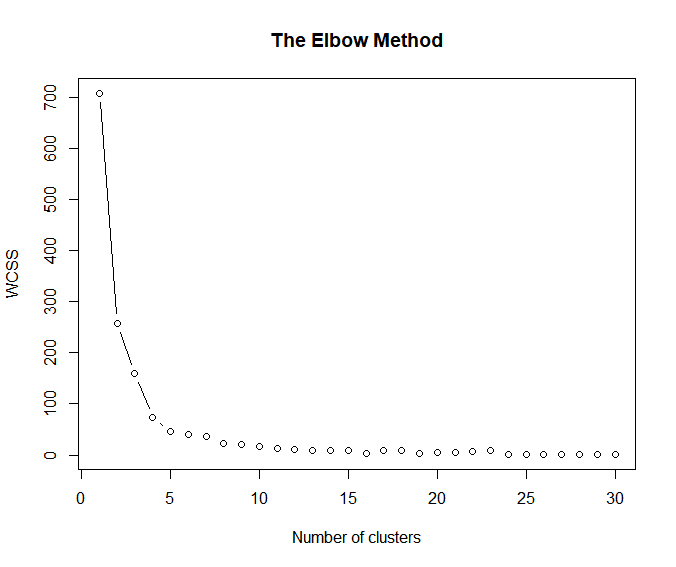
**4.3.3.1 Applying k means clustering:**

**Output:**

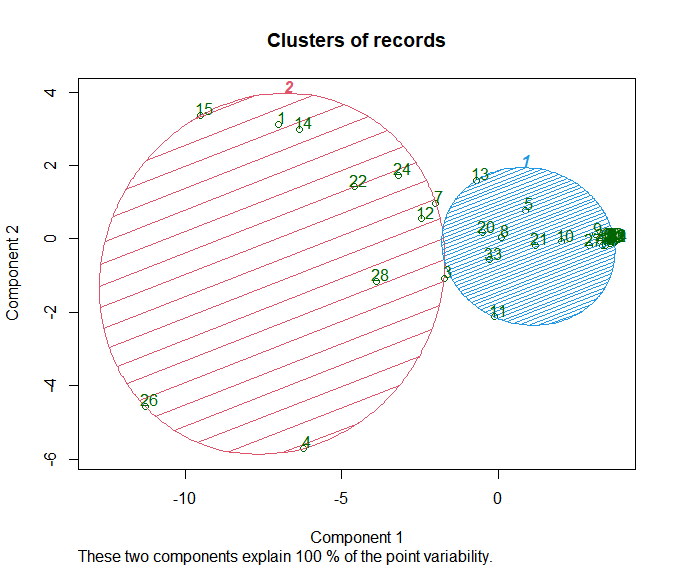
****

PC1explains 87.99 % variability of the data.

PC2 explains 7.96 % variability of the data.



From the above elbow Method we can conclude that optimal number of clusters is 2.

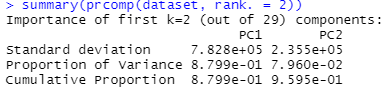


**Conclusion:**

1. As we know that the data set on which we applied clustering has 35 records, each record corresponding to a state or union territory, so by applying clustering we have successfully obtained two separable clusters. The states are grouped on the basis of total number of crimes committed in those states in the period of 2001 to 2012.
2. We can observe that cluster 1 contains majority of the grouped records with 24 records in it and cluster 2 has 11 records in it.
3. Cluster 1 = {Andhra Pradesh, Bihar, Assam, Goa, Jharkhand, Kerala, Madhya Pradesh, Punjab, Sikkim, Tripura, Uttarakhand}.
4. Cluster 2 = {ARUNACHAL PRADESH, CHHATTISGARH, HARYANA, HIMACHAL PRADESH, JAMMU & KASHMIR,KARNATAKA,MAHARASHTRA,MANIPUR,MEGHALAYA,MIZORAM,NAGALAND,ODISHA,RAJASTHAN,TAMIL NADU,UTTAR PRADESH,WEST BENGAL,A & N ISLANDS,CHANDIGARH,D & N HAVELI,DAMAN & DIU,DELHI UT,LAKSHADWEEP,PUDUCHERRY}
5. The relation between principal components and original variables can be analyzed using ggbiplot.

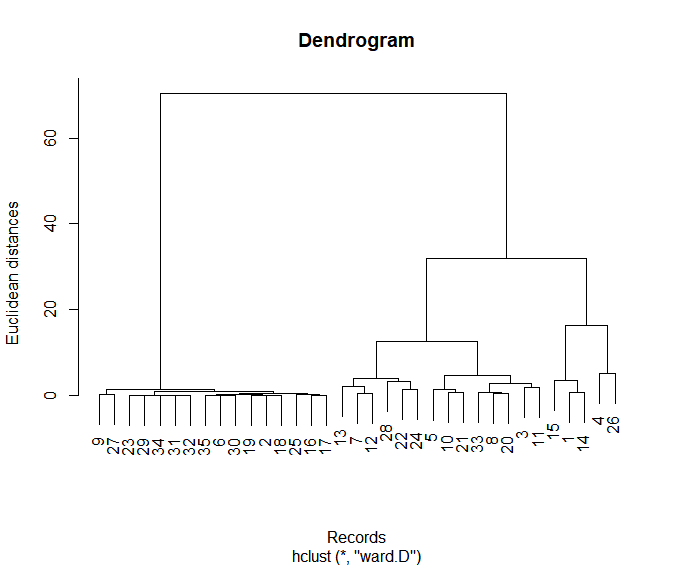
**4.3.3.2 Applying hierarchical clustering:**

**Output:**

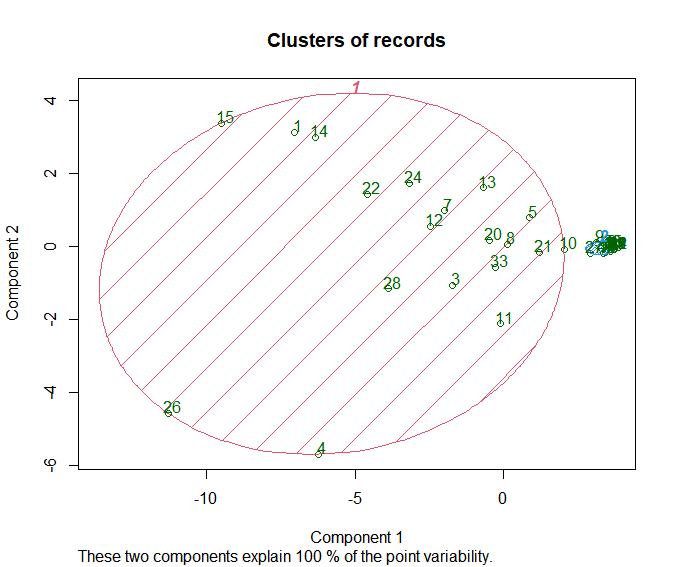
****

PC1explains 87.99 % variability of the data.

PC2 explains 7.96 % variability of the data.



From the above dendrogram we can conclude that the optimal number of clusters/centres is 2.



**Conclusion:**

1. As we know that the data set on which we applied clustering has 35 records, each record corresponding to a state or union territory, so by applying clustering we have successfully obtained two separable clusters. The states are grouped on the basis of total number of crimes committed in those states in the period of 2001 to 2012.
2. Clusters obtained by hierarchical clustering has better clusters, they easily and 100 percent linearly separable.
3. Cluster 1 has 18 records, hence cluster 2 has 17 records.
4. The relation between principal components and original variables can be analyzed using ggbiplot.
   1. **ICDA implementation of regression algorithms**

We have implemented regression algorithms to 3 data sets:

## 01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv:

Since it has 30 dependent variables so we built (has 3 independent variables, after dummy coding 844 independent variables):

1. 30 multiple regression models.
2. 30 Support Vector regression models.
3. 30 Decision tree regression models.
4. 30 Random Forest Regression models.
5. DataGroupedByYEAR.csv:

It has 1 independent variable year but has 30 dependent variables:

1. 30 Simple linear regression models.
2. 30 Polynomial regression models.
3. 30 Support Vector regression models
4. 30 Decision Tree regression models.
5. 30 Random Forest Regression models.
6. DataGroupedBySTATE\_Year\_Names.csv:
7. 30 multiple regression models.
8. 30 Support Vector regression models.
9. 30 Decision tree regression models.
10. 30 Random Forest Regression models.

R code we are going to use to evaluate r-squared and adjusted r-squared value of non linear model is as follows:

-------------------------------------------------------------------------

arr = array(dataset$dependentVariableName)

ssres = 0

sst = 0

yavg = mean(dataset$dependentVariableName)

ss2 <- array(predict(regressor, newdata = dataset))

for (i in seq(1,length(arr))){

ssres = ssres + (arr[i] - ss2[i])^2

}

for (i in seq(1,length(arr))){

sst = sst + (arr[i] - yavg)^2

}

R2 = 1 - (ssres/sst)

n = nrow(dataset)

AdjR2 = 1 - (1 - R2) \* ((n - 1)/ (n - 1 - 1))

---------------------------------------------------------------------------

Here R2 and AdjR2 are the r-squared and adjusted r-squared value.

This section is divided further into sub sections and sub sections of subsections on the basis of different data set and regression models used.

**4.4.1 Implementation of 4 different regression models on:**

**01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv data set, with dependent variable = TOTAL\_IPC\_CRIMES:**

**Note :** MAPE (Mean Absolute percentage error) is used as a parameter to compare the performance of the models on test\_ set:

IMG_256

Where At is the actual value and Ft is the forecast value and n is total number of data points.

**R code:**

------------------------------------------------------------------------

arr1 = array(test\_set$dependentVariableName)

ssres1 = 0

ss <- array(predict(regressor, newdata = test\_set))

for (i in seq(1,length(arr1))){

ssres1 = ssres1 + ((abs(arr1[i] - ss[i]))/abs(arr1[i]))

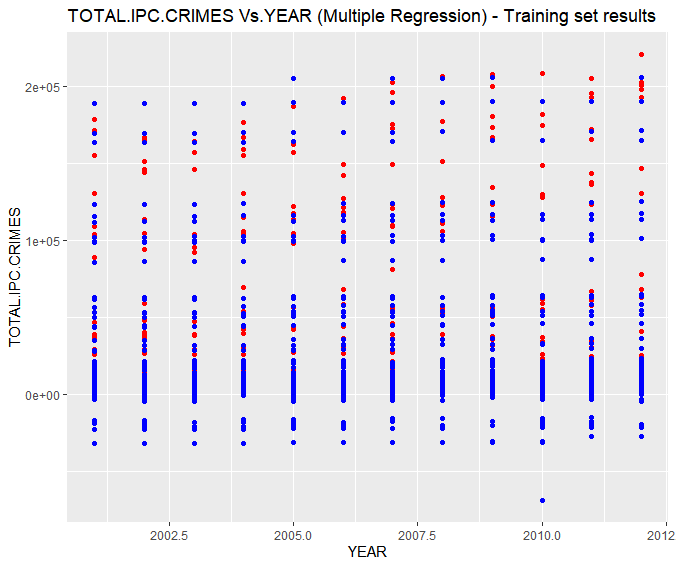
}

m = nrow(test\_set)

mape = (ssres1/m)

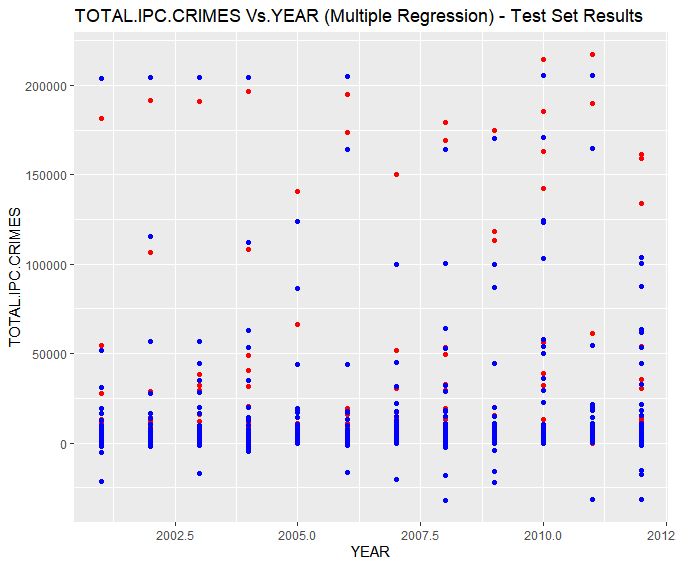
**4.4.1.1 Multiple regression:**

**Output:**



**Conclusion:**

1. In the above Number of TOTAL.IPC.CRIMES vs. YEAR plot, red points are the actual data points that were observed and belongs to training\_set whereas blue points are the predicted values predicted by our Multiple Linear Regression model.
2. Although this was the best option to visualize predicted and observed but still there is a big problem with this graph.
3. Since, the training\_set contains each year value repeated more than 800\*2/3 times corresponding to distinct state, district pair. Therefore, we got almost more than 1 observed and predicted points corresponding to each year from 2001 to 2012.
4. This makes it tough to visualize that how good or poor the model fitted into the training\_set.
5. But no worries we have other parameters to check this, viz. R-squared value and Adjusted R-Squared value.



**Conclusion:**

1. In the above Number of TOTAL.IPC.CRIMES vs. YEAR plot, red points are the actual data points that were observed and belongs to test\_set whereas blue points are the predicted values predicted by our Multiple Linear Regression model on the test\_set.
2. Although this was the best option to visualize predicted and observed but still there is a big problem with this graph.
3. Since, the test\_set contains each year value repeated more than 800/3 times corresponding to distinct state, district pair. Therefore, we got almost more than 1 observed and predicted points corresponding to each year from 2001 to 2012.
4. This makes it tough to visualize that how good or poor the model fitted into the test\_set.
5. But no worries we have other parameters to check this, viz. R-squared value and Adjusted R-Squared value.

>summary(regressor)

Call:

lm(formula = TOTAL.IPC.CRIMES ~ ., data = training\_set)

Residuals:

Min 1Q Median 3Q Max

-45745 -438 -26 383 97210

Residual standard error: 5487 on 6377 degrees of freedom

Multiple R-squared: 0.92, Adjusted R-squared: 0.9095

1. statistic: 87.83 on 835 and 6377 DF, p-value: < 2.2e-16

**Conclusion:**

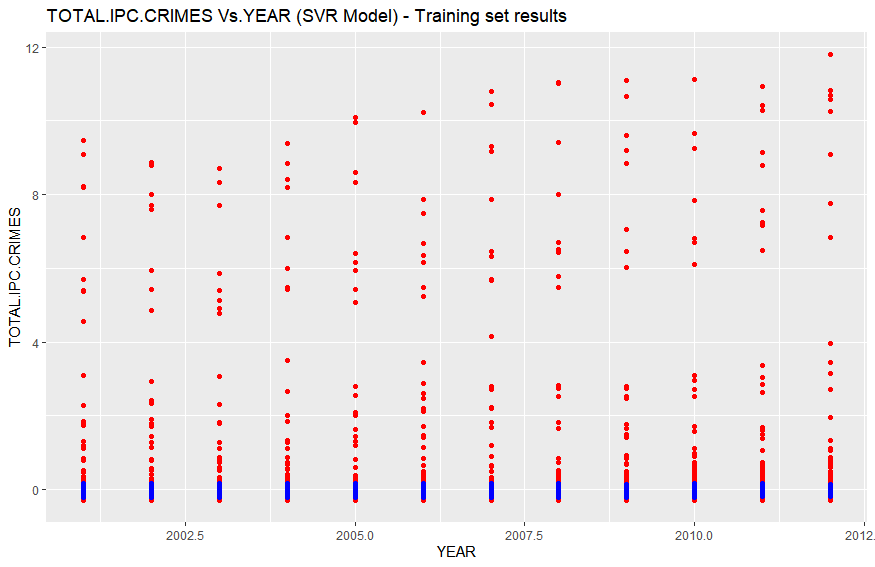
1. From the above output we can observe that majority of the variables out of 866 independent variables have p-value less than 0.05, hence, they are highly contribute to the model.
2. We have, R-squared value = 0.92 and adjusted R-squared value = 0.9095.

**>mape**

1. 8.667558

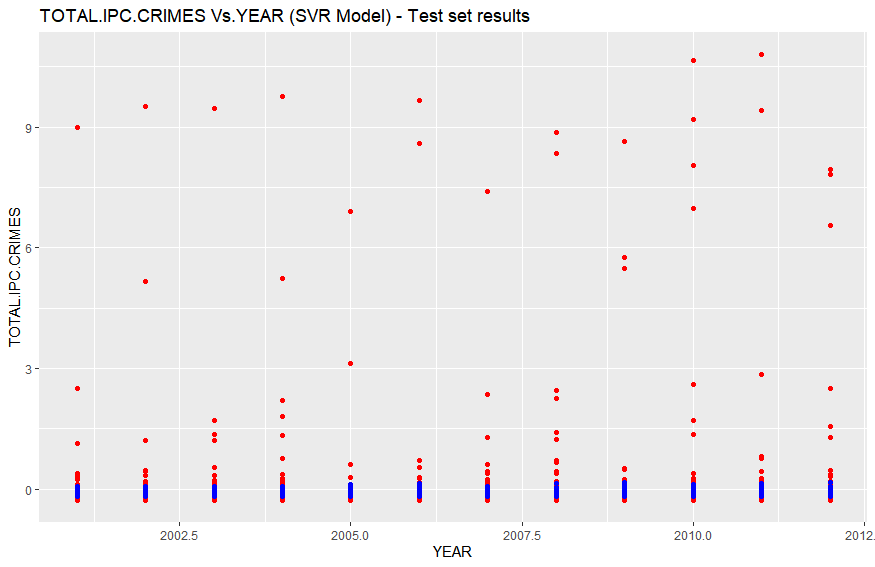
**4.4.1.2 Support vector regression:**

**Output:**



**Conclusion:**

1. In the above Number of TOTAL.IPC.CRIMES vs. YEAR plot, red points are the actual data points that were observed and belongs to training\_set whereas blue points are the predicted values predicted by our Support Vector Regression model.
2. Although this was the best option to visualize predicted and observed but still there is a big problem with this graph.
3. Since, the training\_set contains each year value repeated more than 800\*2/3 times corresponding to distinct state, district pair. Therefore, we got almost more than 1 observed and predicted points corresponding to each year from 2001 to 2012.
4. This makes it tough to visualize that how good or poor the model fitted into the training\_set.
5. But no worries we have other parameters to check this, viz. R-squared value and Adjusted R-Squared value, which in this case we will calculate through our code.



1. In the above Number of TOTAL.IPC.CRIMES vs. YEAR plot, red points are the actual data points that were observed and belongs to test\_set whereas blue points are the predicted values predicted by our Support Vector Regression model on the test\_set.
2. Although this was the best option to visualize predicted and observed but still there is a big problem with this graph.
3. Since, the test\_set contains each year value repeated more than 800/3 times corresponding to distinct state, district pair. Therefore, we got almost more than 1 observed and predicted points corresponding to each year from 2001 to 2012.
4. This makes it tough to visualize that how good or poor the model fitted into the test\_set.
5. But no worries we have other parameters to check this, viz. R-squared value and Adjusted R-Squared value, which in this case we will calculate through our code.

**> mape**

1. 18.15701

**>R2**

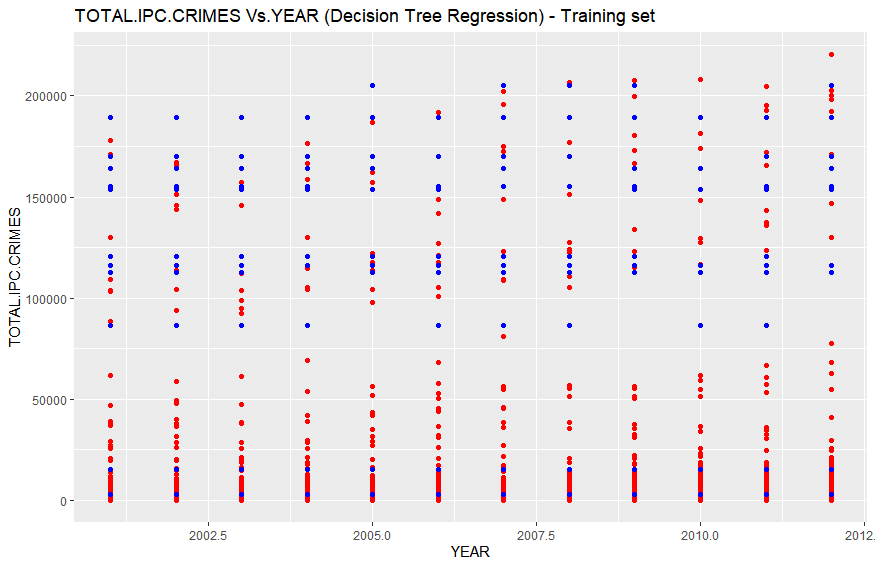
1. 0.0453099

**>AdjR2**

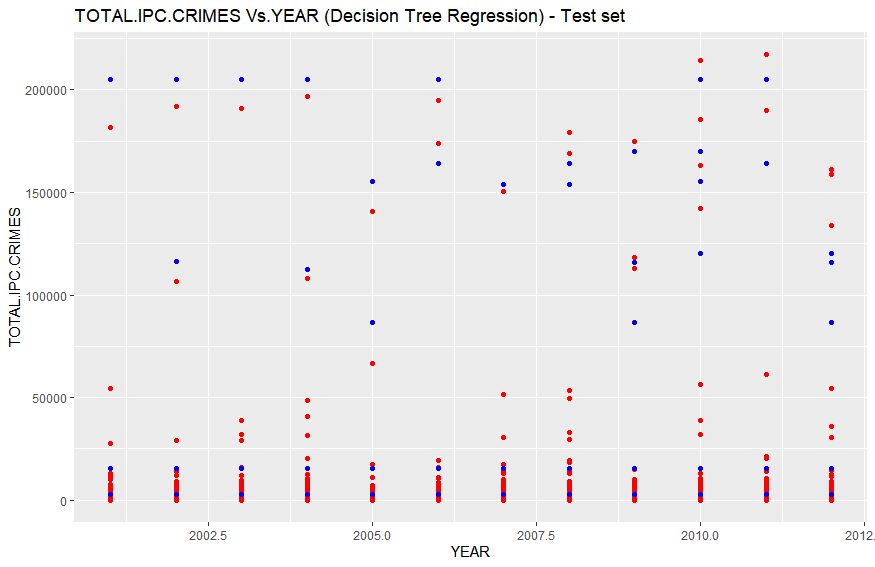
1. 0.04517751

**4.4.1.3 Decision Tree:**

**Output:**



1. In the above Number of TOTAL.IPC.CRIMES vs. YEAR plot, red points are the actual data points that were observed and belongs to training\_set whereas blue points are the predicted values predicted by our Decision Tree Regression model.
2. Although this was the best option to visualize predicted and observed but still there is a big problem with this graph.
3. Since, the training\_set contains each year value repeated more than 800\*2/3 times corresponding to distinct state, district pair. Therefore, we got almost more than 1 observed and predicted points corresponding to each year from 2001 to 2012.
4. This makes it tough to visualize that how good or poor the model fitted into the training\_set.
5. But no worries we have other parameters to check this, viz. R-squared value and Adjusted R-Squared value, which in this case we will calculate through our code.



**Conclusion:**

1. In the above Number of TOTAL.IPC.CRIMES vs. YEAR plot, red points are the actual data points that were observed and belongs to test\_set whereas blue points are the predicted values predicted by our Decision Tree Regression model on the test\_set.
2. Although this was the best option to visualize predicted and observed but still there is a big problem with this graph.
3. Since, the test\_set contains each year value repeated more than 800/3 times corresponding to distinct state,district pair. Therefore, we got almost more than 1 observed and predicted points corresponding to each year from 2001 to 2012.
4. This makes it tough to visualize that how good or poor the model fitted into the test\_set.
5. But no worries we have other parameters to check this, viz. R-squared value and Adjusted R-Squared value, which in this case we will calculate through our code.

**>mape**

1. 8.739453

**> R2**

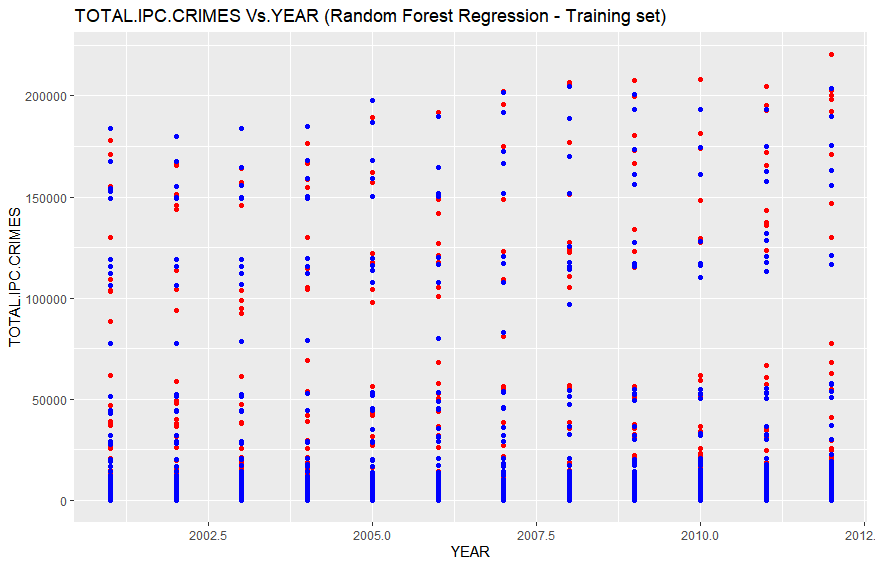
1. 0.9127783

**> AdjR2**

1. 0.9127663

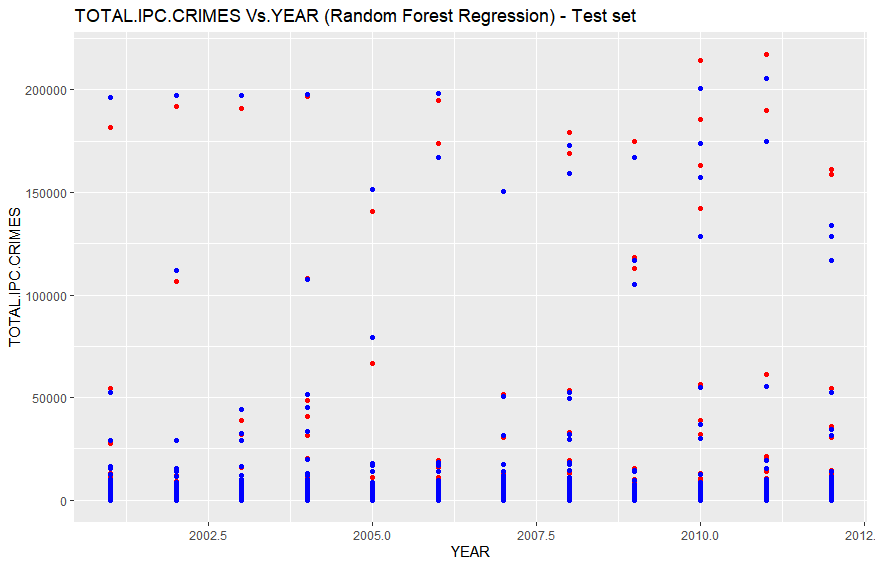
**4.4.1.4 Random Forest Regression:**

**Output:**



**Conclusion:**

1. In the above Number of TOTAL.IPC.CRIMES vs. YEAR plot, red points are the actual data points that were observed and belongs to training\_set whereas blue points are the predicted values predicted by our Random Forest Regression model.
2. Although this was the best option to visualize predicted and observed but still there is a big problem with this graph.
3. Since, the training\_set contains each year value repeated more than 800\*2/3 times corresponding to distinct state, district pair. Therefore, we got almost more than 1 observed and predicted points corresponding to each year from 2001 to 2012.
4. This makes it tough to visualize that how good or poor the model fitted into the training\_set.
5. But no worries we have other parameters to check this, viz. R-squared value and Adjusted R-Squared value, which in this case we will calculate through our code.



**Conclusion:**

1. In the above Number of TOTAL.IPC.CRIMES vs. YEAR plot, red points are the actual data points that were observed and belongs to test\_set whereas blue points are the predicted values predicted by our Random Forest Regression model.
2. Although this was the best option to visualize predicted and observed but still there is a big problem with this graph.
3. Since, the test\_set contains each year value repeated more than 800\*2/3 times corresponding to distinct state, district pair. Therefore, we got almost more than 1 observed and predicted points corresponding to each year from 2001 to 2012.
4. This makes it tough to visualize that how good or poor the model fitted into the training\_set.
5. But no worries we have other parameters to check this, viz. R-squared value and Adjusted R-Squared value, which in this case we will calculate through our code.

**> mape**

[1] 0.2334625

**> R2**

[1] 0.9891969

**> AdjR2**

[1] 0.9891954

**COMPARING THE FOUR REGRESSION WE BUILT ON THE CHOSEN DATA SET**

|  |  |  |  |
| --- | --- | --- | --- |
| **Regression Model Used** | **R-Squared Value** | **Adjusted-R-Squared Value** | **MAPE** |
| Multiple Linear Regression | **0.92** | **0.9095** | **8.666558** |
| Support Vector Regression | **0.0453099** | **0.04517751** | **18.15701** |
| Decision Tree Regression | **0.9127783** | **0.9127663** | **8.739453** |
| Random Forest Regression | **0.989169** | **0.9891954** | **0.23** |

**Conclusion:**

1. We got the best fit with minimum error calculated on predicted values on test\_set with **Random Forest Regression model.**
2. We got worst fit and highest error with **SVR model.**
3. The order of best fitting models is as follows:

**SVR << MULTIPLE REGRESSION MODEL < DECISION TREE REGRESSION < Random Forest Regression**

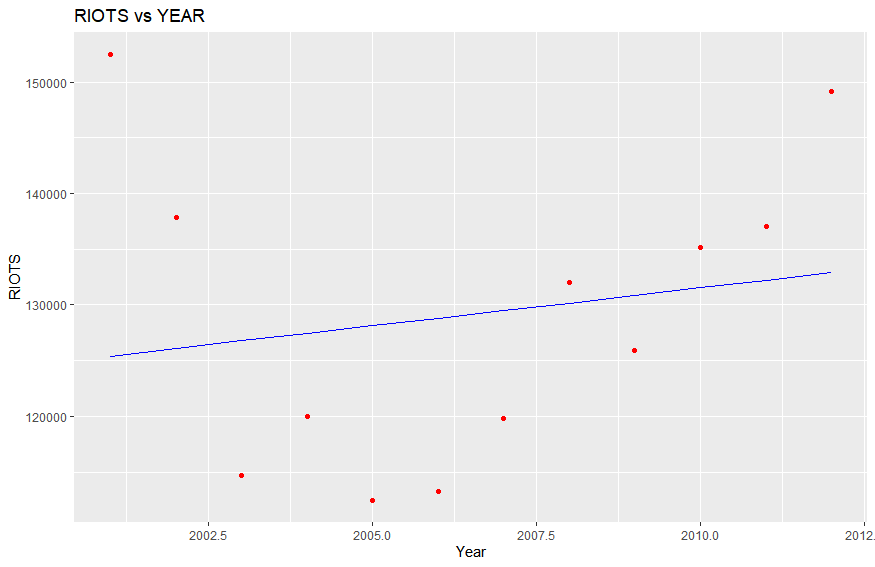
1. The order of models on the basis of decreasing in their error is as follows:

**SVR > DECISION TREE REGRESSION > MULTIPLE REGRESSION MODEL >> Random Forest Regression**

**4.4.2 Implementation of 5 different regression models on DataGroupedByYEAR.csv data set, with dependent variable RIOTS :**

**4.4.2.1 Simple Linear Regression:**

**Output:**



**Conclusion:**

1. We have only one plot this time, as we had only 12 rows, so we didn’t split the data set into training\_set and test\_set.
2. In the above plot we have red points as the actual observed points from the data set and blue line is the regression line that is plot on the basis of predicted values of number of riots cases per year from 2001 to 2012 by our model.
3. As we can observe that the simple linear model poorly fits into the data set with high residual values for most of the data points.

**> summary(regressor)**

Call:

lm(formula = RIOTS ~ YEAR, data = dataset)

Residuals:

Min 1Q Median 3Q Max

-15658 -10269 -1549 6543 27037

Coefficients:

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

(Intercept) -1235874.5 2362661.4 -0.523 0.612

YEAR 680.3 1177.5 0.578 0.576

Residual standard error: 14080 on 10 degrees of freedom

Multiple R-squared: 0.0323, Adjusted R-squared: -0.06447

1. statistic: 0.3338 on 1 and 10 DF, p-value: 0.5762

**Conclusion:**

1. We can see from the above output that the year independent variable has very high value 0.576.
2. Also the r-squared value 0.0323 and adjusted-r-squared value -0.06447 are too low that too wit negative adjusted-r-squared value which also implies that the model poorly fits into the data set.

**> y\_pred**

1

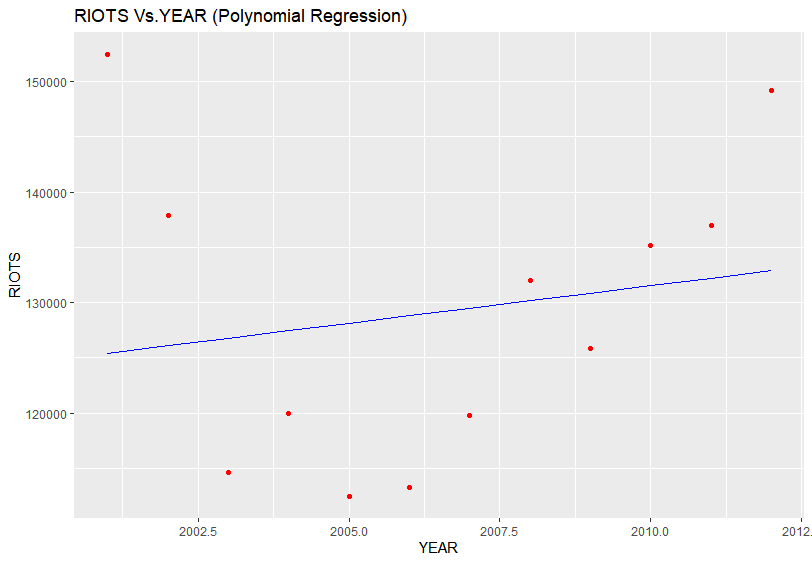
133570.8

**Conclusion:**

For the year 2013, number of riots cases registered is predicted 1,33,570 by the model.

**4.4.2.2 Polynomial Regression:**

**Output:**



**Conclusion:**

1. We have only one plot this time, as we had only 12 rows, so we didn’t split the data set into training\_set and test\_set.
2. In the above plot we have red points as the actual observed points from the data set and blue line is the regression line that is plot on the basis of predicted values of number of riots cases per year from 2001 to 2012 by our model.
3. As we can observe that the polynomial regression model poorly fits into the data set with high residual values for most of the data points.

**> predict(poly\_reg, data.frame(YEAR = 2013,**

**+ YEAR1 = 2013^2,**

**+ YEAR2 = 2013^3,**

**+ YEAR3 = 2013^4))**

1

133570.8

**> summary(regressor)**

Call:

lm(formula = RIOTS ~ YEAR, data = dataset)

Residuals:

Min 1Q Median 3Q Max

-15658 -10269 -1549 6543 27037

Coefficients:

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

(Intercept) -1235874.5 2362661.4 -0.523 0.612

YEAR 680.3 1177.5 0.578 0.576

Residual standard error: 14080 on 10 degrees of freedom

Multiple R-squared: 0.0323, Adjusted R-squared: -0.06447

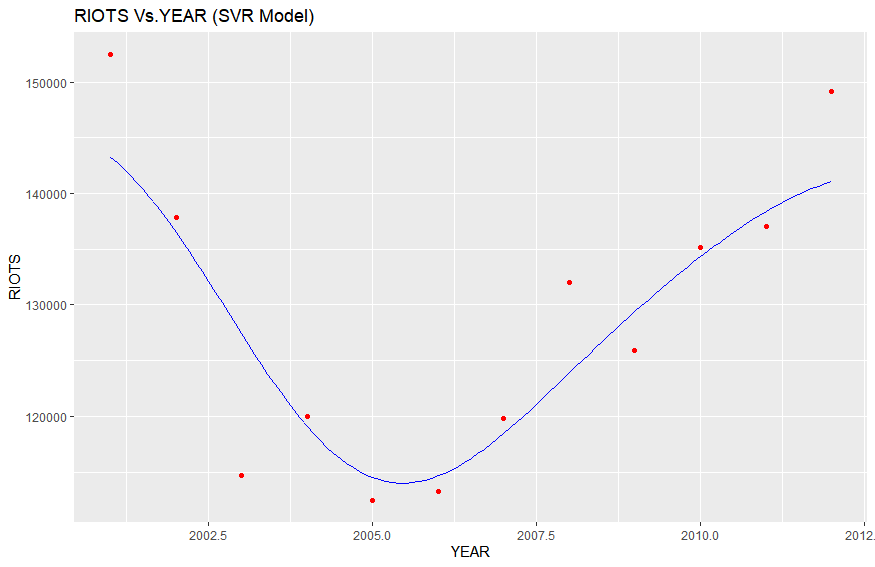
1. statistic: 0.3338 on 1 and 10 DF, p-value: 0.5762

**Conclusion:**

1. As we can see that YEAR1, YEAR2 AND YEAR3 does not play any role in the model, hence removed from the summary of the polynomial regressor.
2. We can see that even YEAR also has very high p-value of 0.576.
3. Also the r-squared = 0.0323 and adjusted r-squared = - 0.06447 values are too low, implying model barely fits into the data set.

**4.4.2.3 Support Vector Regression:**

**Output:**



1. We have only one plot this time, as we had only 12 rows, so we didn’t split the data set into training\_set and test\_set.
2. In the above plot we have red points as the actual observed points from the data set and blue line is the regression line that is plot on the basis of predicted values of number of riots cases per year from 2001 to 2012 by our model.
3. As we can observe that the support vector regression model very nicely fits into the data set.

**> R2**

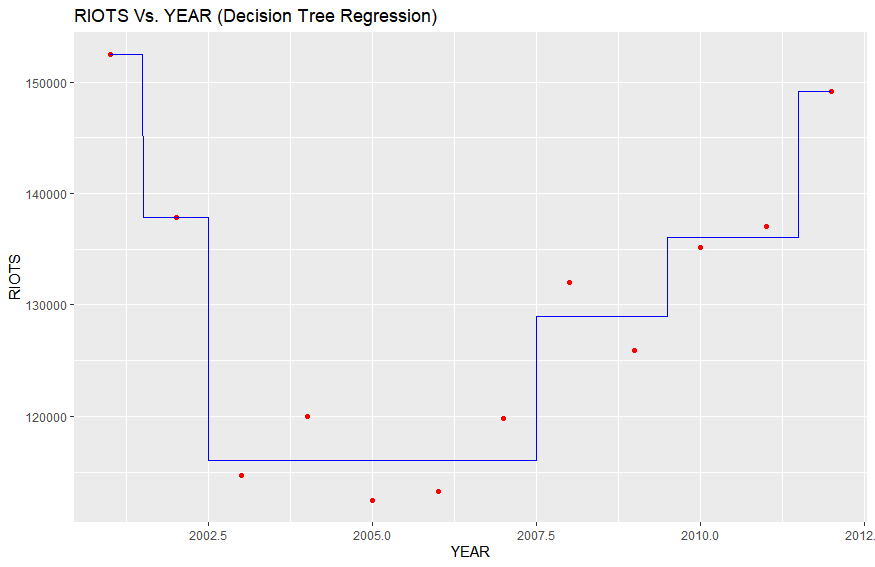
1. 0.8026651

**> AdjR2**

[1] 0.7829316

**4.4.2.4 Decision Tree Regression:**

**Output:**



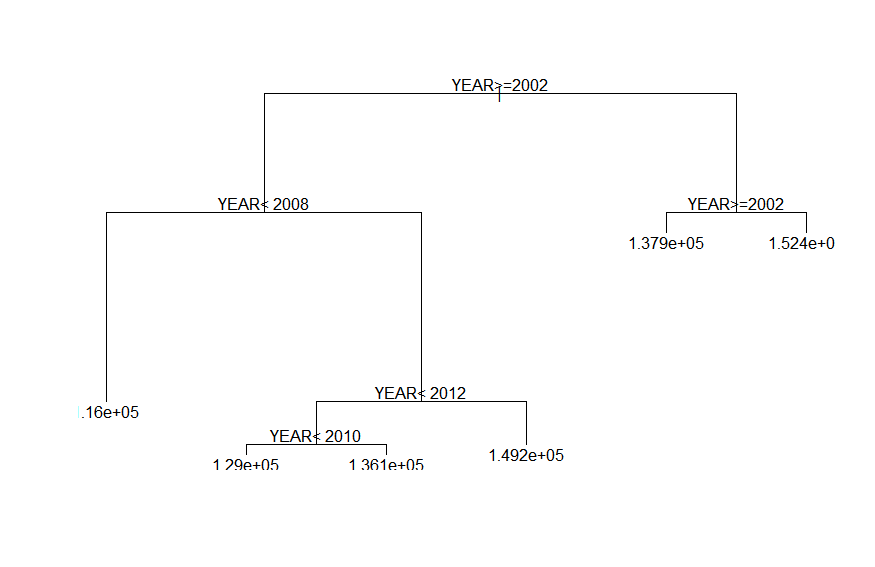
1. We have only one plot this time, as we had only 12 rows, so we didn’t split the data set into training\_set and test\_set.
2. In the above plot we have red points as the actual observed points from the data set and blue line is the regression line that is plot on the basis of predicted values of number of riots cases per year from 2001 to 2012 by our model.
3. As we can observe that the decision tree regression model very nicely fits into the data set with minimum split value = 2.

**> R2**

1. 0.9646263

**> AdjR2**

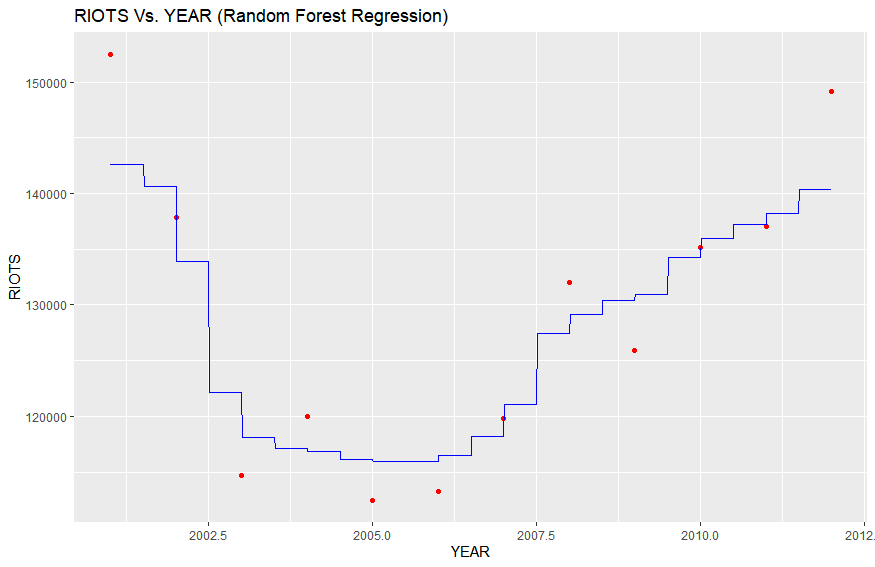
[1] 0.9610889



Above is the decision tree formed by our model on the basis of which it predicts the target variable.

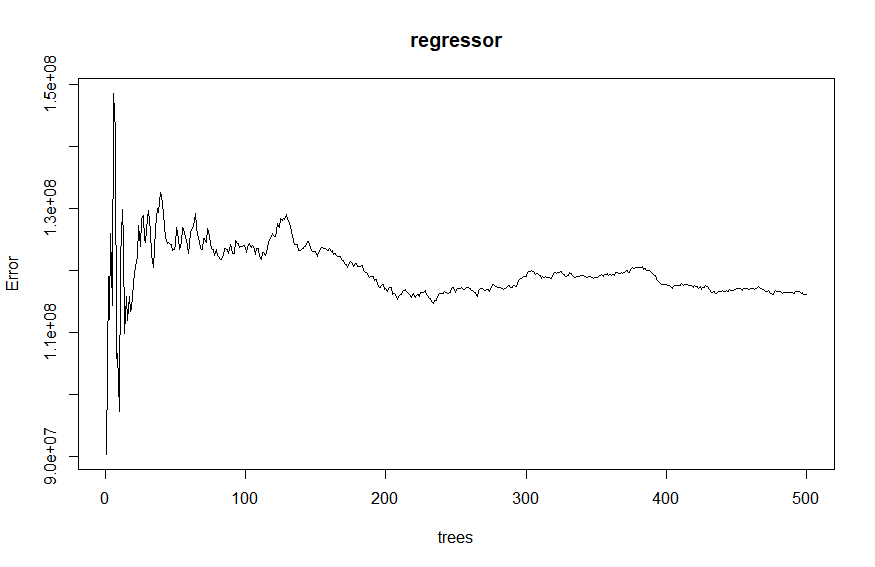
**4.4.2.5 Random Forest regression:**

**Output:**



**Conclusion:**

1. We have only one plot this time, as we had only 12 rows, so we didn’t split the data set into training\_set and test\_set.
2. In the above plot we have red points as the actual observed points from the data set and blue line is the regression line that is plot on the basis of predicted values of number of riots cases per year from 2001 to 2012 by our model.
3. As we can observe that the Random Forest regression model very nicely fits into the data set with total number of 500 different regression trees calculated and then taken average of the predicted value of target variable by all the 500 regression trees.



Above plot is between different number regression trees made by the random forest algorithm and the respective error.

**> R2**

1. 0.8484668

**> AdjR2**

1. 0.8333135

**Comparing the five regression model that we built on the data set**

|  |  |  |
| --- | --- | --- |
| **Regression Model Used** | **R-Squared Value** | **Adjusted- R Squared Value** |
| Multiple Linear Regression | **0.0323** | **- 0.06447** |
| Polynomial Regression | **0.0323** | **- 0.06447** |
| Support Vector Regression | **0.8026651** | **0.782916** |
| Decision Tree Regression (min. Split = 2) | **0.9646263** | **0.9610889** |
| Random Forest Regression | **0.8484668** | **0.8333135** |

**Conclusion:**

1. We got the best fit with **Decision Tree Regression model.**
2. We got worst fit with **MLR and PLR.**
3. The order of best fitting models is as follows:

**MLR = PLR < SVR < Random Forest < Decision tree regression**

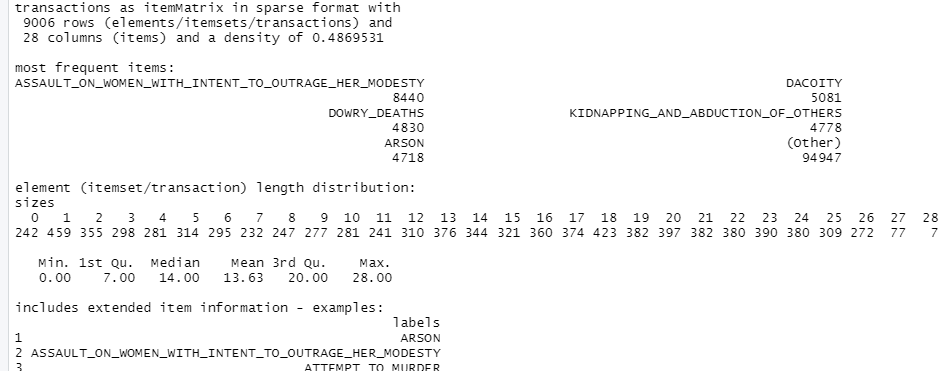
**4.5 ICDA implementing Association Rule Mining**

Dataset used: SparseDataToTransaction.csv

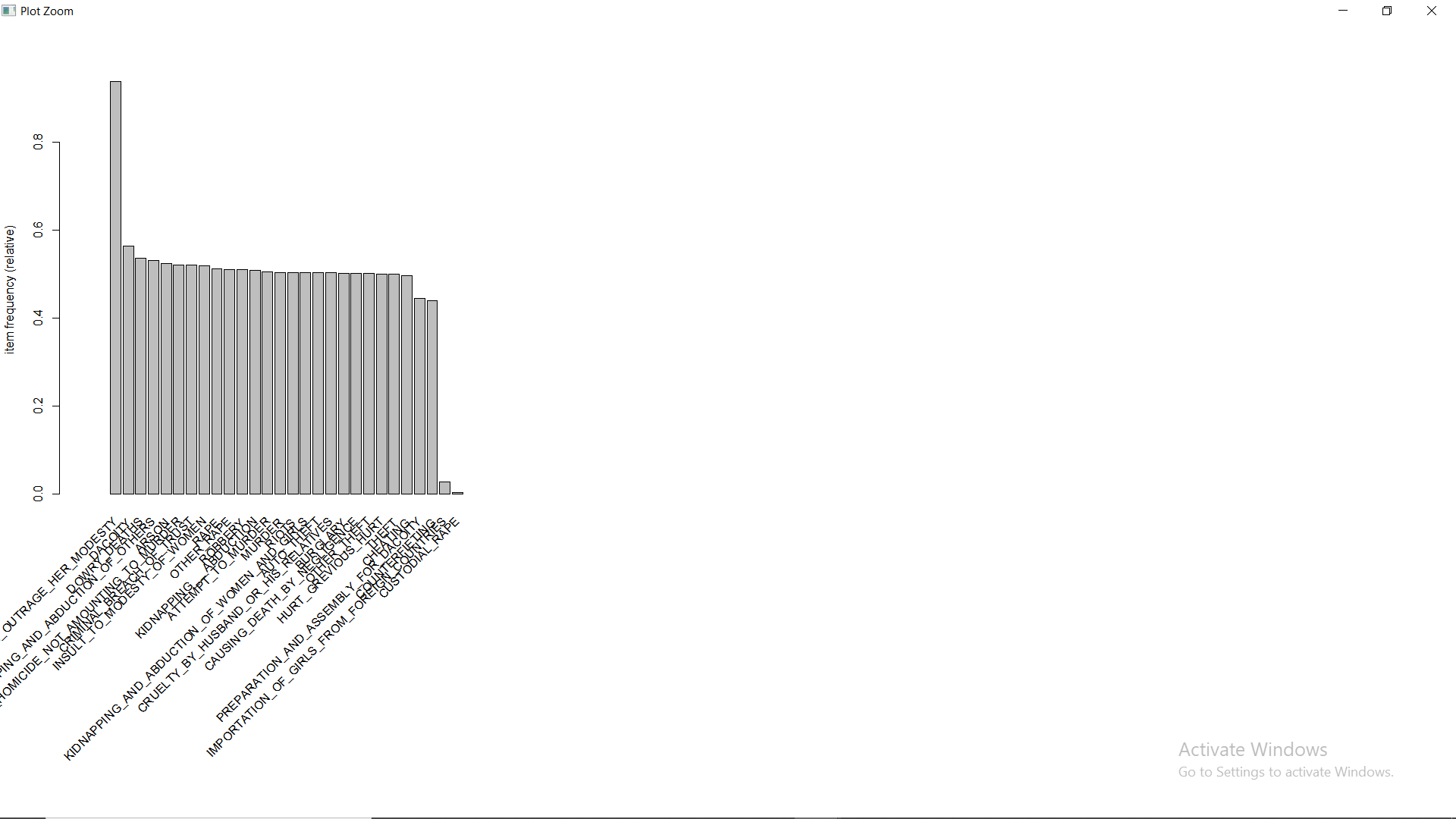
**4.5.1 Implementing Apriori algorithm on the data set**

**Output:**

**> summary(dataset)**



**> itemFrequencyPlot(dataset, topN = 100)**

****

**Conclusion:**

On the basis of transaction data set we produced from the data set:

* The crime that occurred on high scale in most of the part in the country is: ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY.
* Followed by crimes: DOWRY\_DEATHS, ARSON, DACOITY, KIDNAPPING AND ABDUCTION OF OTHERS and so on.
* With 242 representing at that are in a specific year no crime committed greater than the median value of that crime count in 12 years in different areas all over the country and 7 records representing just the opposite extreme.

**> rules = apriori(data = dataset, parameter = list(support = 0.3, confidence = 0.8))**

**Parameter specification:**

confidence minval smax arem aval originalSupport maxtime support minlen maxlen target ext

0.8 0.1 1 none FALSE TRUE 5 0.3 1 10 rules TRUE

**Algorithmic control:**

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 2701

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[28 item(s), 9006 transaction(s)] done [0.01s].

sorting and recoding items ... [26 item(s)] done [0.00s].

creating transaction tree ... done [0.00s].

checking subsets of size 1 2 3 4 5 done [0.04s].

writing ... [1852 rule(s)] done [0.00s].

creating S4 object ... done [0.00s]

**Minimum Support parameter value is calculated on the basis on the crimes whose crime count greater than median at least in 225 region,year pair out of 9008 records in 12 years,**

**225\*12/9008 ~ 0.3**

On the basis of minimum support = 0.3 and confidence = 0.8 we obtained 1852 rules, with top rules based on their lift value:

**> # Visualising the results**

> inspect(sort(rules, by = 'lift')[1:10])

Lhs rhs support confidence coverage lift count

[1] {AUTO\_THEFT,

CRIMINAL\_BREACH\_OF\_TRUST,

OTHER\_THEFT} => {THEFT} 0.3089052 0.9971326 0.3097935 1.992053 2782

[2] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

AUTO\_THEFT,

CRIMINAL\_BREACH\_OF\_TRUST,

OTHER\_THEFT} => {THEFT} 0.3070175 0.9971150 0.3079058 1.992018 2765

[3] {AUTO\_THEFT,

CHEATING,

OTHER\_THEFT} => {THEFT} 0.3159005 0.9957998 0.3172330 1.989391 2845

[4] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

AUTO\_THEFT,

CHEATING,

OTHER\_THEFT} => {THEFT} 0.3139018 0.9957732 0.3152343 1.989337 2827

[5] {AUTO\_THEFT,

CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES,

OTHER\_THEFT} => {THEFT} 0.3218965 0.9955357 0.3233400 1.988863 2899

[6] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

AUTO\_THEFT,

CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES,

OTHER\_THEFT} => {THEFT} 0.3200089 0.9955095 0.3214524 1.988811 2882

[7] {AUTO\_THEFT,

OTHER\_THEFT,

ROBBERY} => {THEFT} 0.3197868 0.9951624 0.3213413 1.988117 2880

[8] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

AUTO\_THEFT,

OTHER\_THEFT,

ROBBERY} => {THEFT} 0.3174550 0.9951270 0.3190095 1.988047 2859

[9] {AUTO\_THEFT,

BURGLARY,

OTHER\_THEFT} => {THEFT} 0.3315567 0.9940080 0.3335554 1.985811 2986

[10] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

AUTO\_THEFT,

BURGLARY,

OTHER\_THEFT} => {THEFT} 0.3295581 0.9939719 0.3315567 1.985739 2968

Any rule from the above rules can be explained as if a set of crimes satisfies the condition than another set of crimes will also satisfy the condition:

For ex:

[1] {AUTO\_THEFT,

CRIMINAL\_BREACH\_OF\_TRUST,

OTHER\_THEFT}

=> {THEFT}

The region, year pairs where

AUTO\_THEFT, CRIMINAL\_BREACH\_OF\_TRUST, OTHER\_THEFT

Count have values greater than the median or we can say where they occurred with high counts on those region, year pair THEFT crimes also occurred in high counts.

**4.5.2 Implementing Eclat algorithm on the data set**

Dataset used: SparseDataToTransaction.csv

**Output:**

**> rules = eclat(data = dataset, parameter = list(support = 0.3, minlen = 2))**

Eclat

parameter specification:

tidLists support minlen maxlen target ext

FALSE 0.3 2 10 frequent itemsets TRUE

algorithmic control:

sparse sort verbose

7 -2 TRUE

Absolute minimum support count: 2701

create itemset ...

set transactions ...[28 item(s), 9006 transaction(s)] done [0.01s].

sorting and recoding items ... [26 item(s)] done [0.00s].

creating bit matrix ... [26 row(s), 9006 column(s)] done [0.00s].

writing ... [1083 set(s)] done [0.00s].

Creating S4 object ... done [0.00s].

With minimum support value = 0.3 and constraint that the set should have at least two elements we obtained 1083 sets, out of which top 10 sets on the basis of their support value are:

> # Visualising the results

**> inspect(sort(rules, by = 'support')[1:10])**

items support transIdenticalToItemsets count

[1] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

DACOITY} 0.5450811 4909 4909

[2] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

DOWRY\_DEATHS} 0.5258716 4736 4736

[3] {ARSON,

ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY} 0.5173218 4659 4659

[4] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS} 0.5166556 4653 4653

[5] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

INSULT\_TO\_MODESTY\_OF\_WOMEN} 0.5146569 4635 4635

[6] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

CRIMINAL\_BREACH\_OF\_TRUST} 0.5129913 4620 4620

[7] {OTHER\_RAPE,

RAPE} 0.5111037 4603 4603

[8] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER} 0.5107706 4600 4600

[9] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

RAPE} 0.5065512 4562 4562

[10] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

OTHER\_RAPE,

RAPE}

Any of the above sets can be explained as the crimes in each set occur on high scale and together on the region,year pairs:

For e.g :

[10] {ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,

OTHER\_RAPE,

RAPE}

=> ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY, OTHER\_RAPE,

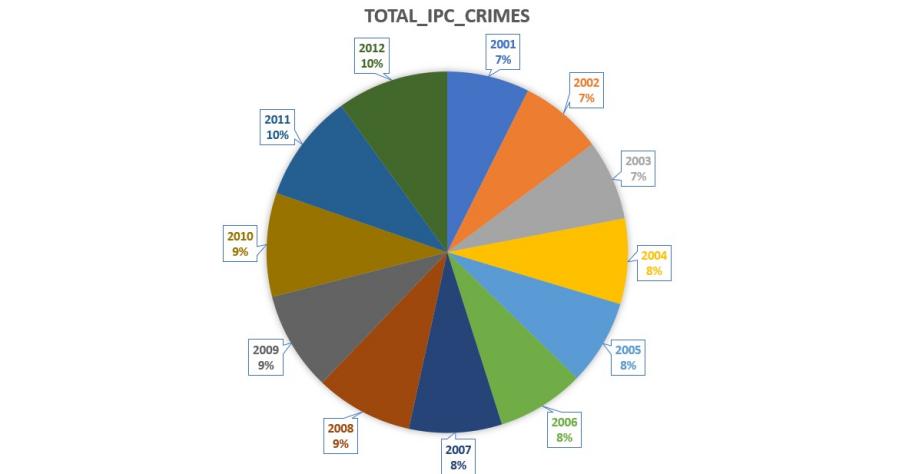
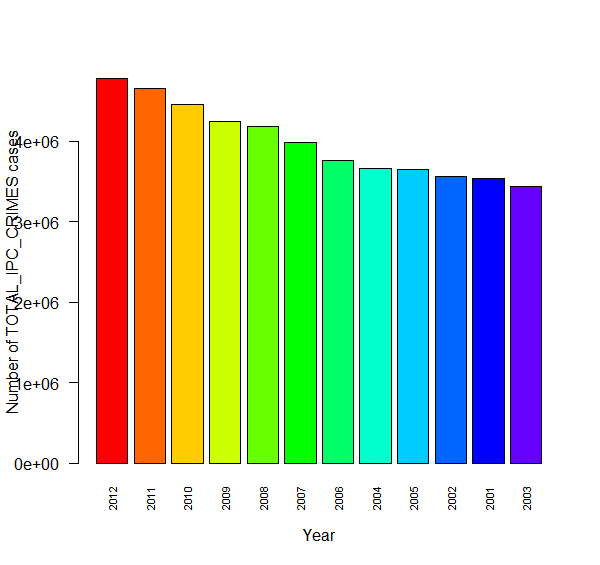
RAPE occurs on high scale together where any one of them occur on high scale.

**4.6 ICDA data visualization**

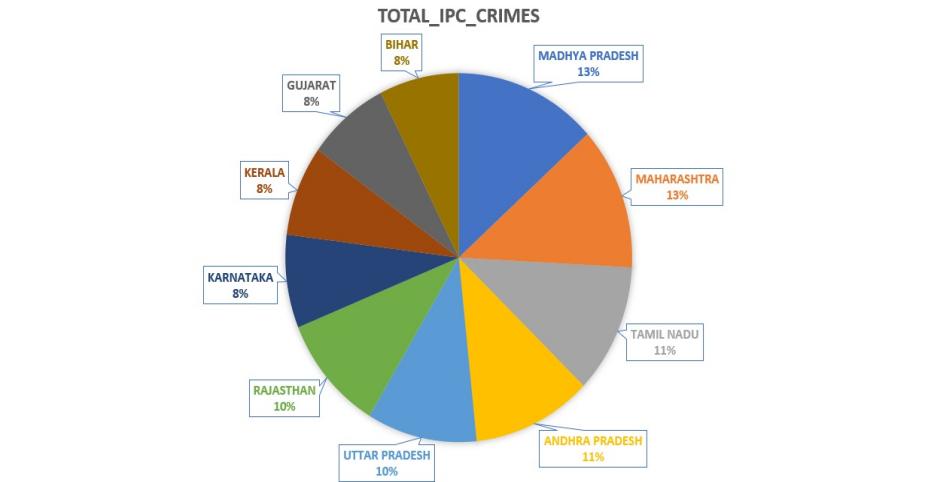
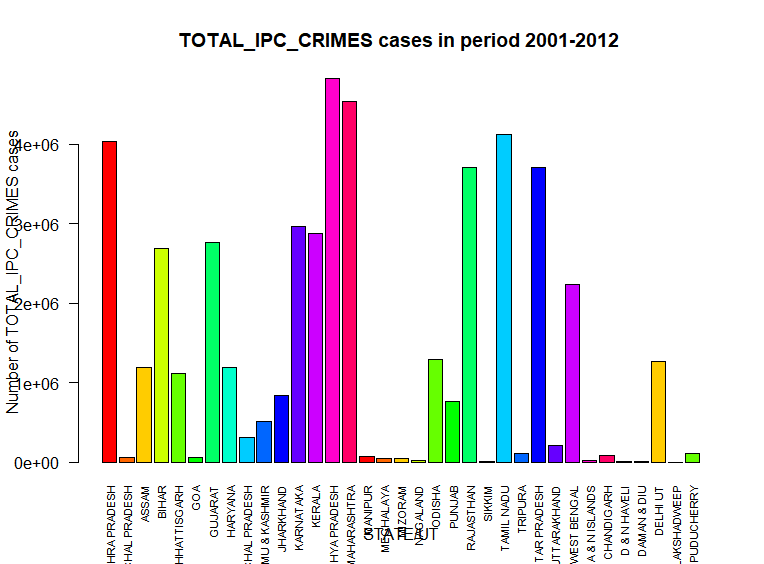
We have R codes and Microsoft excel data visualization tools to visualize the ICDA data set and other data sets derived from it using SQL. This section contains some sub sections based on different type of data visualization techniques.

**4.6.1 Pie Charts and bar plots:**

Microsoft excel is used for pie charts and R for bar plots.

**Fig. 4.6.1.1** Number of crimes in 2001-12 **Fig 4.6.1.2** Number of crimes vs year

**Fig. 4.6.1.3** Top 10 states, number of crimes in 2001-12 **Fig. 4.6.1.4** Number of crimes vs states in 2001-12

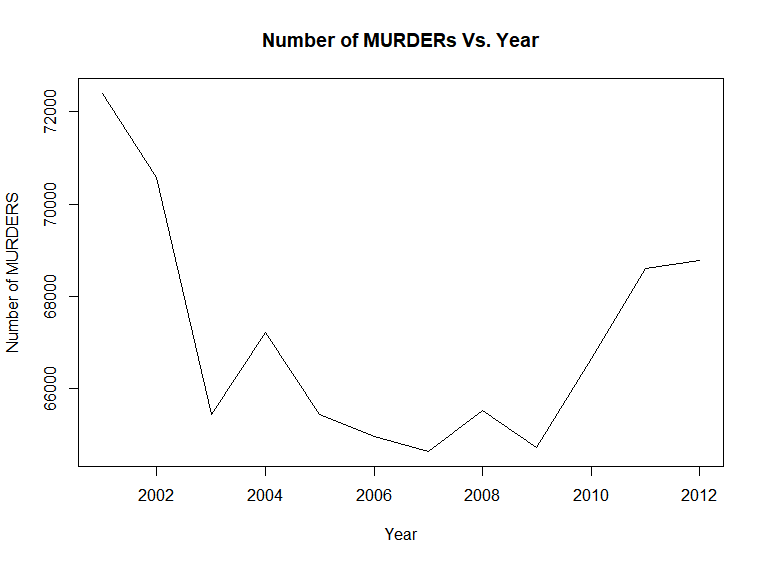
Fig 4.6.1.1 is the pie chart for number of total ipc crimes in different years. We can observe that in number of total ipc crime cases registered in 2001-2012, years 2011-2012 contributes the most with 10% each. 2008-2010 makes second position in the list of contributing 9% each to total number of ipc crimes registered in 2001-12 whereas 2004-2007 makes it third ranking by contributing 8% each to the total number of registered ipc crimes in 2001-12.

And years 2001-2004 contributing relatively the least with 7% each to the total number of ipc crimes in 2001-12.

Fig 4.6.1.2 is the bar plot on the same data which has been used to plot the pie chart in fig 4.6.1.1. From figure we can analyze that 2012 has the highest number of crimes registered among years 2001-12 whereas 2003 has the lowest. Ascending order of years based on total number of ipc crimes registered in those years is 2003, 2001, 2002, 2005, 2004, 2006, 2007, 2008, 2009, 2010, 2011, 2012. From this order we can conclude that there was monotonically increase in number of crimes from 2006-2012. Although we can also conclude that number of crimes have increased from 2001 to 2012 yearly.

**4.6.2 Line plots:**

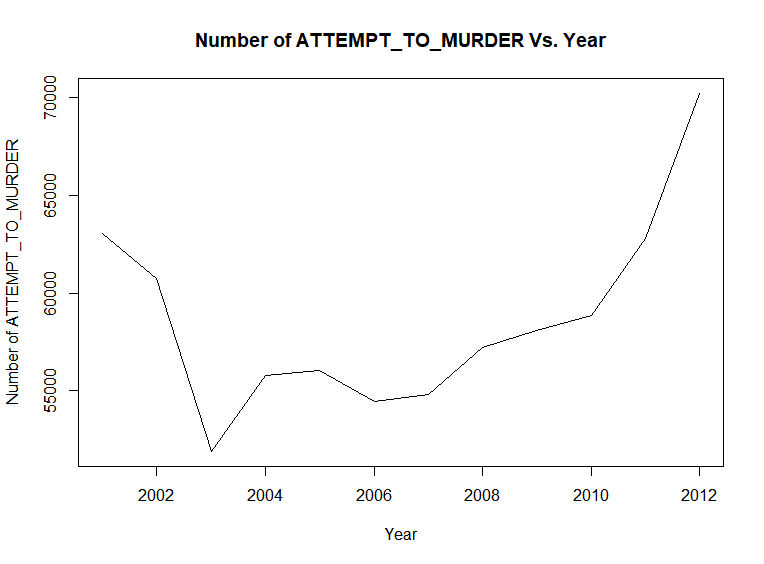
We have used R for line plots. Specific crime vs. Year plots and their analysis are given in this section.



**Fig 4.6.2.1** Number of murder cases vs. Year

**Analysis:**

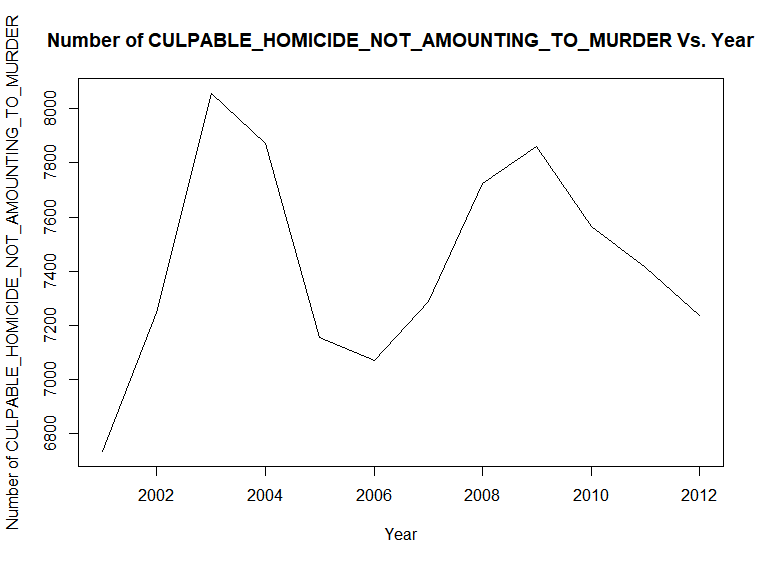
1. No Trend is observed.
2. From 2001 to 2012 count of murder cases are lower with a difference of around 4700.
3. In 2001 to 2012, Lowest number of murder cases was recorded in 2007 with around 64,400 registered cases.
4. In 2001 to 2012, Highest number of murder cases was recorded in year 2001 with around 72,400 registered cases.



**Fig 4.6.2.2** Number of attempt to murder cases vs. Year

**Analysis:**

1. From 2001 to 2003 decrease in attempt to murder cases is observed and from 2003 to 2012 increase in attempt to murder cases is observed expect with decrease in number of cases in 2005 and 2006.
2. Lowest number of cases registered in 2003 with in the period of 2001 to 2012.
3. Highest number of cases registered in 2012 with around 51,000 registered cases.

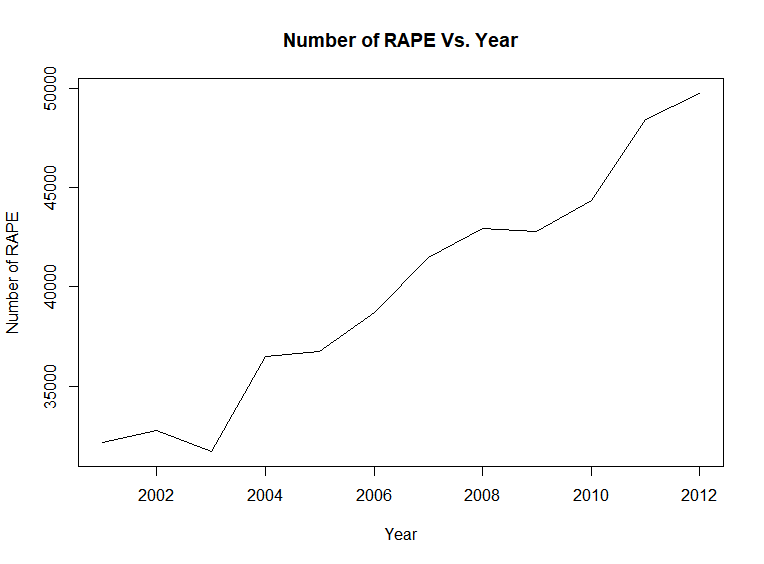


CYCLIC PATTERN

**Fig 4.6.2.3** Number of culpable homicide not amounting to murder cases vs. Year

**Analysis:**

1. A **Cyclic pattern** can be observed from the above plot.
2. Lowest number of culpable homicide not amounting to murder cases in 2001 with around 6660 numbers of registered cases.
3. Highest number of culpable homicide not amounting to murder cases observed in year 2003 with around 8080 registered cases.



Random movement

**Fig 4.6.2.4** Number of rape cases vs. Year

**Analysis:**

1. A **Trend** can be observed from the above plot, number of registered rape cases was increasing in the period of 2001 to 2012 in India.
2. But we have a **random movement** due to decrease of number of registered rape cases in between year 2002 and 2004 due to lowest number of registered rape cases around 36,000 cases registered in the year 2003.
3. Lowest number of registered rape cases, around 36,000 registered cases in 2003 and highest around 50,000 registered cases in 2012.
4. **One of the important observation is that the plot between total ipc crimes vs. Year and Number of rape cases vs. Year are identical:**

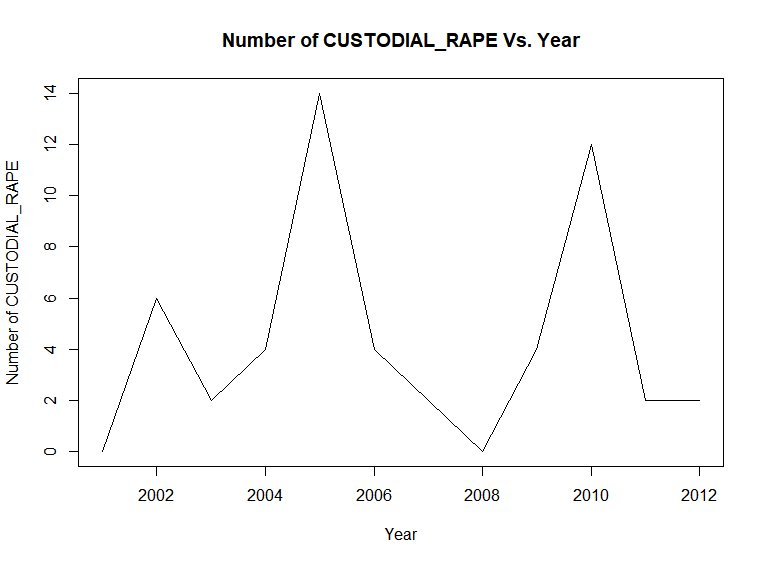
* With same trend.
* Random movement, highest y value and lowest y value at same corresponding years and that too with very less difference in ratio, following ratios:

Highest total ipc crime cases registered = 4768564/49752 = 95.8

Highest number of registered rape cases

Lowest rape cases registered =3432240/31694=108.2930523127406

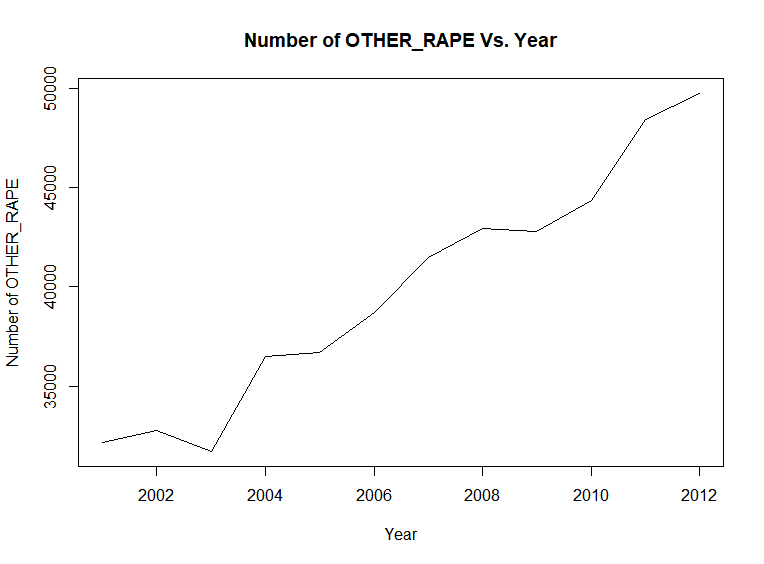
Lowest number of registered rape cases



**Fig 4.6.2.5** Number of custodial\_rape vs. Year

**Analysis:**

1. **Cyclic pattern** can be observed in above plot with random movements at 2008 and 2012.
2. Lowest number of custodial rape cases registered in 2003 with 0 number of cases registered.
3. Highest number of custodial rape cases registered in 2005 with 14 number of cases registered.



Random movement

**Fig 4.6.2.6** Number of other\_rape cases vs. Year

**Analysis:**

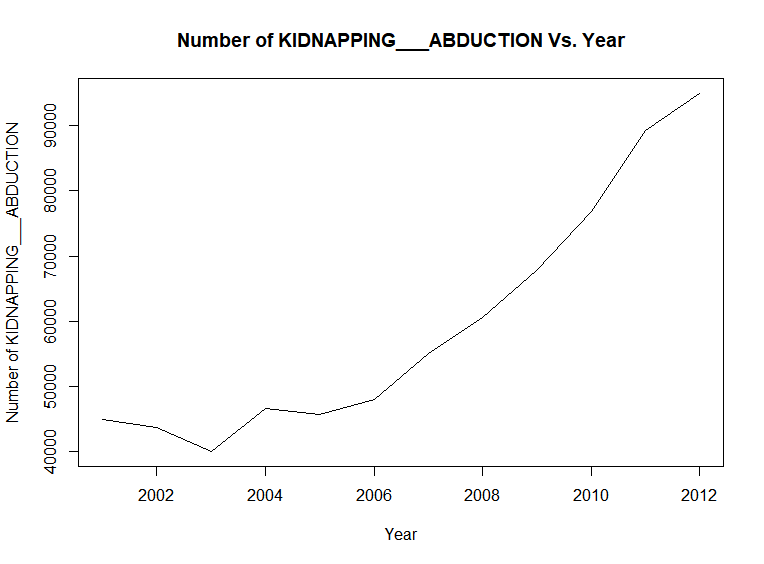
1)A **Trend** can be observed from the above plot, number of registered other\_rape cases was increasing in the period of 2001 to 2012 in India.

2) But we have a **random movement** due to decrease of number of registered other\_rape cases in between year 2002 and 2004 due to lowest number of registered other\_rape cases around 36,000 cases registered in the year 2003.

3) Lowest number of registered other\_rape cases, around 36,000 registered cases in 2003 and highest around 50,000 registered cases in 2012.

4) **Also, the number of rape cases vs. Year plot and number of other\_rape cases plot are identical and that also with very close y values.**

**Hence, total\_ipc\_crime count, registered rape cases count and registered other\_rape cases count are highly positively correlated.**

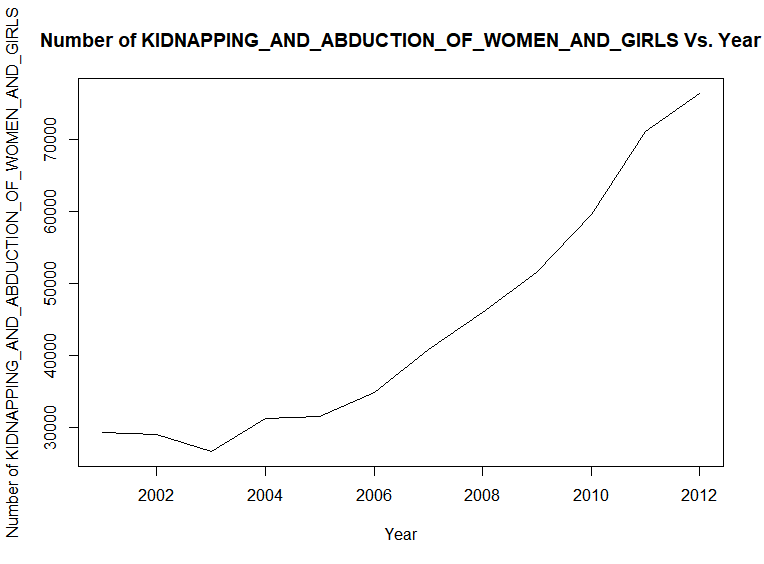


Random movement

**Fig 4.6.2.7** Number of kidnapping abduction cases vs. Year

**Analysis:**

1. A **Trend** can be observed from the above plot, number of registered Kidnapping and Abduction cases was increasing in the period of 2001 to 2012 in India.
2. But we have a **random movement** due to decrease of number of registered Kidnapping and Abduction cases in between year 2002 and 2004 due to lowest number of registered Kidnapping and Abduction cases around 40,000 cases registered in the year 2003.
3. Lowest number of registered Kidnapping and Abduction cases, around 40,000 registered cases in 2003 and highest around 95,000 registered cases in 2012.

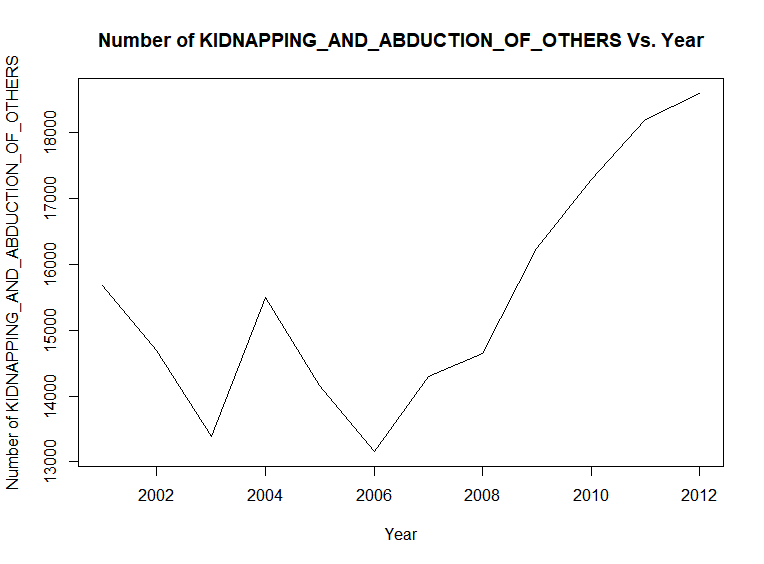


Random movement

**Fig 4.6.2.8** Number of kidnapping and abduction of women and girls cases vs. Year

**Analysis:**

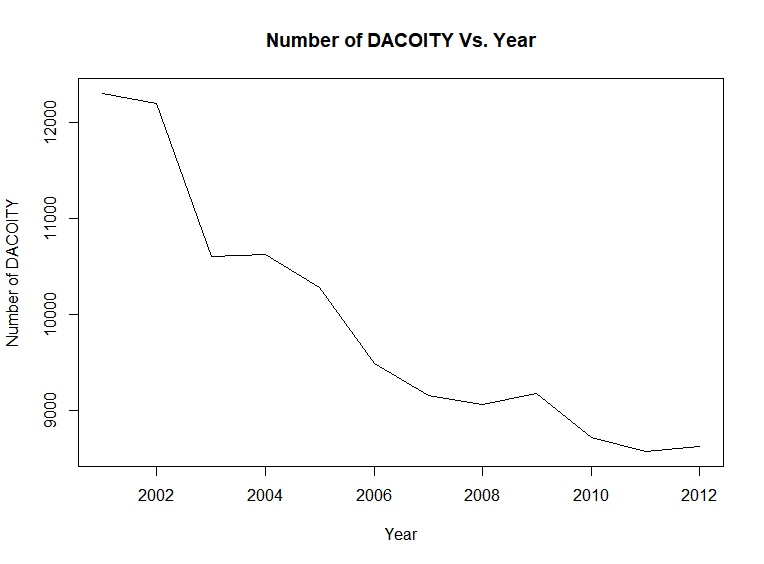
1. A **Trend** can be observed from the above plot, number of registered Kidnapping and Abduction of women and girls cases was increasing in the period of 2001 to 2012 in India.
2. But we have a **random movement** due to decrease of number of registered Kidnapping and Abduction of women and girls cases in between year 2002 and 2004 due to lowest number of registered Kidnapping and Abduction of women and girls cases around 34,000 cases registered in the year 2003.
3. Lowest number of registered Kidnapping and Abduction of women and girls cases, around 34,000 registered cases in 2003 and highest around 75,000 registered cases in 2012.
4. **The kidnapping and abduction vs. Year plot and kidnapping and abduction of women and girls vs. Year are identical, thus, number of cases registered under kidnapping and abduction and kidnapping and abduction of women and girls are highly correlated**



**Fig 4.6.2.9** Number of kidnapping and abduction of others cases vs. Year

**Analysis:**

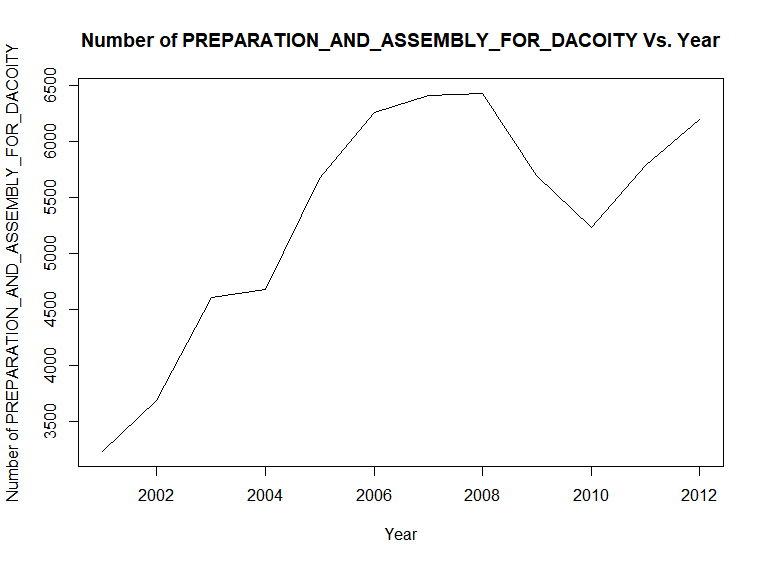
1. A Trend can be observed from Year 2006 to 2012, the number of registered kidnapping and abduction of others cases increasing throughout the period of 2006 to 2012.
2. Lowest number of cases registered in 2006, around 13,300 and highest number of cases registered in 2012, around 18,500 cases.



**Fig 4.6.2.10** Number of dacoity cases vs. Year

**Analysis:**

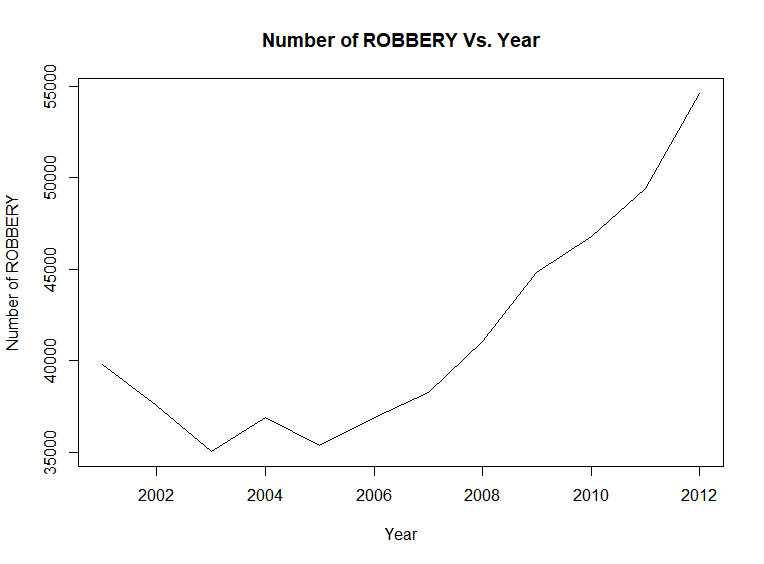
1. A **trend** can be observed from the above graph, number of cases registered under dacoity has decreased from 2001 to 2012.
2. Highest number of cases registered in 2001, around 12,500.
3. Lowest number of cases registered in 2012, around 9,400 cases.



**Fig 4.6.2.11** Number of preparation and assembly for dacoity cases vs. Year

**Analysis:**

1. A trend can be observed from 2001 to 2008, number of cases registered under preparation and assembly for dacoity was increasing in 2001 to 2008.
2. Lowest number of cases registered in 2001, around 3,550 cases.
3. Highest number of cases registered in 2008, around 6,125 cases.

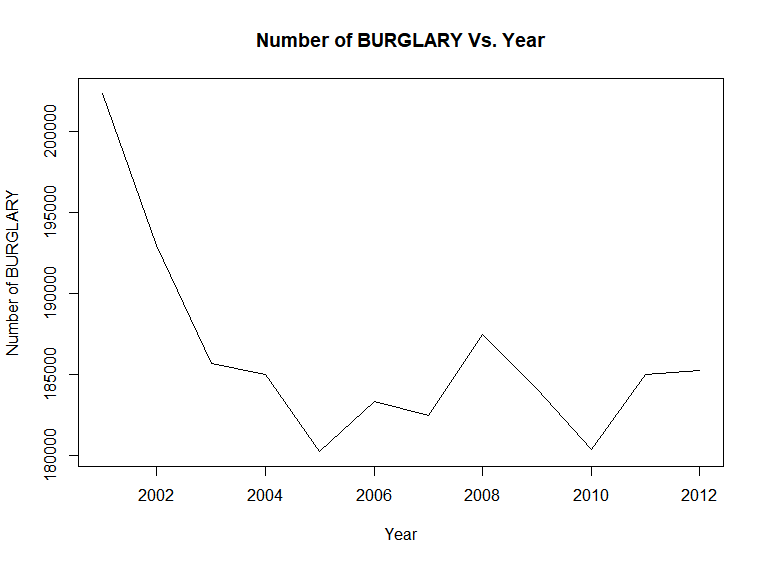


Random movement

**Fig 4.6.2.12** Number of robbery cases vs. Year

**Analysis:**

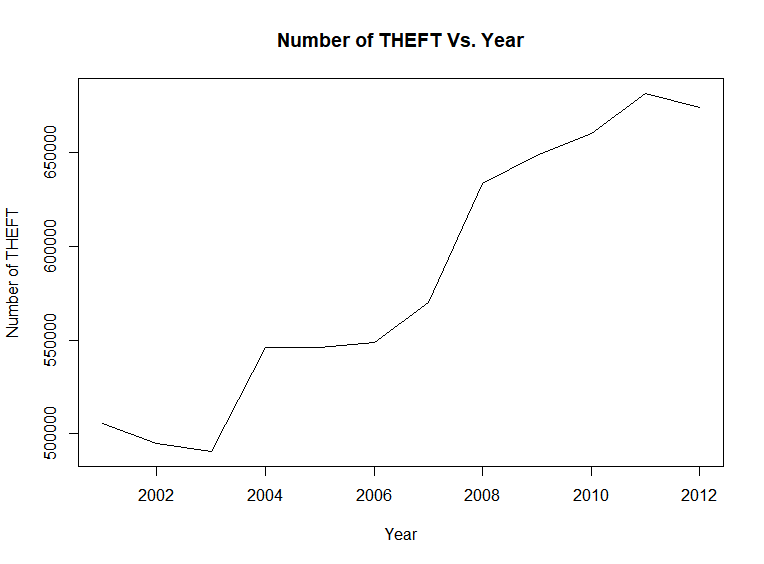
1. A trend can be observed in the above plot, number of cases registered under ROBBERY was increasing from 2001 to 2012 with few random movements due to some y values.
2. But we can say that the graph is monotonically increasing from 2005 to 2012.
3. Lowest number of cases registered in 2003, around 35,000 cases.
4. Highest number of cases registered in 2012, around 55,000 cases.



**Fig 4.6.2.13** Number of burglary cases vs. Year

**Analysis:**

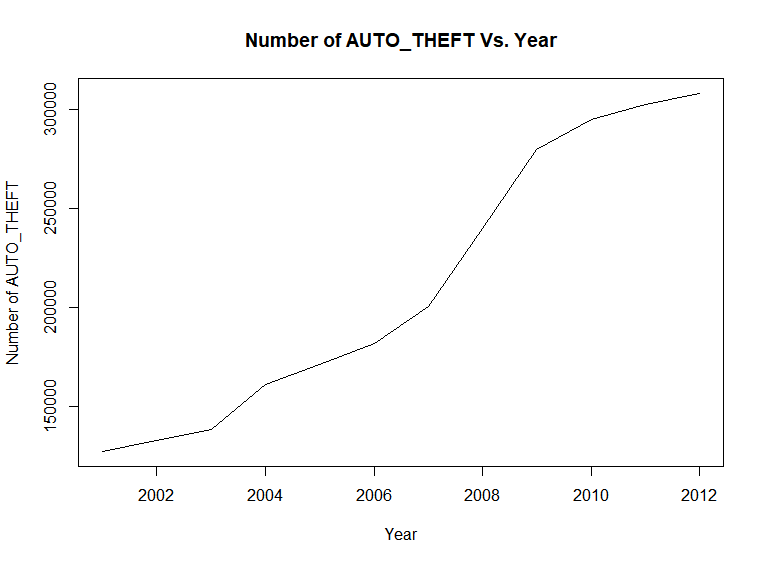
1. No trend can be observed but we can say that number of cases registered under burglary was lower in 2012 as compared to that in 2001.
2. Highest number of cases registered in 2001, around 2,02,500.
3. Lowest number of cases registered in 2010, around 9,400 cases.



**Fig 4.6.2.14** Number of theft cases vs. Year

**Analysis:**

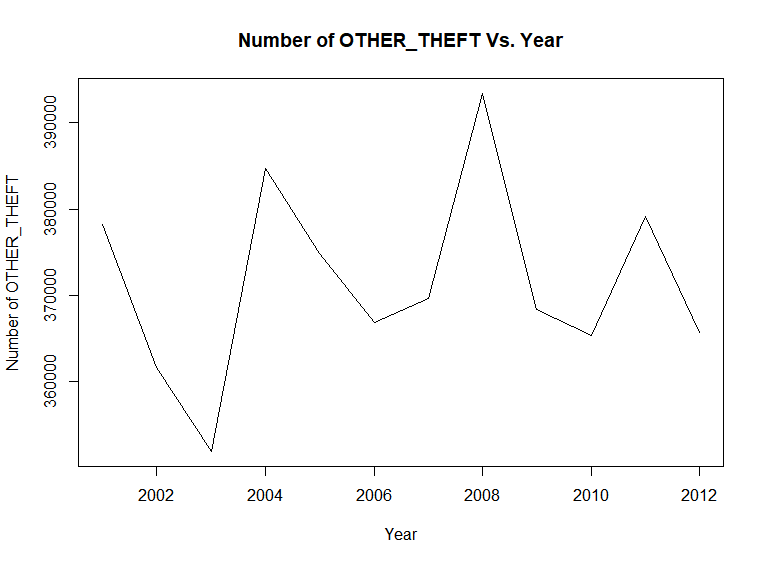
1. A Trend can be observed in the above plot, number cases registered under THEFT was increasing from 2001 to 2012 with a considerable cyclic pattern.
2. Lowest number of cases registered under theft in 2003, around 5,67,000 cases.
3. Highest number of cases registered under theft in 2011, around 7,00,000 cases.



**Fig 4.6.2.15** Number of auto\_theft cases vs. Year

**Analysis:**

1. A trend can be observed in the above plot, number of cases registered under Auto theft was monotonically increasing from 2001 to 2012.
2. Lowest number of cases registered under auto theft in 2001, around 3,00,000 cases.
3. Highest number of cases registered under auto theft in 2012, around 1,60,000 cases.



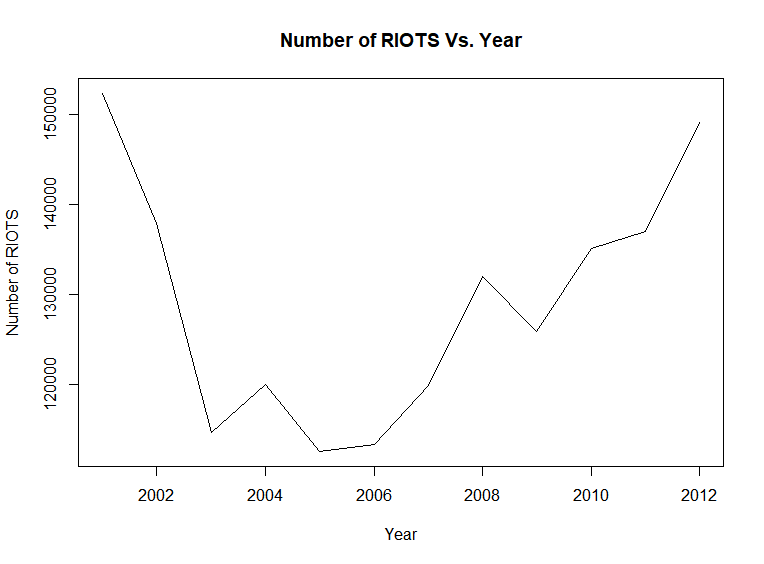
CYCLIC

PATTERN

**Fig 4.6.2.16** Number of murder cases vs. Year

**Analysis:**

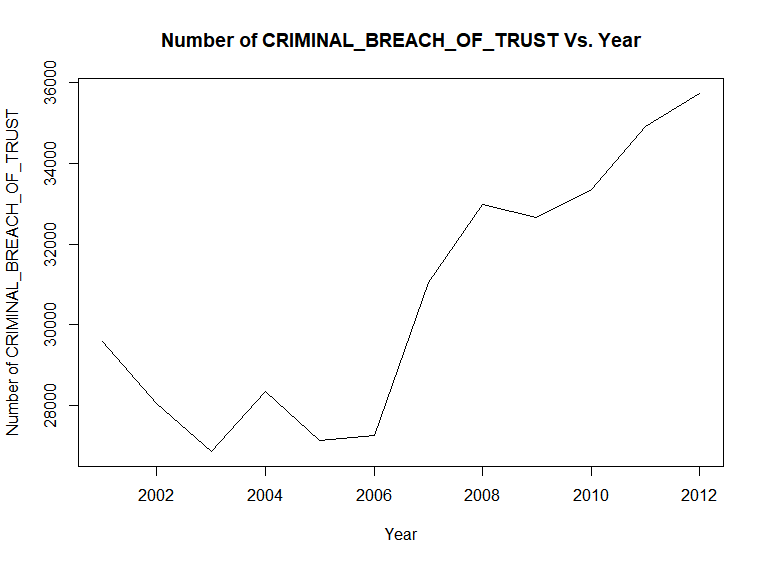
1. **Cyclic pattern** can be observed in the above plot.
2. Number of other\_theft cases varied in a cyclic pattern from year 2001 to 2012.
3. Lowest number of cases registered under other\_theft is observed in 2003 with around 3,62,000 registered cases.
4. Highest number of cases registered under other\_theft is observed in 2008 with around 3,94,000 registered cases.



**Fig 4.6.2.16** Number of riots cases vs. Year

**Analysis:**

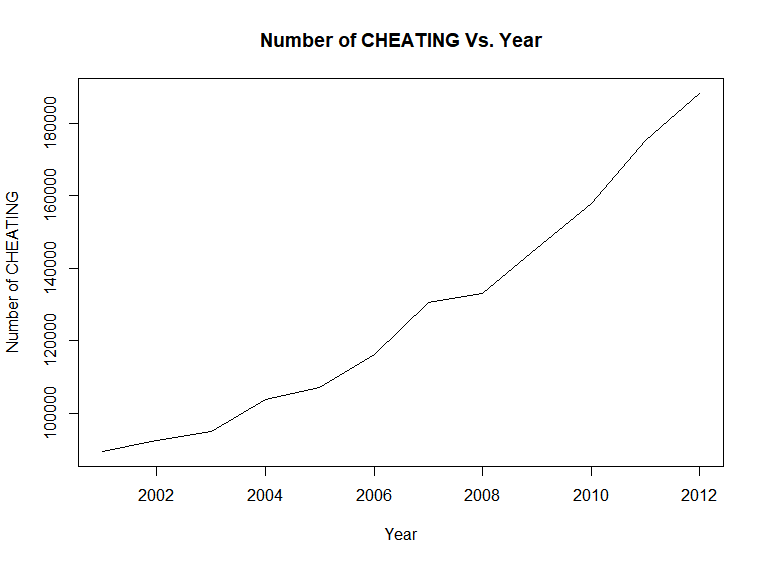
1. No Trend can be observed, but we can observe that from 2001 to 2005 number of cases registered under riots was decreasing and from 2006 to 2012 it was increasing.
2. Lowest number of cases registered under RIOTS is observed in 2005 with around 1,13,000 registered cases.
3. Highest number of cases registered under RIOTS is observed in 2001 with around 1,53,000 registered cases.



**Fig 4.6.2.17** Number of criminal breach of trust cases vs. Year

**Analysis:**

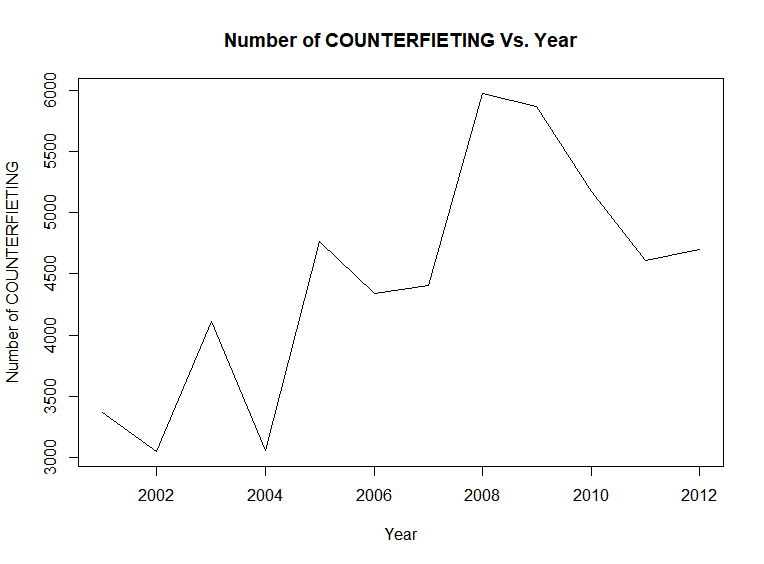
1. No Trend can be observed, but we can observe that from 2006 to 2012 number of cases registered under criminal breach of trust was increasing.
2. Lowest number of cases registered under criminal breach of trust is observed in 2003 with around 26,500 registered cases.
3. Highest number of cases registered under criminal breach of trust is observed in 2012 with around 35,800 registered cases.



**Fig 4.6.2.18** Number of cheating cases vs. Year

**Analysis:**

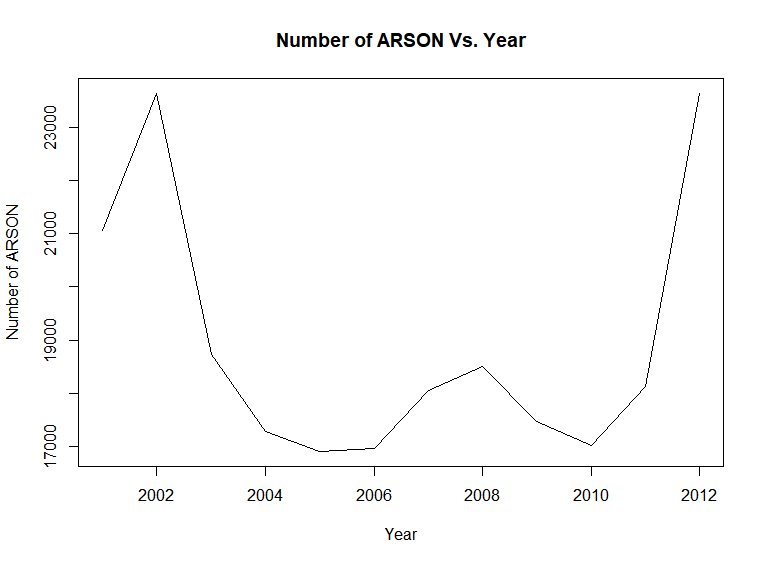
1. A trend can be observed from the above plot, number of cases registered under cheating was increasing from 2001 to 2012.
2. Lowest number of cheating cases registered in 2001 with around 99,400 cases.
3. Highest number of cheating cases registered in 2012 with around 1,85,000 cases.



**Fig 4.6.2.19** Number of counterfeiting cases vs. Year

**Analysis:**

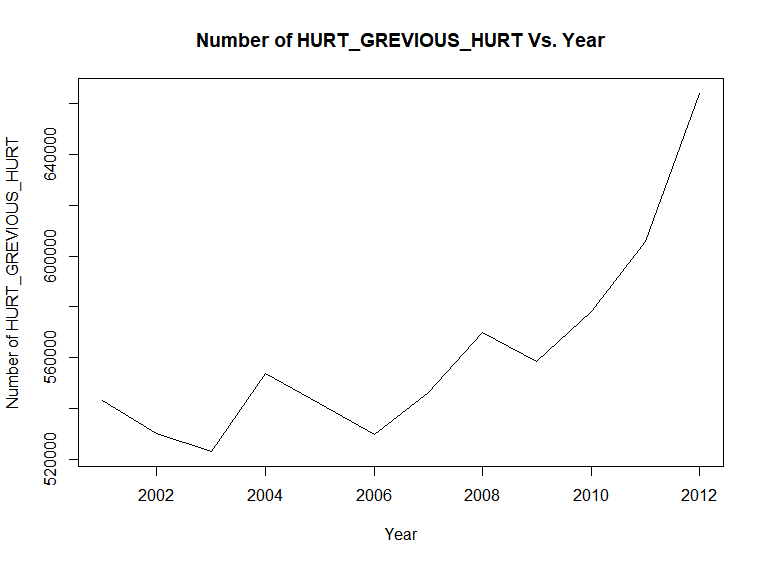
1. Trend with cyclic pattern can be observed plot, number of counterfieting cases increased with a cyclic variation from 2001 to 2012.
2. Lowest number of cases counterfieting registered in 2004 with around 3,100 cases.
3. Highest number of counterfieting cases registered in 2008 with around 5,900 cases.



**Fig 4.6.2.20** Number of arson cases vs. Year

**Analysis:**

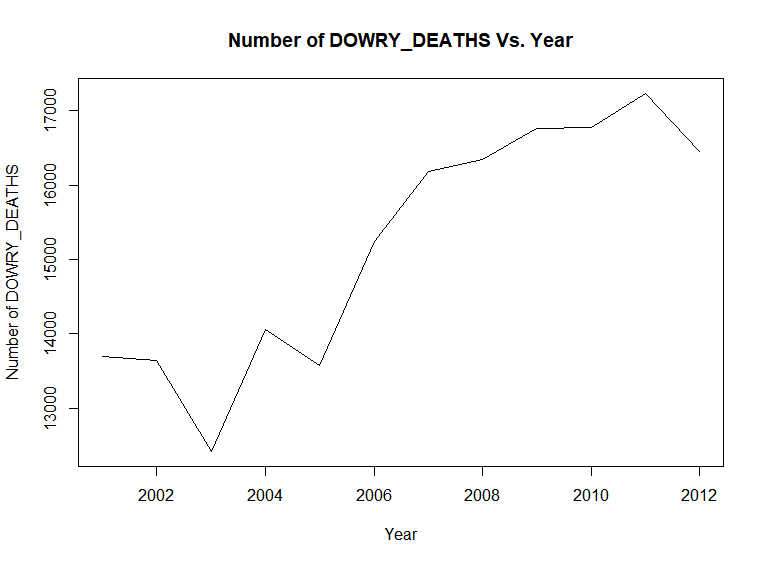
1. No trend can be observed.
2. Lowest number of ARSON cases registered in 2005 with around 17,000 cases.
3. Highest number of ARSON cases registered in 2002 with around 23,700 cases.
4. Also we can observe after achieving the highest count in the period 2001 to 2012 in 2001, the count decreased dramatically till 2005 and again reached a count closed to the highest count in 2012.



**Fig 4.6.2.21** Number of hurt grevious hurt cases vs. Year

**Analysis:**

1. A trend with cyclic variation can be observed in the above plot, from 2001 to 2009 number of cases was increasing with a cyclic variation and after that the number of cases monotonically increased.
2. Lowest number of cases registered in 2003 with around 5,20,000 cases.
3. Highest number of cases registered in 2012 with around 6,60,000 cases.

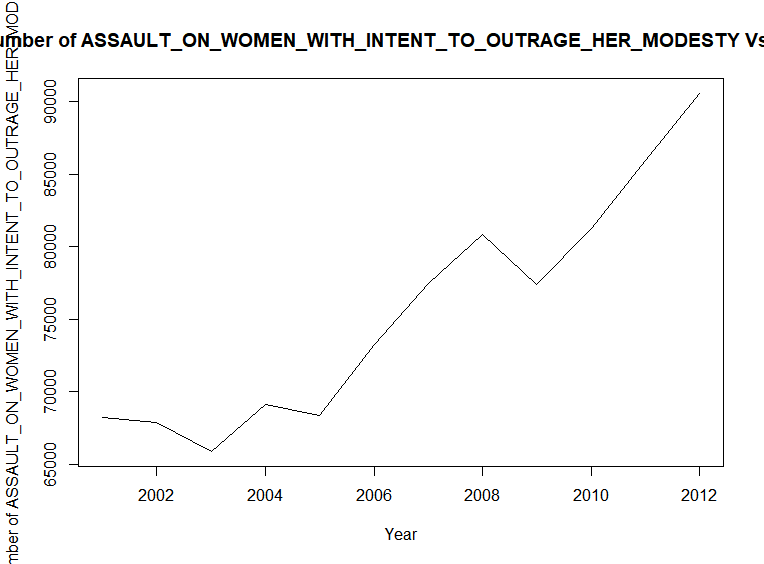


Random Movements

**Fig 4.6.2.22** Number of dowry deaths cases vs. Year

**Analysis:**

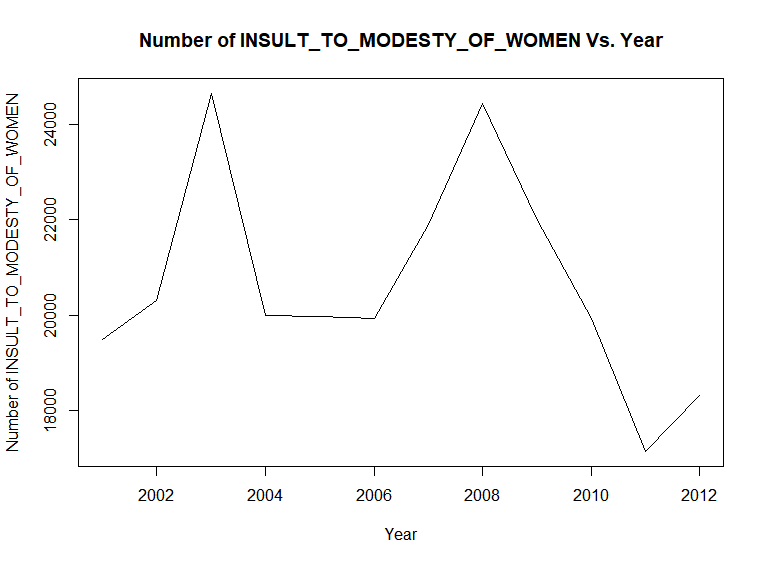
1. Trend can be observed from with y values increasing with few random movements.
2. Lowest number of cases registered in 2003 with around 12,250 cases.
3. Highest number of cases registered in 2011 with 17,200 around cases.



**Fig 4.6.2.23** Number of assault on women with intent to outrage her modesty cases vs. Year

**Analysis:**

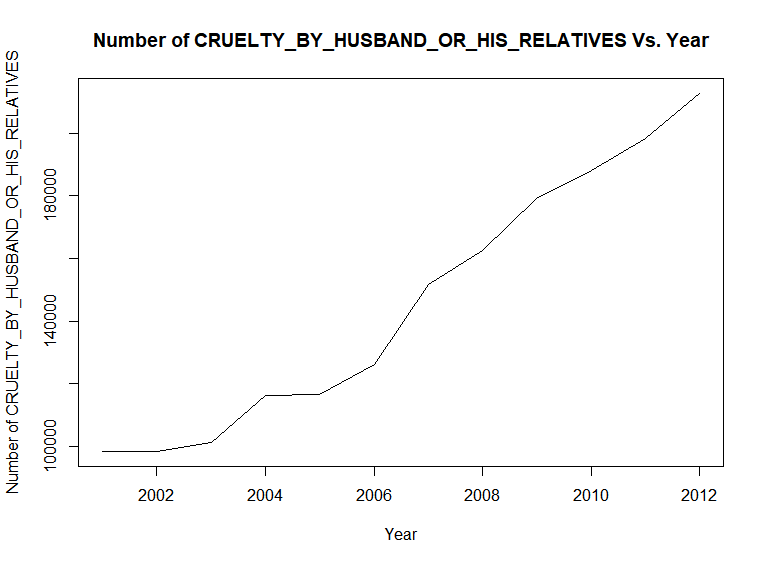
1. A trend can be observed in the above plot, the number of assault on women with intent to outrage her modesty cases is increasing from 2001 to 2012.
2. Lowest number of cases registered in 2003 with around 65,500 cases.
3. Highest number of cases registered in 2012 with 91,000 around cases.



**Fig 4.6.2.24** Number of insult to modesty of women cases vs. Year

**Analysis:**

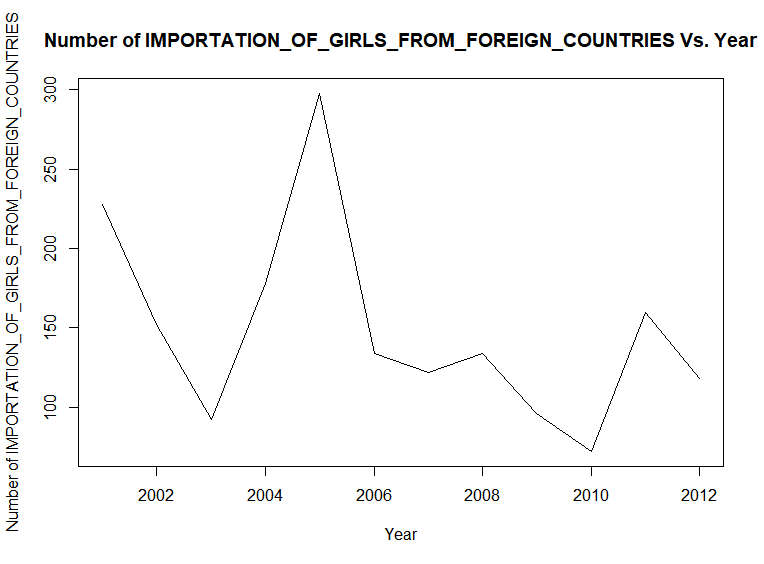
1. From the above plot, a trend with cyclic variation is estimated, number of cases of insult to modesty of women is decreasing with a cyclic variation from 2001 - 2012 and after 2012.
2. Lowest number of cases registered in 2011 with around 17,300 cases.
3. Highest number of cases registered in 2003 with 24,600 around cases.



**Fig 4.6.2.25** Number of cruelty by husband or his relatives cases vs. Year

**Analysis:**

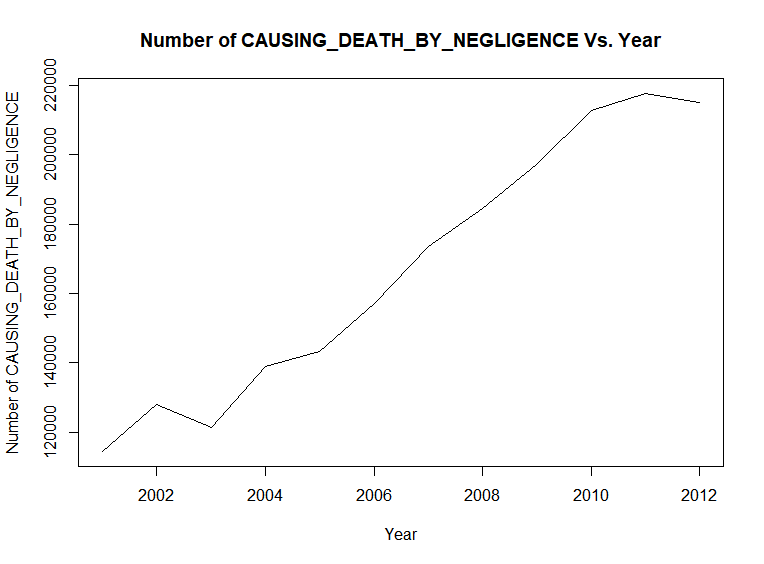
1. A trend can be observed in the above plot, the number of cases registered is increasing from 2001 to 2012.
2. Lowest number of cases registered in 2001 with around 99,800 cases.
3. Highest number of cases registered in 2012 with around 1,90,000 cases.



**Fig 4.6.2.26** Number of importation of girls from foreign countries cases vs. Year

**Analysis:**

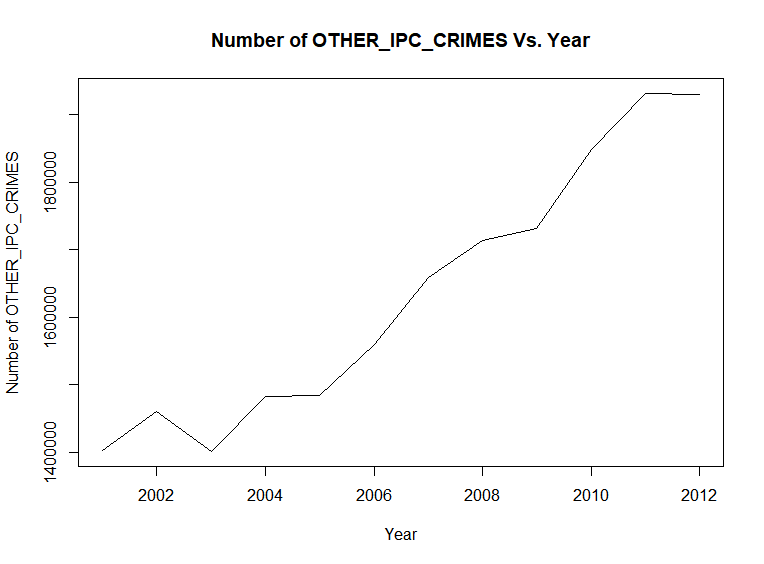
1. Trend with cyclic variation can be observed in the above plot, the number of importation of girls from foreign countries cases was decreasing with a cyclic variation from 2001 to 2012.
2. Lowest number of cases registered in 2005 with around 300 cases.
3. Highest number of cases registered in 2011 with around 70 cases.



**Fig 4.6.2.27** Number of causing death by negligence cases vs. Year

**Analysis:**

1. A trend can be observed in the above plot, number of cases registered under causing death by negligence was increasing from 2001 to 2012.
2. Lowest number of registered cases in 2001 with around 1,27,000 cases.
3. Highest number of cases registered in 2011 with around 2,17,000 cases.



**Fig 4.6.2.28** Number of other ipc crime cases vs. Year

**Analysis:**

1. A trend can be observed in the above plot, the number of cases under other ipc crimes increased from 2001 to 2012.
2. Lowest number of cases registered in 2003 with around 1,40,000 cases.
3. Highest number of cases registered in 2011 with around 1,90,000 cases.
4. **Scope and Limitations**

**Scope :-**

**Deliverables:**

* Our project will deliver regression models which will predict the total number of specific crimes and total number of IPC crimes provided state, district and year as input or any combination of these (according depiction changes).
* Our project will set of crimes committed in high numbers and most frequently within the time period 2001-2012 through eclat and rules representing this through apriori algorithm.
* Our project will deliver the clusters of districts/area with the similarities in the number of different crimes committed there yearly or in the whole period of 2001-2012.
* Time Series analysis on the data set.
* Data visualizations : year vs. Specific crime count (National Level/State level), bar plots on specific crime counts (National Level/State level) to draw useful inferences.

**Limitations:**

* We have less amount data if we see the data set as a historical data as it contains year wise records distinct on the basis of state and district pair value, also it would have better if we had a data set with case wise historical data, due to this we end up building weak regression models.
* After hot encoding the categorical variables before fitting the regression algorithms, data frame object size increases with 35 +808 (state +district) + other numerical variables, which results in high execution time of algorithms.
* To create transaction data for association rule mining we mapped numerical values to binary categorical variable by comparing the numerical values of each record to the median of all values of that variable, this is not a great approach to produce good association rules.
* If any new district/Area or state value is given as input to the regression model then it would not be able to predict.