# PROJECT: ANALYZING INDIAN CRIME DATA

## GUIDED BY:

## Dr. Nilamadhab Mishra

# ABSTRACT

Analyzing the Indian crime data set to produce some useful inferences: 01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv (available at data.gov.in ).

In our project we have used the above mentioned data set and other transformed and processed forms of the data set to produce useful data visualization and also have applied different machine learning algorithms namely, Linear regression, Multiple linear regression, Polynomial regression, support vector regression, decision tree regression and random forest regression to build some useful regression models. We have also implemented clustering algorithms k means and hierarchical clustering with the given data set and data sets derived from it. We have also performed time series analysis and association rule mining algorithms, namely, eclat and apriori algorithms on the transformed form of the data set.

We have implemented all the data visualization, machine learning algorithms, time series analysis and association rule mining in R and for deriving the needed data set and data set transformations we have used SQL.

# Problem Statement

## Problem:

With the chosen data set we have to build a predictive model which can predict the total number of specific crimes and total number of total IPC crimes for a given year, state, district/area or any combination of these three provided as input. Other than this, time series analysis and possible set of crimes committed frequently in India on the basis of given data. And to group different districts(treated as instances with year).

**Our project as a solution:**

We have used different regression, association rule mining and clustering algorithms on the chosen data set to produce solutions to the given problem.

# The Objective Of The Project

**General objective :-**

To build a regression model 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). Also to find out set of crimes committed in high numbers and most frequently.

**Specific objectives :-**

1. To build a regression model 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).
2. Find out set of crimes committed in high numbers and most frequently.
3. Find 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.
4. Time Series analysis on the data set.
5. 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.

# Approach/Methodology:

## Tools Used :

Rstudio, Notepad and MySql DBMS

R packages used:

caTools, e1071, rpart, randomForest, arules, ggplot2, dummies and cluster.

**Approach/Methodology:**

1. Needed data pre-processing (according to the algorithms to be applied or according to the specific objective), data transformation/deriving new data sets from the chosen data set using R and SQL.
2. Data visualization: bar plots: Specific crime count/Total crime count (State level/ National Level) / and line plots : Specific crime/Total Crimes vs. Year (State level/National Level).
3. Fitting our data sets to different algorithms according to our objectives, visualizing the results, measuring and comparing the models.

# Literature Review

Crime in India: Understanding human Behavior through data

There is so much we can understand from the data, Some prediction models and their uses are shown below:

1. Crime Analysis and Prediction System (CAPS) : By Suryansh Singh Raghuvanshi (141349) Deepak Kumar (141370)

This system can predict regions which have high probability for crime occurrence and can visualize crime prone areas. Using the concept of Machine Learning in R we can extract previously unknown, useful information from an unstructured data.

1. PREDICTION of CRIME RATE ANALYSIS using MACHINE LEARNING APPROACH :

The aim is to investigate machine learning based techniques for crime rate by prediction results in best accuracy and the analysis of dataset by supervised machine learning technique(SMLT) to capture several information’s like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset.

1. USING MACHINE LEARNING ALGORITHMS TO ANALYZE CRIME DATA : By Lawrence McClendon and Natarajan Meghanathan

Model has implementation of Linear Regression, Additive Regression, and Decision Stump algorithms, on the Crime Dataset. The scope of this project is to prove how effective and accurate the machine learning algorithms used in data mining analysis can be at predicting violent crime patterns.

1. Spatio-Temporal Crime HotSpot Detection and Prediction :

This study unfolds the following major aspects:

1) The impact of data mining and machine learning approaches

2) The utility of time series analysis techniques and deep learning techniques in crime trend prediction.

3) The inclusion of spatial and temporal information in crime datasets making the crime prediction systems more accurate and reliable.

# Implementation Details

**About data set:**

The data set contains 9017 records and 33 columns (variables), each record 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 are 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.

**Note : We have zipped the data set with the report.**

## Taking care of missing values using R:

We replaced the na values with the mean of the all column values of the corresponding column:

# Taking care of missing data

dataset$MURDER = ifelse(is.na(dataset$MURDER),

ave(dataset$MURDER, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$MURDER)

dataset$ATTEMPT.TO.MURDER = ifelse(is.na(dataset$ATTEMPT.TO.MURDER),

ave(dataset$ATTEMPT.TO.MURDER, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$ATTEMPT.TO.MURDER)

dataset$CULPABLE.HOMICIDE.NOT.AMOUNTING.TO.MURDER = ifelse(is.na(dataset$CULPABLE.HOMICIDE.NOT.AMOUNTING.TO.MURDER),

ave(dataset$CULPABLE.HOMICIDE.NOT.AMOUNTING.TO.MURDER, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$CULPABLE.HOMICIDE.NOT.AMOUNTING.TO.MURDER)

dataset$RAPE = ifelse(is.na(dataset$RAPE),

ave(dataset$RAPE, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$RAPE)

dataset$CUSTODIAL.RAPE = ifelse(is.na(dataset$CUSTODIAL.RAPE),

ave(dataset$CUSTODIAL.RAPE, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$CUSTODIAL.RAPE)

dataset$OTHER.RAPE = ifelse(is.na(dataset$OTHER.RAPE),

ave(dataset$OTHER.RAPE, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$OTHER.RAPE)

dataset$KIDNAPPING...ABDUCTION = ifelse(is.na(dataset$KIDNAPPING...ABDUCTION),

ave(dataset$KIDNAPPING...ABDUCTION, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$KIDNAPPING...ABDUCTION)

dataset$KIDNAPPING.AND.ABDUCTION.OF.WOMEN.AND.GIRLS = ifelse(is.na(dataset$KIDNAPPING.AND.ABDUCTION.OF.WOMEN.AND.GIRLS),

ave(dataset$KIDNAPPING.AND.ABDUCTION.OF.WOMEN.AND.GIRLS, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$KIDNAPPING.AND.ABDUCTION.OF.WOMEN.AND.GIRLS)

dataset$KIDNAPPING.AND.ABDUCTION.OF.OTHERS = ifelse(is.na(dataset$KIDNAPPING.AND.ABDUCTION.OF.OTHERS),

ave(dataset$KIDNAPPING.AND.ABDUCTION.OF.OTHERS, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$KIDNAPPING.AND.ABDUCTION.OF.OTHERS)

dataset$DACOITY = ifelse(is.na(dataset$DACOITY),

ave(dataset$DACOITY, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$DACOITY)

dataset$PREPARATION.AND.ASSEMBLY.FOR.DACOITY = ifelse(is.na(dataset$PREPARATION.AND.ASSEMBLY.FOR.DACOITY),

ave(dataset$PREPARATION.AND.ASSEMBLY.FOR.DACOITY, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$PREPARATION.AND.ASSEMBLY.FOR.DACOITY)

dataset$ROBBERY = ifelse(is.na(dataset$ROBBERY),

ave(dataset$ROBBERY, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$ROBBERY)

dataset$BURGLARY = ifelse(is.na(dataset$BURGLARY),

ave(dataset$BURGLARY, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$BURGLARY)

dataset$THEFT = ifelse(is.na(dataset$THEFT),

ave(dataset$THEFT, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$THEFT)

dataset$AUTO.THEFT = ifelse(is.na(dataset$AUTO.THEFT),

ave(dataset$AUTO.THEFT, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$AUTO.THEFT)

dataset$OTHER.THEFT = ifelse(is.na(dataset$OTHER.THEFT),

ave(dataset$OTHER.THEFT, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$OTHER.THEFT)

dataset$RIOTS = ifelse(is.na(dataset$RIOTS),

ave(dataset$RIOTS, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$RIOTS)

dataset$CRIMINAL.BREACH.OF.TRUST = ifelse(is.na(dataset$CRIMINAL.BREACH.OF.TRUST),

ave(dataset$CRIMINAL.BREACH.OF.TRUST, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$CRIMINAL.BREACH.OF.TRUST)

dataset$CHEATING = ifelse(is.na(dataset$CHEATING),

ave(dataset$CHEATING, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$CHEATING)

dataset$COUNTERFIETING = ifelse(is.na(dataset$COUNTERFIETING),

ave(dataset$COUNTERFIETING, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$COUNTERFIETING)

dataset$ARSON = ifelse(is.na(dataset$ARSON),

ave(dataset$ARSON, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$ARSON)

dataset$HURT.GREVIOUS.HURT = ifelse(is.na(dataset$HURT.GREVIOUS.HURT),

ave(dataset$HURT.GREVIOUS.HURT, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$HURT.GREVIOUS.HURT)

dataset$DOWRY.DEATHS = ifelse(is.na(dataset$DOWRY.DEATHS),

ave(dataset$DOWRY.DEATHS, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$DOWRY.DEATHS)

dataset$ASSAULT.ON.WOMEN.WITH.INTENT.TO.OUTRAGE.HER.MODESTY = ifelse(is.na(dataset$ASSAULT.ON.WOMEN.WITH.INTENT.TO.OUTRAGE.HER.MODESTY),

ave(dataset$ASSAULT.ON.WOMEN.WITH.INTENT.TO.OUTRAGE.HER.MODESTY, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$ASSAULT.ON.WOMEN.WITH.INTENT.TO.OUTRAGE.HER.MODESTY)

dataset$INSULT.TO.MODESTY.OF.WOMEN = ifelse(is.na(dataset$INSULT.TO.MODESTY.OF.WOMEN),

ave(dataset$INSULT.TO.MODESTY.OF.WOMEN, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$INSULT.TO.MODESTY.OF.WOMEN)

dataset$CRUELTY.BY.HUSBAND.OR.HIS.RELATIVES = ifelse(is.na(dataset$CRUELTY.BY.HUSBAND.OR.HIS.RELATIVES),ave(dataset$CRUELTY.BY.HUSBAND.OR.HIS.RELATIVES, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$CRUELTY.BY.HUSBAND.OR.HIS.RELATIVES)

dataset$IMPORTATION.OF.GIRLS.FROM.FOREIGN.COUNTRIES = ifelse(is.na(dataset$IMPORTATION.OF.GIRLS.FROM.FOREIGN.COUNTRIES),ave(dataset$IMPORTATION.OF.GIRLS.FROM.FOREIGN.COUNTRIES, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$IMPORTATION.OF.GIRLS.FROM.FOREIGN.COUNTRIES)

dataset$CAUSING.DEATH.BY.NEGLIGENCE = ifelse(is.na(dataset$CAUSING.DEATH.BY.NEGLIGENCE),ave(dataset$CAUSING.DEATH.BY.NEGLIGENCE, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$CAUSING.DEATH.BY.NEGLIGENCE)

dataset$OTHER.IPC.CRIMES = ifelse(is.na(dataset$OTHER.IPC.CRIMES),ave(dataset$OTHER.IPC.CRIMES, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$OTHER.IPC.CRIMES)

dataset$TOTAL.IPC.CRIMES = ifelse(is.na(dataset$TOTAL.IPC.CRIMES),ave(dataset$TOTAL.IPC.CRIMES, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$TOTAL.IPC.CRIMES)

dataset$STATE.UT = ifelse(is.na(dataset$STATE.UT),ave(dataset$STATE.UT, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$STATE.UT)

dataset$DISTRICT = ifelse(is.na(dataset$DISTRICT),ave(dataset$DISTRICT, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$DISTRICT)

dataset$YEAR = ifelse(is.na(dataset$YEAR),ave(dataset$YEAR, FUN = function(x) mean(x, na.rm = TRUE)),

dataset$YEAR)

## Dummy coding the categorical variables: STATE/UT and DISTRICT before applying regression algorithms on the data set:

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):**

library(dummies)

dataset = read.csv("01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv")

dataset <- dummy.data.frame(dataset, sep= ".")

write.csv(row.names = FALSE, dataset, file = "dataa.csv")

We stored the dummy encoded data into a new csv file named dataa.csv (attached with the report) so that we can use it further in other codes directly.

## Implementing association rule mining on the chosen data set.

1. **Producing transaction data set from the chosen data set:**

Before applying any association rule mining 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:**

use mydb;

SET SQL\_SAFE\_UPDATES = 0;

update data set MURDER = 0 where MURDER = 1;

update data set MURDER = 0 where MURDER<38;

update data set MURDER = 1 where MURDER>38 or MURDER= 38;

update data set ATTEMPT\_TO\_MURDER = 0 where ATTEMPT\_TO\_MURDER = 1;

update data set ATTEMPT\_TO\_MURDER = 0 where ATTEMPT\_TO\_MURDER < 28;

update data set ATTEMPT\_TO\_MURDER = 1 where ATTEMPT\_TO\_MURDER > 28 or ATTEMPT\_TO\_MURDER = 28;

update data set CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = 0 where CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = 1;

update data set CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = 0 where CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER < 2;

update data set CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = 1 where CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER > 2 or CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = 2;

update data set CUSTODIAL\_RAPE = 1 where CUSTODIAL\_RAPE > 0;

update data set RAPE = 0 where RAPE = 1;

update data set RAPE = 0 where RAPE <20;

update data set RAPE = 1 where RAPE > 20 or RAPE = 20;

update data set OTHER\_RAPE = 0 where OTHER\_RAPE = 1;

update data set OTHER\_RAPE = 0 where OTHER\_RAPE<20;

update data set OTHER\_RAPE= 1 where OTHER\_RAPE>20 and OTHER\_RAPE= 20 ;

update data set PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY = 1 where PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY>0;

update data set KIDNAPPING\_\_\_ABDUCTION = 0 where KIDNAPPING\_\_\_ABDUCTION = 1;

update data set KIDNAPPING\_\_\_ABDUCTION = 0 where KIDNAPPING\_\_\_ABDUCTION <25;

update data set KIDNAPPING\_\_\_ABDUCTION = 1 where KIDNAPPING\_\_\_ABDUCTION >25 or KIDNAPPING\_\_\_ABDUCTION = 25;

update data set KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS = 0 where KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS= 1;

update data set KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS = 0 where KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS<18;

update data set KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS = 1 where KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS > 18 or KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS = 18;

update data set KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS = 0 where KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS = 1;

update data set KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS = 0 where KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS<5;

update data set KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS= 1 where KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS>5 and KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS = 5;

update data set DACOITY = 0 where DACOITY = 1;

update data set DACOITY = 0 where DACOITY<3;

update data set DACOITY= 1 where DACOITY>3 and DACOITY = 3;

update data set ROBBERY = 0 where ROBBERY = 1;

update data set ROBBERY = 0 where ROBBERY< 17;

update data set ROBBERY = 1 where ROBBERY>17 or ROBBERY = 17;

update data set BURGLARY = 0 where BURGLARY = 1;

update data set BURGLARY = 0 where BURGLARY < 83;

update data set BURGLARY = 1 where BURGLARY > 83 or BURGLARY = 83;

update data set THEFT = 0 where THEFT = 1;

update data set THEFT = 0 where THEFT<217;

update data set THEFT = 1 where THEFT >217 or THEFT = 217 ;

update data set AUTO\_THEFT = 0 where AUTO\_THEFT = 1;

update data set AUTO\_THEFT= 0 where AUTO\_THEFT<48;

update data set AUTO\_THEFT = 1 where AUTO\_THEFT>48 or AUTO\_THEFT= 48 ;

update data set OTHER\_THEFT= 0 where OTHER\_THEFT= 1;

update data set OTHER\_THEFT= 0 where OTHER\_THEFT<152;

update data set OTHER\_THEFT= 1 where OTHER\_THEFT>152 or OTHER\_THEFT= 152 ;

update data set RIOTS= 0 where RIOTS= 1;

update data set RIOTS= 0 where RIOTS<46;

update data set RIOTS= 1 where RIOTS>46 or RIOTS= 46 ;

update data set CRIMINAL\_BREACH\_OF\_TRUST= 0 where CRIMINAL\_BREACH\_OF\_TRUST= 1;

update data set CRIMINAL\_BREACH\_OF\_TRUST= 0 where CRIMINAL\_BREACH\_OF\_TRUST<11;

update data set CRIMINAL\_BREACH\_OF\_TRUST= 1 where CRIMINAL\_BREACH\_OF\_TRUST>11 or CRIMINAL\_BREACH\_OF\_TRUST= 11 ;

update data set CHEATING= 0 where CHEATING= 37;

update data set CHEATING= 0 where CHEATING<37;

update data set CHEATING= 1 where CHEATING>37 or CHEATING = 37;

update data set COUNTERFIETING = 0 where COUNTERFIETING = 1;

update data set COUNTERFIETING = 0 where COUNTERFIETING <1;

update data set COUNTERFIETING= 1 where COUNTERFIETING>1 or COUNTERFIETING= 1;

update data set ARSON= 0 where ARSON= 1;

update data set ARSON= 0 where ARSON<8;

update data set ARSON = 1 where ARSON >8 or ARSON=8 ;

update data set HURT\_GREVIOUS\_HURT= 0 where HURT\_GREVIOUS\_HURT= 1;

update data set HURT\_GREVIOUS\_HURT= 0 where HURT\_GREVIOUS\_HURT<199;

update data set HURT\_GREVIOUS\_HURT= 1 where HURT\_GREVIOUS\_HURT>199 or HURT\_GREVIOUS\_HURT= 199;

update data set DOWRY\_DEATHS = 0 where DOWRY\_DEATHS= 1;

update data set DOWRY\_DEATHS= 0 where DOWRY\_DEATHS<5;

update data set DOWRY\_DEATHS= 1 where DOWRY\_DEATHS >5 or DOWRY\_DEATHS = 5 ;

update data set ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY = 1 where ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY>0;

update data set INSULT\_TO\_MODESTY\_OF\_WOMEN = 0 where INSULT\_TO\_MODESTY\_OF\_WOMEN = 1;

update data set INSULT\_TO\_MODESTY\_OF\_WOMEN = 0 where INSULT\_TO\_MODESTY\_OF\_WOMEN<2;

update data set INSULT\_TO\_MODESTY\_OF\_WOMEN = 1 where INSULT\_TO\_MODESTY\_OF\_WOMEN> 2 or INSULT\_TO\_MODESTY\_OF\_WOMEN= 2 ;

update data set CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = 0 where CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = 1;

update data set CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = 0 where CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES <47;

update data set CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = 1 where CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES >47 or CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = 47 ;

update data set IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES = 1 where IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES > 0;

update data set CAUSING\_DEATH\_BY\_NEGLIGENCE = 0 where CAUSING\_DEATH\_BY\_NEGLIGENCE = 1;

update data set CAUSING\_DEATH\_BY\_NEGLIGENCE = 0 where CAUSING\_DEATH\_BY\_NEGLIGENCE <68;

update data set CAUSING\_DEATH\_BY\_NEGLIGENCE = 1 where CAUSING\_DEATH\_BY\_NEGLIGENCE > 68 or CAUSING\_DEATH\_BY\_NEGLIGENCE = 68 ;

select \* from data LIMIT 0,9012;

SET SQL\_SAFE\_UPDATES = 1;

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:**

use mydb;

SET SQL\_SAFE\_UPDATES = 0;

update ssparsecrimedata set MURDER = "MURDER" where MURDER = "1";

update ssparsecrimedata set MURDER = "" where MURDER = "0";

update ssparsecrimedata set ATTEMPT\_TO\_MURDER = "ATTEMPT\_TO\_MURDER" where ATTEMPT\_TO\_MURDER = "1";

update ssparsecrimedata set ATTEMPT\_TO\_MURDER = "" where ATTEMPT\_TO\_MURDER = "0";

update ssparsecrimedata set CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = "CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER" where CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = "1";

update ssparsecrimedata set CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = "" where CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER = "0";

update ssparsecrimedata set RAPE = "RAPE" where RAPE = "1";

update ssparsecrimedata set RAPE = "" where RAPE = "0";

update ssparsecrimedata set CUSTODIAL\_RAPE = "CUSTODIAL\_RAPE" where CUSTODIAL\_RAPE = "1";

update ssparsecrimedata set CUSTODIAL\_RAPE = "" where CUSTODIAL\_RAPE = "0";

update ssparsecrimedata set OTHER\_RAPE = "OTHER\_RAPE" where OTHER\_RAPE = "1";

update ssparsecrimedata set OTHER\_RAPE = "" where OTHER\_RAPE = "0";

update ssparsecrimedata set KIDNAPPING\_\_\_ABDUCTION = "KIDNAPPING\_\_\_ABDUCTION" where KIDNAPPING\_\_\_ABDUCTION = "1";

update ssparsecrimedata set KIDNAPPING\_\_\_ABDUCTION = "" where KIDNAPPING\_\_\_ABDUCTION = "0";

update ssparsecrimedata set KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS = "KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS" where KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS = "1";

update ssparsecrimedata set KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS = "" where KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS = "0";

update ssparsecrimedata set KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS = "KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS" where KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS = "1";

update ssparsecrimedata set KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS = "" where KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS = "0";

update ssparsecrimedata set DACOITY = "DACOITY" where DACOITY = "1";

update ssparsecrimedata set DACOITY = "" where DACOITY = "0";

update ssparsecrimedata set PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY = "PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY" where PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY = "1";

update ssparsecrimedata set PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY = "" where PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY = "0";

update ssparsecrimedata set ROBBERY = "ROBBERY" where ROBBERY = "1";

update ssparsecrimedata set ROBBERY = "" where ROBBERY = "0";

update ssparsecrimedata set BURGLARY = "BURGLARY" where BURGLARY = "1";

update ssparsecrimedata set BURGLARY = "" where BURGLARY = "0";

update ssparsecrimedata set THEFT = "THEFT" where THEFT = "1";

update ssparsecrimedata set THEFT = "" where THEFT = "0";

update ssparsecrimedata set AUTO\_THEFT = "AUTO\_THEFT" where AUTO\_THEFT = "1";

update ssparsecrimedata set AUTO\_THEFT = "" where AUTO\_THEFT = "0";

update ssparsecrimedata set OTHER\_THEFT = "OTHER\_THEFT" where OTHER\_THEFT = "1";

update ssparsecrimedata set OTHER\_THEFT = "" where OTHER\_THEFT = "0";

update ssparsecrimedata set RIOTS = "RIOTS" where RIOTS = "1";

update ssparsecrimedata set RIOTS = "" where RIOTS = "0";

update ssparsecrimedata set CRIMINAL\_BREACH\_OF\_TRUST = "CRIMINAL\_BREACH\_OF\_TRUST" where CRIMINAL\_BREACH\_OF\_TRUST = "1";

update ssparsecrimedata set CRIMINAL\_BREACH\_OF\_TRUST = "" where CRIMINAL\_BREACH\_OF\_TRUST = "0";

update ssparsecrimedata set CHEATING = "CHEATING" where CHEATING = "1";

update ssparsecrimedata set CHEATING = "" where CHEATING = "0";

update ssparsecrimedata set CRIMINAL\_BREACH\_OF\_TRUST = "CRIMINAL\_BREACH\_OF\_TRUST" where CRIMINAL\_BREACH\_OF\_TRUST = "1";

update ssparsecrimedata set CRIMINAL\_BREACH\_OF\_TRUST = "" where CRIMINAL\_BREACH\_OF\_TRUST = "0";

update ssparsecrimedata set COUNTERFIETING = "COUNTERFIETING" where COUNTERFIETING = "1";

update ssparsecrimedata set COUNTERFIETING = "" where COUNTERFIETING = "0";

update ssparsecrimedata set ARSON = "ARSON" where ARSON = "1";

update ssparsecrimedata set ARSON = "" where ARSON = "0";

update ssparsecrimedata set HURT\_GREVIOUS\_HURT = "HURT\_GREVIOUS\_HURT" where HURT\_GREVIOUS\_HURT = "1";

update ssparsecrimedata set HURT\_GREVIOUS\_HURT = "" where HURT\_GREVIOUS\_HURT = "0";

update ssparsecrimedata set ARSON = "ARSON" where ARSON = "1";

update ssparsecrimedata set ARSON = "" where ARSON = "0";

update ssparsecrimedata set DOWRY\_DEATHS = "DOWRY\_DEATHS" where DOWRY\_DEATHS = "1";

update ssparsecrimedata set DOWRY\_DEATHS = "" where DOWRY\_DEATHS = "0";

update ssparsecrimedata set ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY = "ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY" where ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY = "1";

update ssparsecrimedata set ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY = "" where ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY = "0";

update ssparsecrimedata set CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = "CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES" where CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = "1";

update ssparsecrimedata set CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = "" where CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES = "0";

update ssparsecrimedata set IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES = "IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES" where IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES = "1";

update ssparsecrimedata set IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES = "" where IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES = "0";

update ssparsecrimedata set CAUSING\_DEATH\_BY\_NEGLIGENCE = "CAUSING\_DEATH\_BY\_NEGLIGENCE" where CAUSING\_DEATH\_BY\_NEGLIGENCE = "1";

update ssparsecrimedata set CAUSING\_DEATH\_BY\_NEGLIGENCE = "" where CAUSING\_DEATH\_BY\_NEGLIGENCE = "0";

update ssparsecrimedata set INSULT\_TO\_MODESTY\_OF\_WOMEN = "INSULT\_TO\_MODESTY\_OF\_WOMEN" where INSULT\_TO\_MODESTY\_OF\_WOMEN = "1";

update ssparsecrimedata set INSULT\_TO\_MODESTY\_OF\_WOMEN = "" where INSULT\_TO\_MODESTY\_OF\_WOMEN = "0";

select \* from ssparsecrimedata LIMIT 0,9012;

Then we stored the data into a csv file named SparseDataToTransaction.csv (attached with the report).

1. **Implementing apriori algorithm on the data set (**SparseDataToTransaction.csv**)(apriori.R):**

# Apriori

# install.packages('arules')

library(arules)

dataset = read.csv('SparseDataToTransaction.csv', header = FALSE)

dataset = read.transactions('SparseDataToTransaction.csv', sep = ',', rm.duplicates = TRUE)

summary(dataset)

itemFrequencyPlot(dataset, topN = 100)

# Training Apriori on the dataset

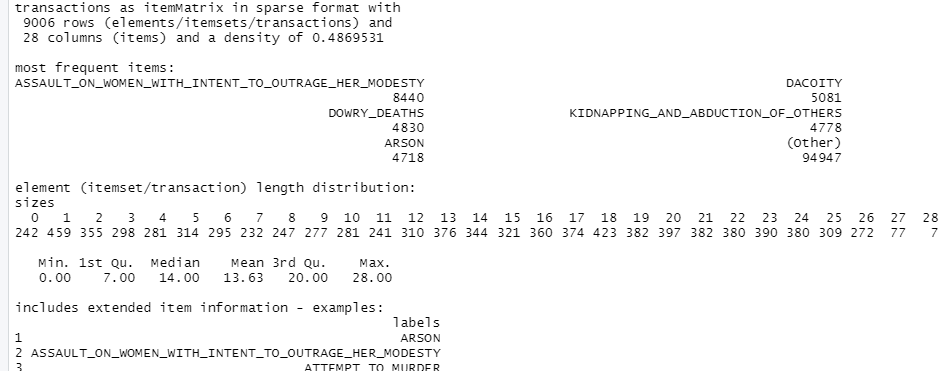
rules = apriori(data = dataset, parameter = list(support = 0.3, confidence = 0.8))

# Visualising the results

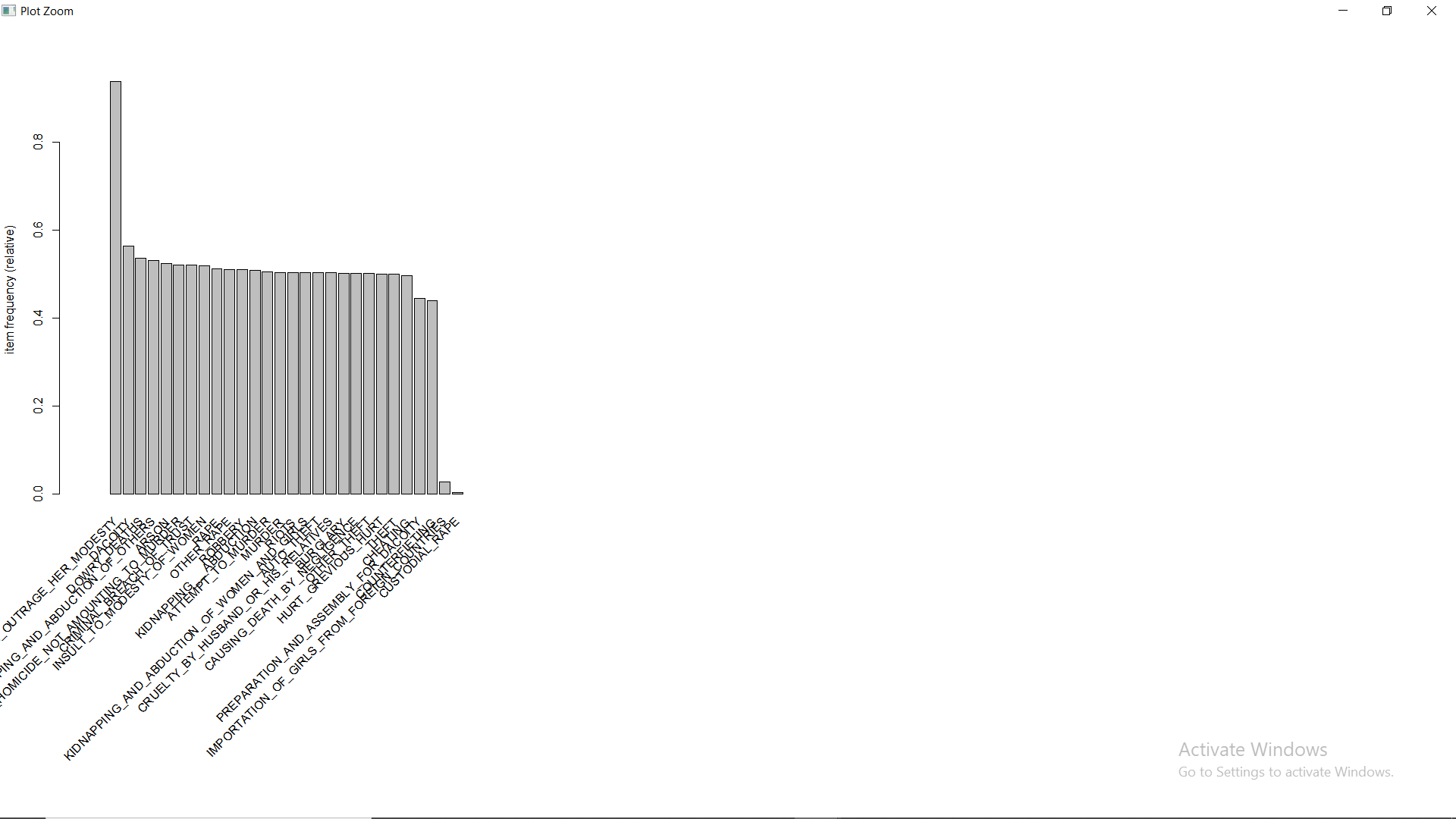
inspect(sort(rules, by = 'lift')[1:10])

**Output:**

**> summary(dataset)**



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

****

**Analysis:**

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))**

Apriori

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 e.g:

[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 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.

1. **Implementing eclat algorithm on the transaction data set (**SparseDataToTransaction.csv)**:**

Implementation in R (eclat.R):

# Eclat

# Data Preprocessing

# install.packages('arules')

library(arules)

dataset = read.csv('SparseDataToTransaction.csv')

dataset = read.transactions('SparseDataToTransaction.csv', sep = ',', rm.duplicates = TRUE)

summary(dataset)

itemFrequencyPlot(dataset, topN = 100)

# Training Eclat on the dataset

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

# Visualising the results

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

**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.

1. **Other derived data sets:**

**SQL for Data grouped by Year with sum of all crimes, grouped by State with sum of all crimes, grouped by state and district and with sum of all crimes, grouped by state and year:**

use mydb;

select STATE\_UT, SUM(MURDER) AS MURDER, SUM(ATTEMPT\_TO\_MURDER) AS ATTEMPT\_TO\_MURDER, SUM(CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER) AS CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER,SUM(RAPE) AS RAPE,SUM(CUSTODIAL\_RAPE) AS CUSTODIAL\_RAPE,SUM(OTHER\_RAPE) AS OTHER\_RAPE,SUM(KIDNAPPING\_\_\_ABDUCTION) AS KIDNAPPING\_\_\_ABDUCTION,SUM(KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS) AS KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS,SUM(KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS) AS KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS,SUM(DACOITY) AS DACOITY,SUM(PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY) AS PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY,SUM(ROBBERY) AS ROBBERY,SUM(BURGLARY) AS BURGLARY,SUM(THEFT) AS THEFT,SUM(AUTO\_THEFT) AS AUTO\_THEFT,SUM(OTHER\_THEFT) AS OTHER\_THEFT,SUM(RIOTS) AS RIOTS,SUM(CRIMINAL\_BREACH\_OF\_TRUST) AS CRIMINAL\_BREACH\_OF\_TRUST,SUM(CHEATING) AS CHEATING,SUM(COUNTERFIETING) AS COUNTERFIETING,SUM(ARSON) AS ARSON,SUM(HURT\_GREVIOUS\_HURT) AS HURT\_GREVIOUS\_HURT,SUM(DOWRY\_DEATHS) AS DOWRY\_DEATHS,SUM(ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY) AS ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,SUM(INSULT\_TO\_MODESTY\_OF\_WOMEN) AS INSULT\_TO\_MODESTY\_OF\_WOMEN,SUM(CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES) AS CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES,SUM(IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES) AS IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES,SUM(CAUSING\_DEATH\_BY\_NEGLIGENCE) AS CAUSING\_DEATH\_BY\_NEGLIGENCE,SUM(OTHER\_IPC\_CRIMES) AS OTHER\_IPC\_CRIMES,SUM(TOTAL\_IPC\_CRIMES) AS TOTAL\_IPC\_CRIMES from filed GROUP BY STATE\_UT;

select YEAR, SUM(MURDER) AS MURDER, SUM(ATTEMPT\_TO\_MURDER) AS ATTEMPT\_TO\_MURDER, SUM(CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER) AS CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER,SUM(RAPE) AS RAPE,SUM(CUSTODIAL\_RAPE) AS CUSTODIAL\_RAPE,SUM(OTHER\_RAPE) AS OTHER\_RAPE,SUM(KIDNAPPING\_\_\_ABDUCTION) AS KIDNAPPING\_\_\_ABDUCTION,SUM(KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS) AS KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS,SUM(KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS) AS KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS,SUM(DACOITY) AS DACOITY,SUM(PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY) AS PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY,SUM(ROBBERY) AS ROBBERY,SUM(BURGLARY) AS BURGLARY,SUM(THEFT) AS THEFT,SUM(AUTO\_THEFT) AS AUTO\_THEFT,SUM(OTHER\_THEFT) AS OTHER\_THEFT,SUM(RIOTS) AS RIOTS,SUM(CRIMINAL\_BREACH\_OF\_TRUST) AS CRIMINAL\_BREACH\_OF\_TRUST,SUM(CHEATING) AS CHEATING,SUM(COUNTERFIETING) AS COUNTERFIETING,SUM(ARSON) AS ARSON,SUM(HURT\_GREVIOUS\_HURT) AS HURT\_GREVIOUS\_HURT,SUM(DOWRY\_DEATHS) AS DOWRY\_DEATHS,SUM(ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY) AS ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,SUM(INSULT\_TO\_MODESTY\_OF\_WOMEN) AS INSULT\_TO\_MODESTY\_OF\_WOMEN,SUM(CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES) AS CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES,SUM(IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES) AS IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES,SUM(CAUSING\_DEATH\_BY\_NEGLIGENCE) AS CAUSING\_DEATH\_BY\_NEGLIGENCE,SUM(OTHER\_IPC\_CRIMES) AS OTHER\_IPC\_CRIMES,SUM(TOTAL\_IPC\_CRIMES) AS TOTAL\_IPC\_CRIMES from filed GROUP BY YEAR;

select STATE\_UT, DISTRICT, SUM(MURDER) AS MURDER, SUM(ATTEMPT\_TO\_MURDER) AS ATTEMPT\_TO\_MURDER, SUM(CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER) AS CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER,SUM(RAPE) AS RAPE,SUM(CUSTODIAL\_RAPE) AS CUSTODIAL\_RAPE,SUM(OTHER\_RAPE) AS OTHER\_RAPE,SUM(KIDNAPPING\_\_\_ABDUCTION) AS KIDNAPPING\_\_\_ABDUCTION,SUM(KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS) AS KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS,SUM(KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS) AS KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS,SUM(DACOITY) AS DACOITY,SUM(PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY) AS PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY,SUM(ROBBERY) AS ROBBERY,SUM(BURGLARY) AS BURGLARY,SUM(THEFT) AS THEFT,SUM(AUTO\_THEFT) AS AUTO\_THEFT,SUM(OTHER\_THEFT) AS OTHER\_THEFT,SUM(RIOTS) AS RIOTS,SUM(CRIMINAL\_BREACH\_OF\_TRUST) AS CRIMINAL\_BREACH\_OF\_TRUST,SUM(CHEATING) AS CHEATING,SUM(COUNTERFIETING) AS COUNTERFIETING,SUM(ARSON) AS ARSON,SUM(HURT\_GREVIOUS\_HURT) AS HURT\_GREVIOUS\_HURT,SUM(DOWRY\_DEATHS) AS DOWRY\_DEATHS,SUM(ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY) AS ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,SUM(INSULT\_TO\_MODESTY\_OF\_WOMEN) AS INSULT\_TO\_MODESTY\_OF\_WOMEN,SUM(CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES) AS CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES,SUM(IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES) AS IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES,SUM(CAUSING\_DEATH\_BY\_NEGLIGENCE) AS CAUSING\_DEATH\_BY\_NEGLIGENCE,SUM(OTHER\_IPC\_CRIMES) AS OTHER\_IPC\_CRIMES,SUM(TOTAL\_IPC\_CRIMES) AS TOTAL\_IPC\_CRIMES from filed GROUP BY DISTRICT, STATE\_UT;

**Derived data sets obtained:**

1. DataGroupedByYEAR.csv
2. DataGroupedBySTATENames.csv
3. DataGroupedBySTATE\_DISTRICT\_Names.csv
4. DataGroupedBySTATE\_Year\_Names.csv

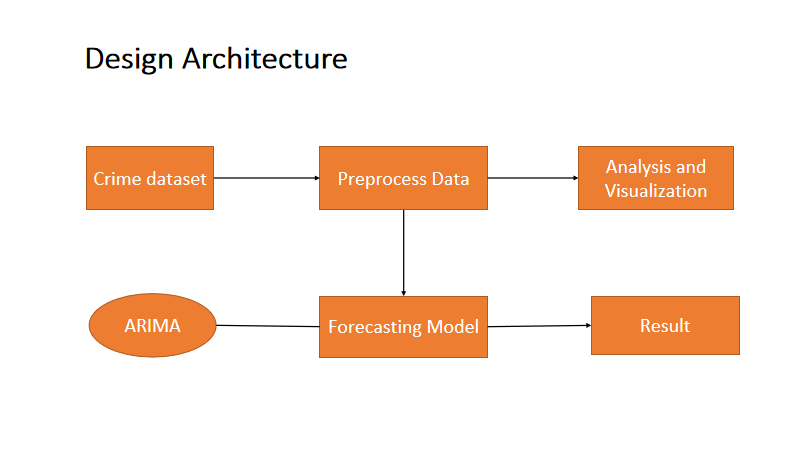
**All the data sets are attached with the report.**

1. **TIME SERIES ANLYSIS AND FORCASTING:**

Introduction :- ‘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 analyzes 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 neighborhood watch or patrol departments during the time of high probability for a crime or suggesting students or business travelers 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.



Technologies Used:-

The technologies used for implementing models and visualizations. Technical programming is performed in ‘R’ using its Machine Learning libraries.

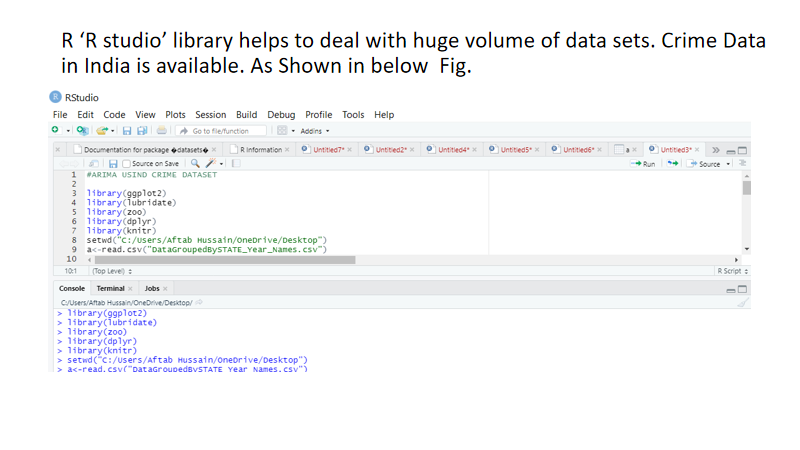
Data Processing :-

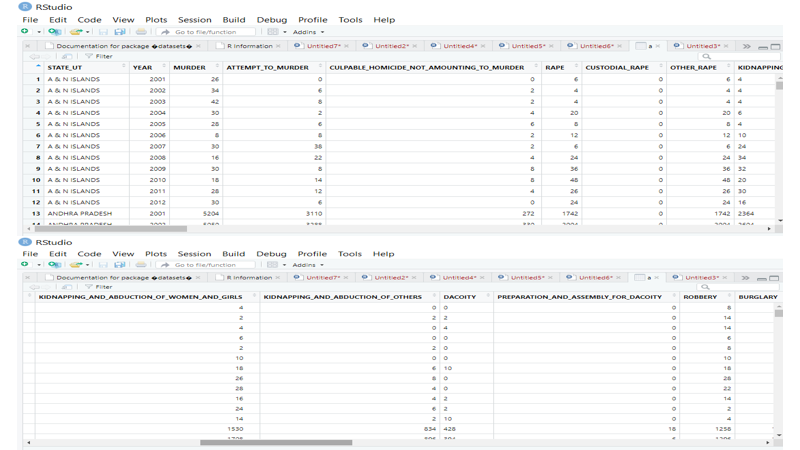
Data Sets considered for this project are crime data information of India. Kaggel gives information about various crimes in different States at India. In this project India with different Year and Total IPC Crime are selected.

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

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

Data Preprocessing 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 analyzes and takes data from 2001 to forecast future crime. So, first step of preprocessing is to extract data from 2001 to 2012.



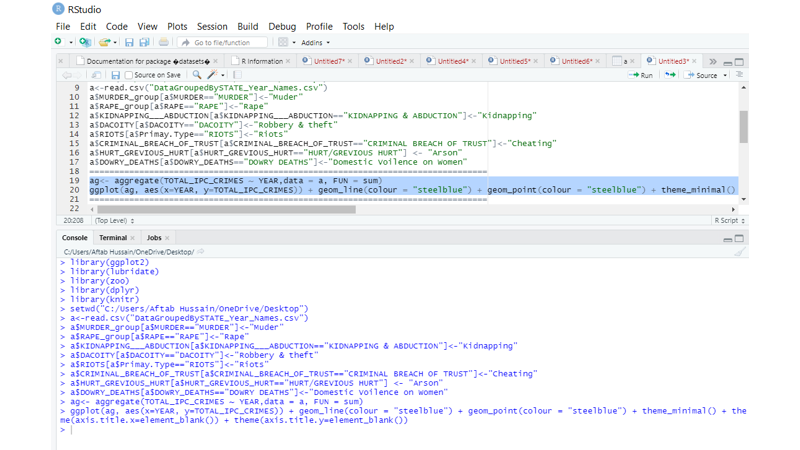


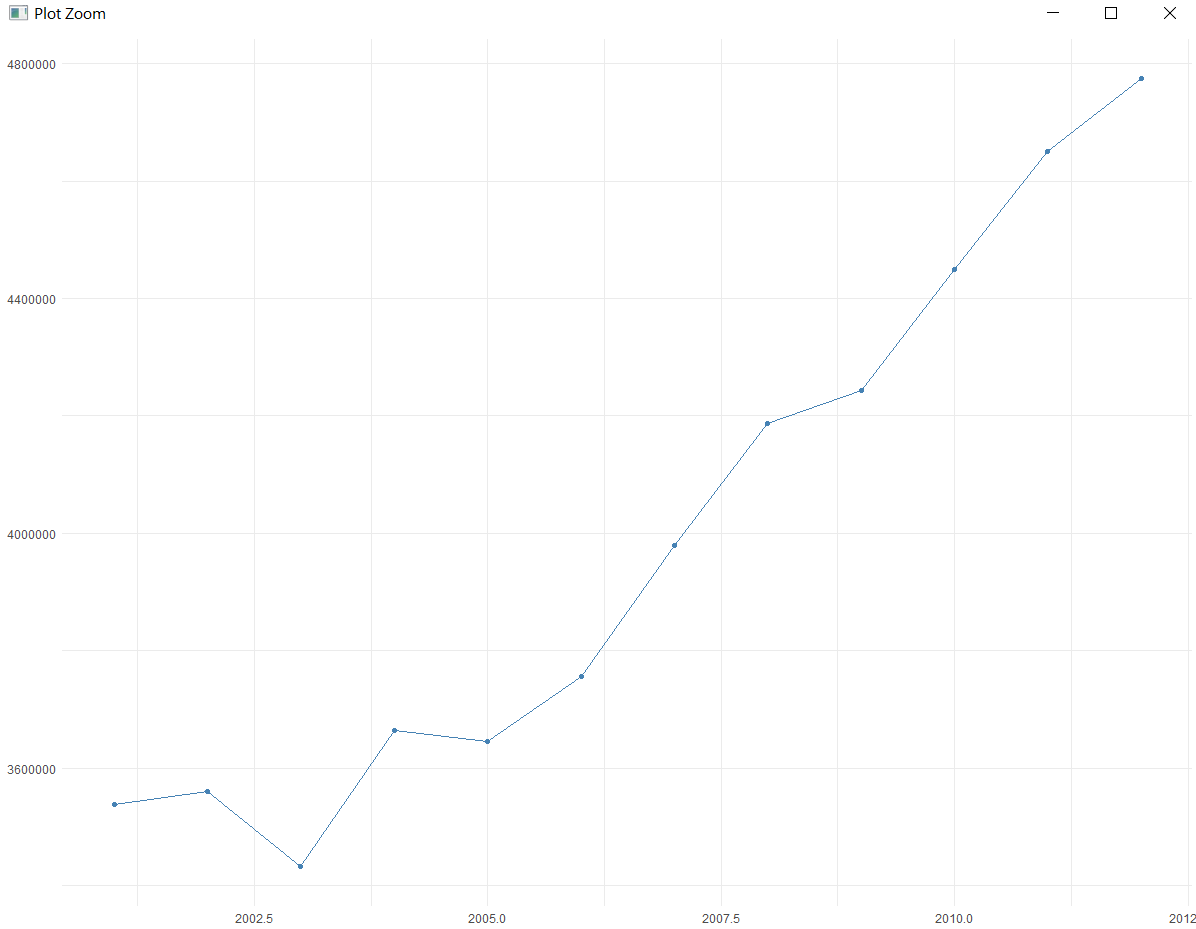
As , we don’t have month and Location So, we can’t Predict the which time and location are dangerous of which time of crime so we go though this dataset only.

**Data Exploration :-**

How has Crime Evolved over time in the India ?

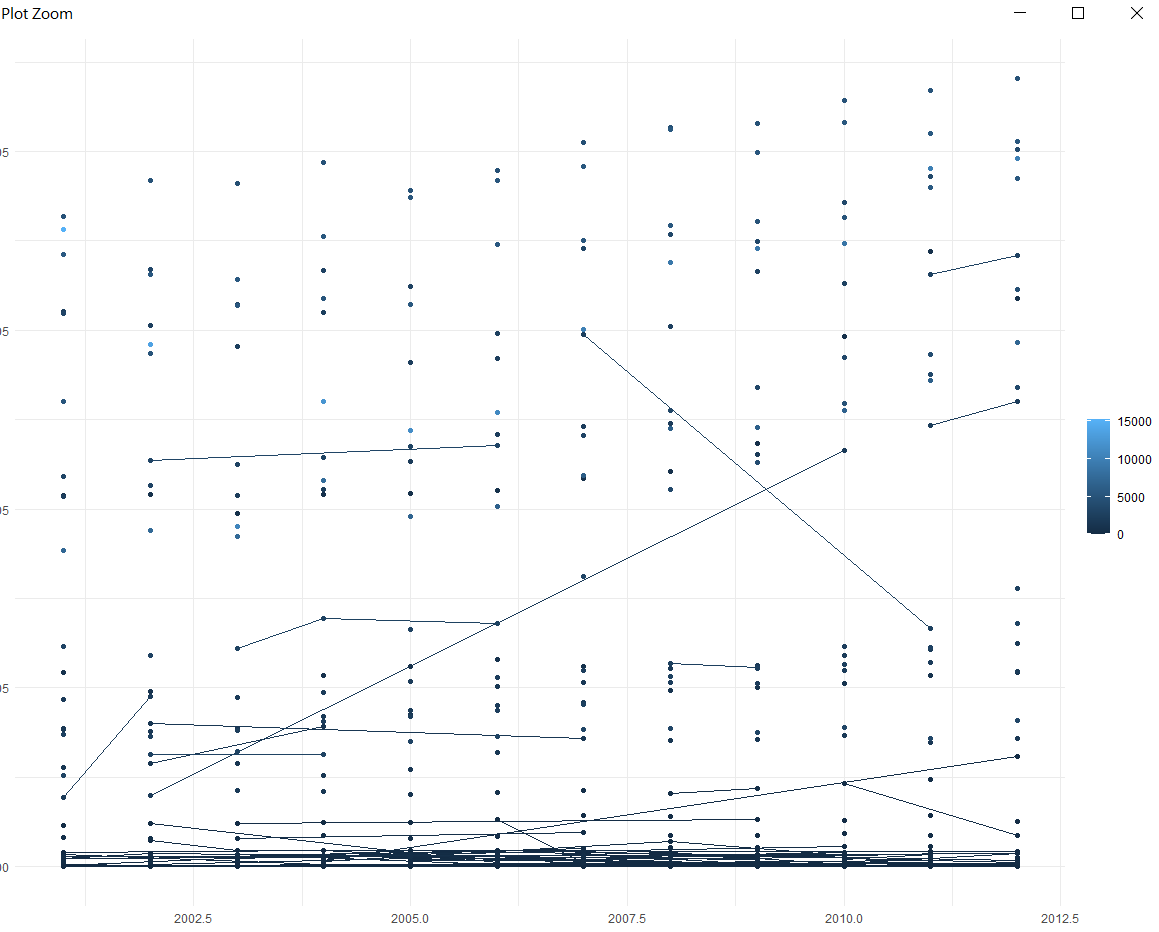
To answer the question we plot the number of crime per year from 2001 to 2012 . The graph shows that Crime in the India has been increased year after year with continuous Incline.



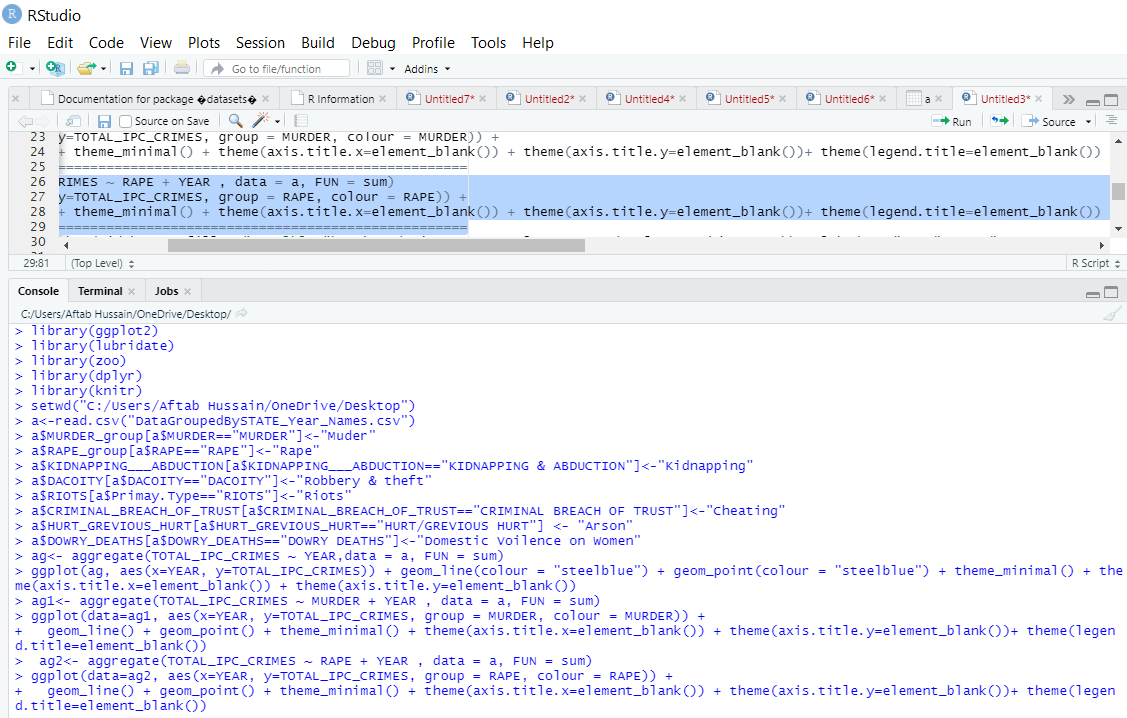


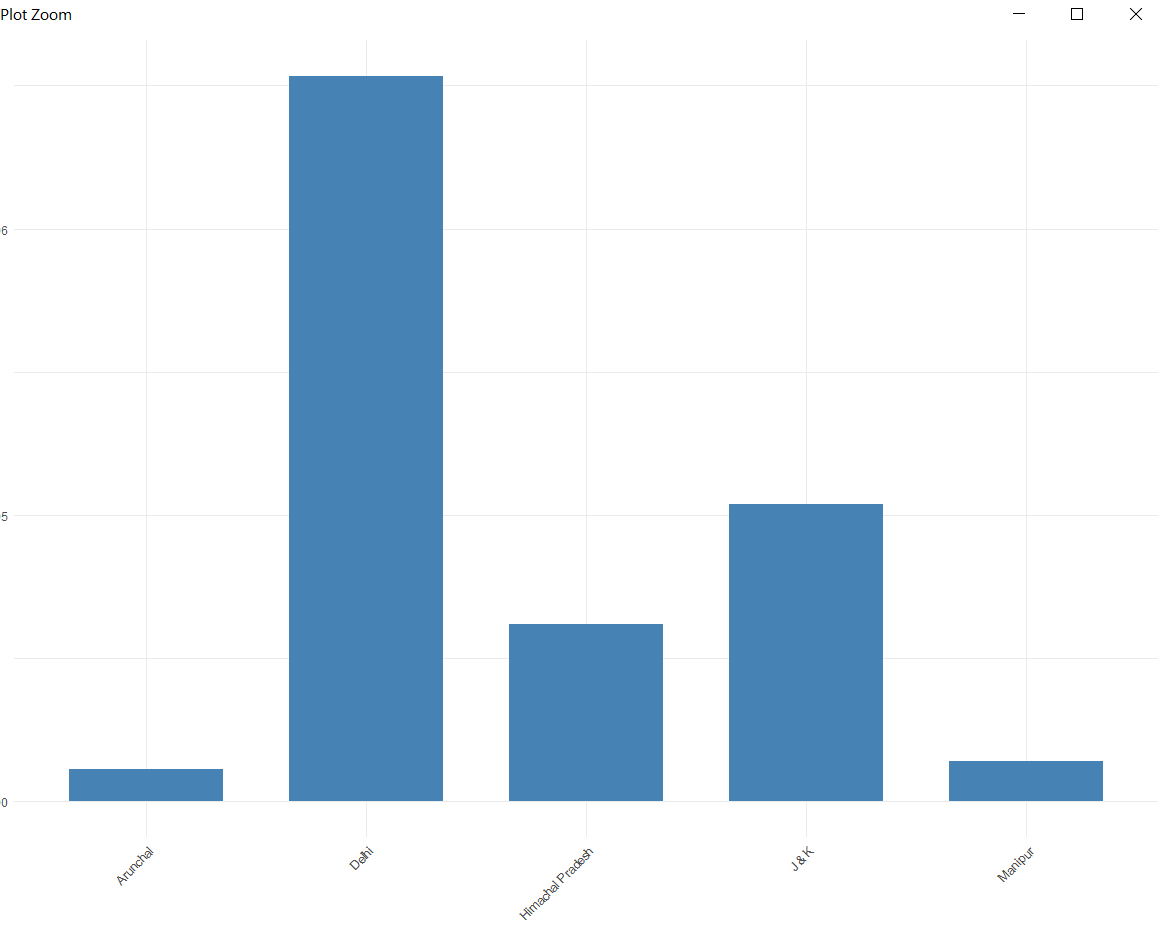
To understand the complexity in numbers, trend in the data and to observe the entire Murder data through a single visualization . Murder crime count in Year over the period of 11 years for the India is represented as map in given figures.

Darker color shades of blue represents lower crime count due to Murder in each Year and lighter color 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.



Like wise Murder this is graph for Rape . Higher crime rate due to rape in Year 2001 – 2012. We can see different crime rate due to different crime in year like Dowry, kidnapping , etc.





**CONCLUSION:**

Time Series Analysis and Forecasting is performed with several visualizations and statistical models in this project.

According to forecasting results after for the year 2002 crimes are slightly decreasing for Some Year then It Increasing as year Increases.

This forecasting results can help police to take necessary precautions according to the crime rate.

Crime Data analyzing with visualizations states that in 2001-2012 Delhi is the dangerous which have high crime occurrences.

Rape occures more than Murder because Murder range comes in (0-1500)

And Rape range has (0-6000).

Future work with this analysis is to predict the location of crime and tag the crime activities to a geographical map.

1. **Crime vs. Year plots using DataGroupedByYear.csv:**

**R code (VsYearPlots.R) :**

dataset = read.csv("DataGroupedByYear.csv")

data = dataset[,c(1,31)]

plot(data$YEAR,data$TOTAL\_IPC\_CRIMES,main=paste("Number of IPC Crimes Vs. Year"), xlab = "Year", ylab = "Number of IPC Crimes", type = "l")

data = dataset[,c(1,2)]

plot(data$YEAR,data$MURDER,main=paste("Number of MURDERs Vs. Year"), xlab = "Year", ylab = "Number of MURDERS", type = "l")

data = dataset[,c(1,3)]

plot(data$YEAR,data$ATTEMPT\_TO\_MURDER,main=paste("Number of ATTEMPT\_TO\_MURDER Vs. Year"), xlab = "Year", ylab = "Number of ATTEMPT\_TO\_MURDER", type = "l")

data = dataset[,c(1,4)]

plot(data$YEAR,data$CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER,main=paste("Number of CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER Vs. Year"), xlab = "Year", ylab = "Number of CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER", type = "l")

data = dataset[,c(1,5)]

plot(data$YEAR,data$RAPE,main=paste("Number of RAPE Vs. Year"), xlab = "Year", ylab = "Number of RAPE", type = "l")

data = dataset[,c(1,6)]

plot(data$YEAR,data$CUSTODIAL\_RAPE,main=paste("Number of CUSTODIAL\_RAPE Vs. Year"), xlab = "Year", ylab = "Number of CUSTODIAL\_RAPE", type = "l")

data = dataset[,c(1,7)]

plot(data$YEAR,data$OTHER\_RAPE,main=paste("Number of OTHER\_RAPE Vs. Year"), xlab = "Year", ylab = "Number of OTHER\_RAPE", type = "l")

data = dataset[,c(1,8)]

plot(data$YEAR,data$KIDNAPPING\_\_\_ABDUCTION,main=paste("Number of KIDNAPPING\_\_\_ABDUCTION Vs. Year"), xlab = "Year", ylab = "Number of KIDNAPPING\_\_\_ABDUCTION", type = "l")

data = dataset[,c(1,9)]

plot(data$YEAR,data$KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS,main=paste("Number of KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS Vs. Year"), xlab = "Year", ylab = "Number of KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS", type = "l")

data = dataset[,c(1,10)]

plot(data$YEAR,data$KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS,main=paste("Number of KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS Vs. Year"), xlab = "Year", ylab = "Number of KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS", type = "l")

data = dataset[,c(1,11)]

plot(data$YEAR,data$DACOITY,main=paste("Number of DACOITY Vs. Year"), xlab = "Year", ylab = "Number of DACOITY", type = "l")

data = dataset[,c(1,12)]

plot(data$YEAR,data$PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY,main=paste("Number of PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY Vs. Year"), xlab = "Year", ylab = "Number of PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY", type = "l")

data = dataset[,c(1,13)]

plot(data$YEAR,data$ROBBERY,main=paste("Number of ROBBERY Vs. Year"), xlab = "Year", ylab = "Number of ROBBERY", type = "l")

data = dataset[,c(1,14)]

plot(data$YEAR,data$BURGLARY,main=paste("Number of BURGLARY Vs. Year"), xlab = "Year", ylab = "Number of BURGLARY", type = "l")

data = dataset[,c(1,15)]

plot(data$YEAR,data$THEFT,main=paste("Number of THEFT Vs. Year"), xlab = "Year", ylab = "Number of THEFT", type = "l")

data = dataset[,c(1,16)]

plot(data$YEAR,data$AUTO\_THEFT,main=paste("Number of AUTO\_THEFT Vs. Year"), xlab = "Year", ylab = "Number of AUTO\_THEFT", type = "l")

data = dataset[,c(1,17)]

plot(data$YEAR,data$OTHER\_THEFT,main=paste("Number of OTHER\_THEFT Vs. Year"), xlab = "Year", ylab = "Number of OTHER\_THEFT", type = "l")

data = dataset[,c(1,18)]

plot(data$YEAR,data$RIOTS,main=paste("Number of RIOTS Vs. Year"), xlab = "Year", ylab = "Number of RIOTS", type = "l")

data = dataset[,c(1,19)]

plot(data$YEAR,data$CRIMINAL\_BREACH\_OF\_TRUST,main=paste("Number of CRIMINAL\_BREACH\_OF\_TRUST Vs. Year"), xlab = "Year", ylab = "Number of CRIMINAL\_BREACH\_OF\_TRUST", type = "l")

data = dataset[,c(1,20)]

plot(data$YEAR,data$CHEATING,main=paste("Number of CHEATING Vs. Year"), xlab = "Year", ylab = "Number of CHEATING", type = "l")

data = dataset[,c(1,21)]

plot(data$YEAR,data$COUNTERFIETING,main=paste("Number of COUNTERFIETING Vs. Year"), xlab = "Year", ylab = "Number of COUNTERFIETING", type = "l")

data = dataset[,c(1,22)]

plot(data$YEAR,data$ARSON,main=paste("Number of ARSON Vs. Year"), xlab = "Year", ylab = "Number of ARSON", type = "l")

data = dataset[,c(1,23)]

plot(data$YEAR,data$HURT\_GREVIOUS\_HURT,main=paste("Number of HURT\_GREVIOUS\_HURT Vs. Year"), xlab = "Year", ylab = "Number of HURT\_GREVIOUS\_HURT", type = "l")

data = dataset[,c(1,24)]

plot(data$YEAR,data$DOWRY\_DEATHS,main=paste("Number of DOWRY\_DEATHS Vs. Year"), xlab = "Year", ylab = "Number of DOWRY\_DEATHS", type = "l")

data = dataset[,c(1,25)]

plot(data$YEAR,data$ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,main=paste("Number of ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY Vs. Year"), xlab = "Year", ylab = "Number of ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY", type = "l")

data = dataset[,c(1,26)]

plot(data$YEAR,data$INSULT\_TO\_MODESTY\_OF\_WOMEN,main=paste("Number of INSULT\_TO\_MODESTY\_OF\_WOMEN Vs. Year"), xlab = "Year", ylab = "Number of INSULT\_TO\_MODESTY\_OF\_WOMEN", type = "l")

data = dataset[,c(1,27)]

plot(data$YEAR,data$CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES,main=paste("Number of CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES Vs. Year"), xlab = "Year", ylab = "Number of CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES", type = "l")

data = dataset[,c(1,28)]

plot(data$YEAR,data$IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES,main=paste("Number of IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES Vs. Year"), xlab = "Year", ylab = "Number of IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES", type = "l")

data = dataset[,c(1,29)]

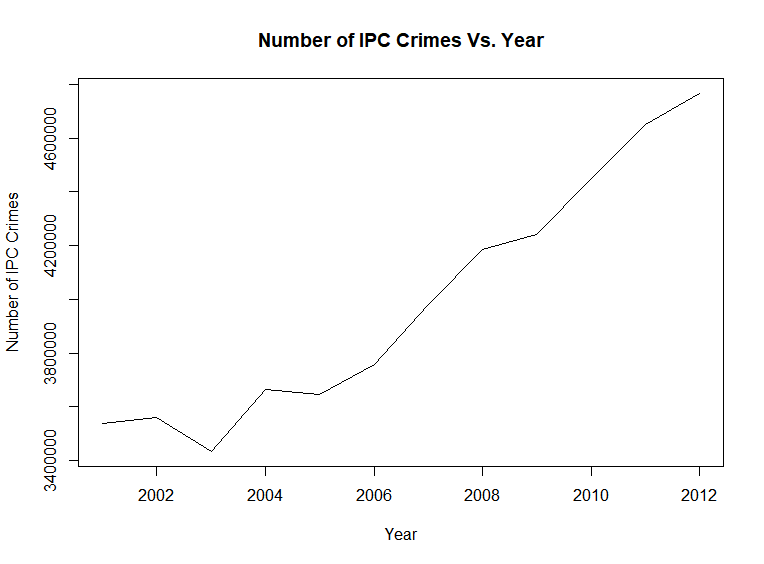
plot(data$YEAR,data$CAUSING\_DEATH\_BY\_NEGLIGENCE,main=paste("Number of CAUSING\_DEATH\_BY\_NEGLIGENCE Vs. Year"), xlab = "Year", ylab = "Number of CAUSING\_DEATH\_BY\_NEGLIGENCE", type = "l")

data = dataset[,c(1,30)]

plot(data$YEAR,data$OTHER\_IPC\_CRIMES,main=paste("Number of OTHER\_IPC\_CRIMES Vs. Year"), xlab = "Year", ylab = "Number of OTHER\_IPC\_CRIMES", type = "l")

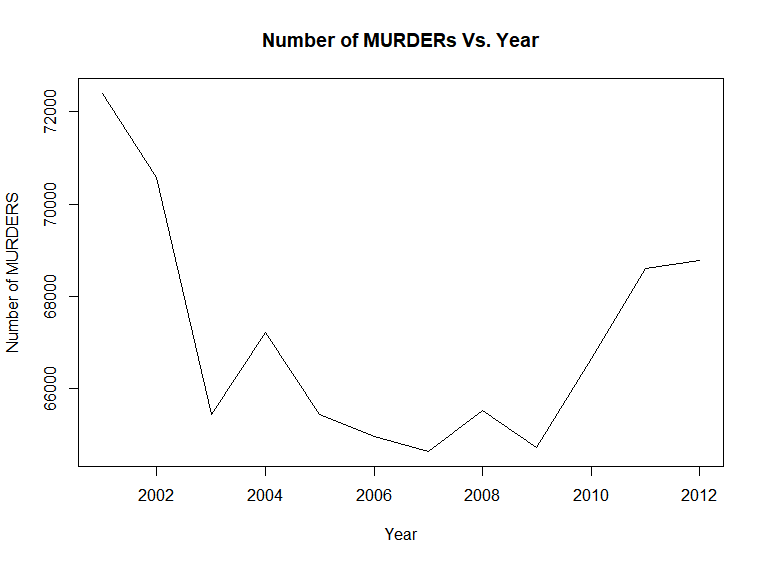
**Plots:**

Random movement



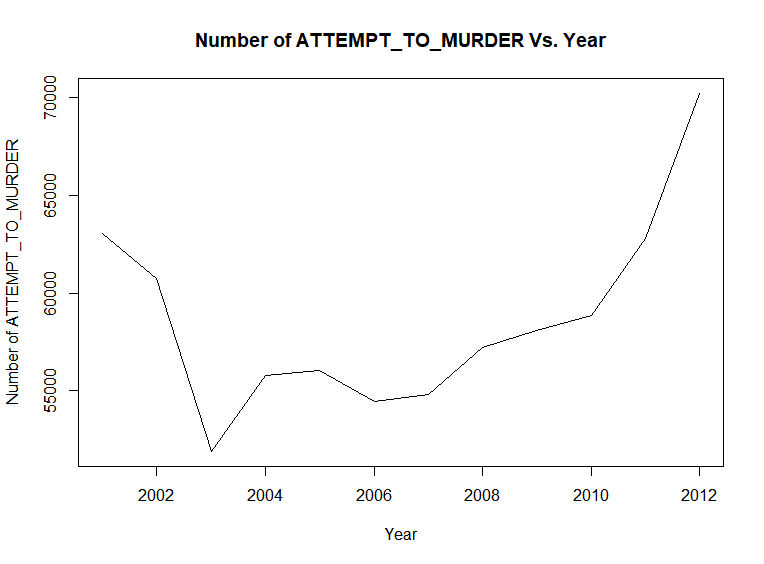
**Analysis**:

1. A **Trend** can be observed from the above plot, ipc crime count was increasing in the period of 2001 to 2012 in India.
2. But we have a **random movement** due to decrease of ipc crime count in between year 2002 and 2004 due to lowest ipc crime count around 3,40,000 cases registered in the year 2003.
3. Lowest ipc crime count, around 34,00,000 registered cases in 2003 and highest around 48,00,000 registered cases in 2012.



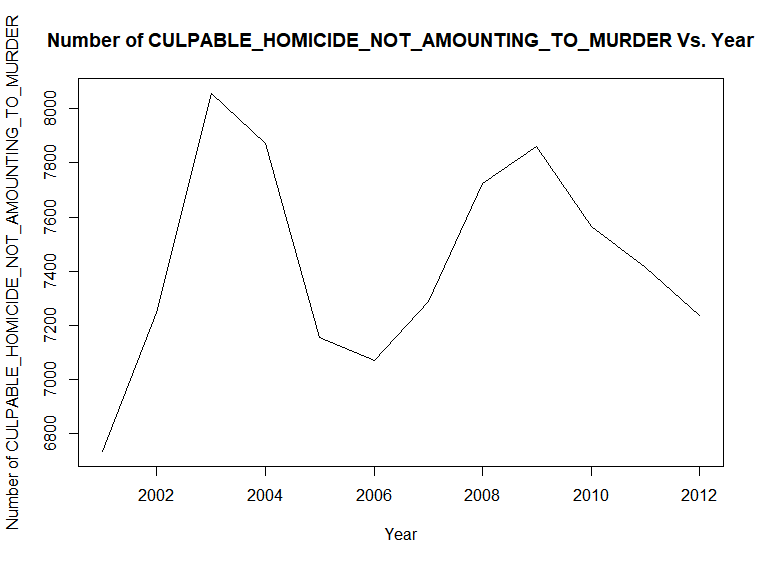
**Analysis:**

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.



**Analysis:**

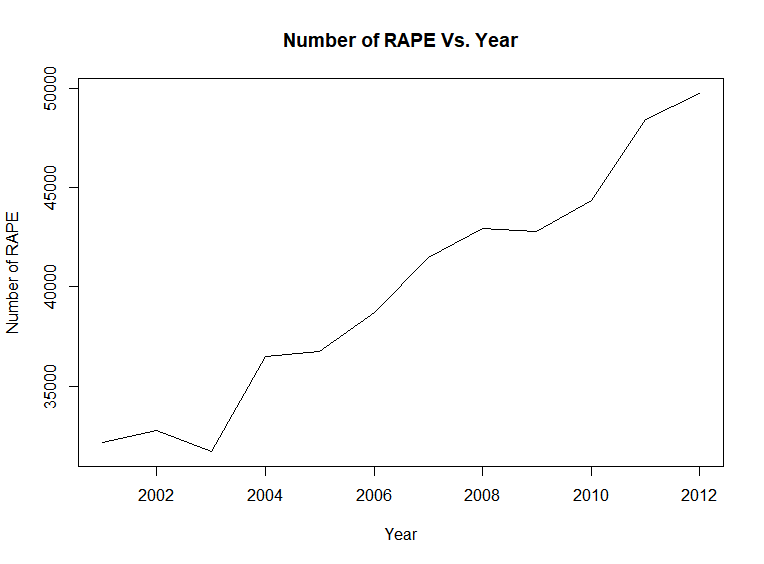
1. From 2001 to 2003 decrease in attempt to murder cases is observed and from 2003 to 2012 increase in attempt to murder cases is observed expect with decrease in number of cases in 2005 and 2006.
2. Lowest number of cases registered in 2003 with in the period of 2001 to 2012.
3. Highest number of cases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



CYCLIC PATTERN

**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

**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:

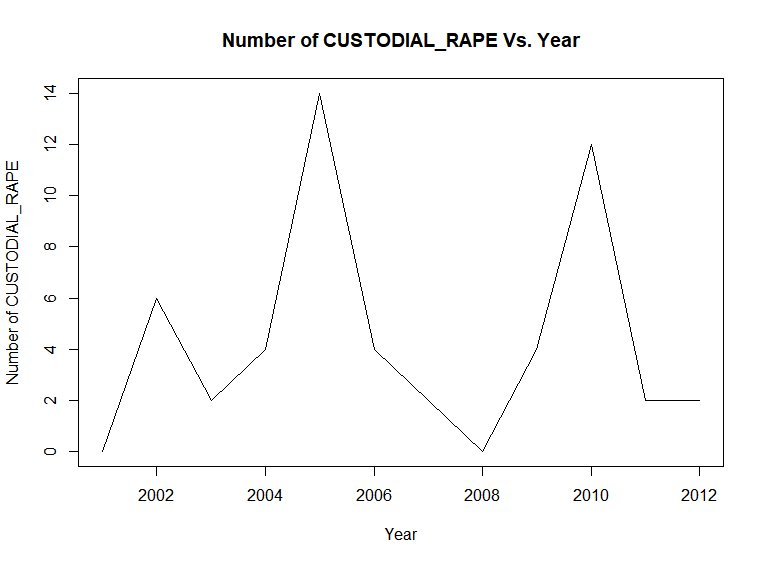
Highest total ipc crime cases registered = 4768564/49752 = 95.8

Highest number of registered rape cases

Lowest rape cases registered =3432240/31694=108.2930523127406

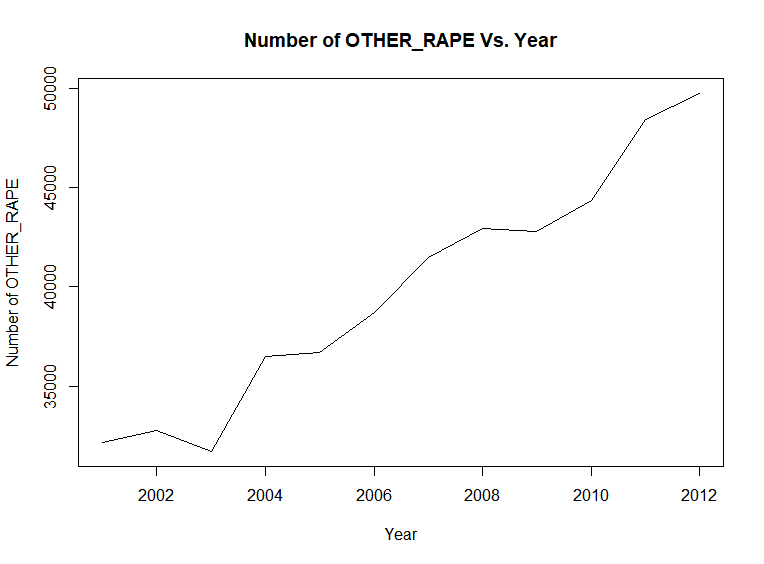
Lowest number of registered rape cases

1. **From this we can conclude that number of total ipc crime cases registered and number of rape cases registered in 2001 to 2012 years are positively correlated and hence, as the total crime increased the rape cases contributing to it also increased.**



**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

**Analysis:**

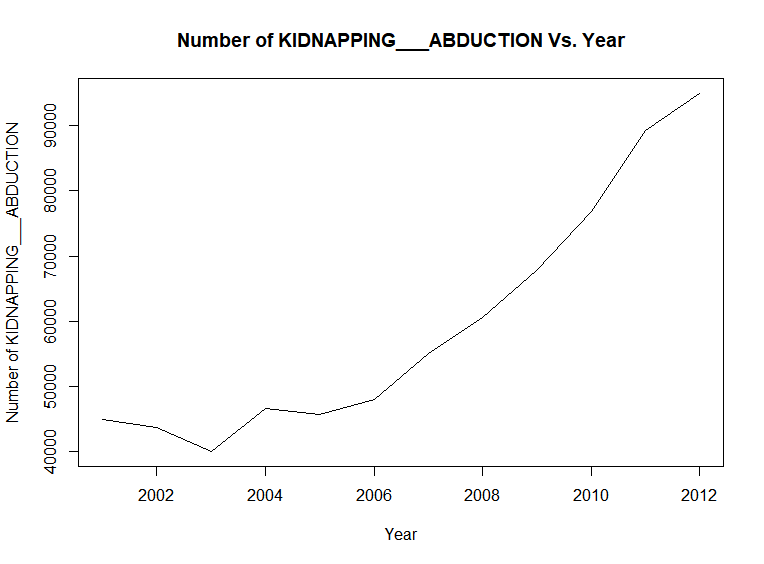
1)A **Trend** can be observed from the above plot, number of registered other\_rape cases was increasing in the period of 2001 to 2012 in India.

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.**

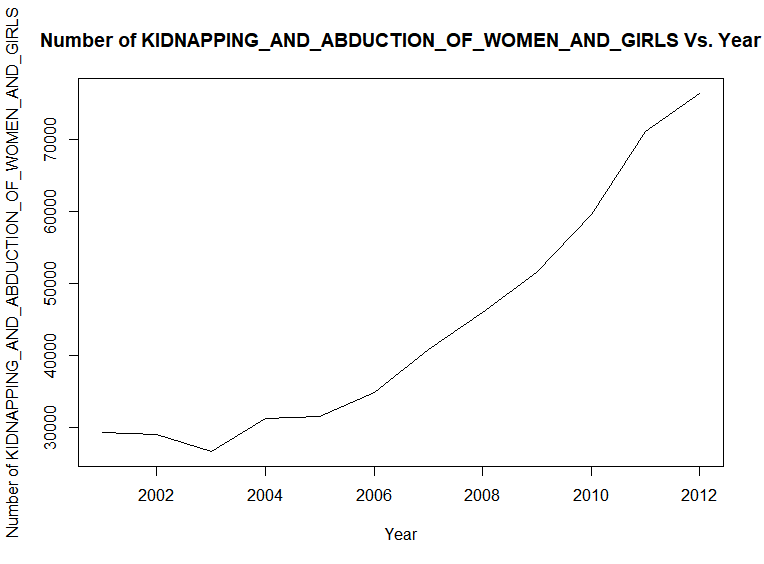
**Hence, total\_ipc\_crime count, registered rape cases count and registered other\_rape cases count are highly positively correlated.**



Random movement

**Analysis:**

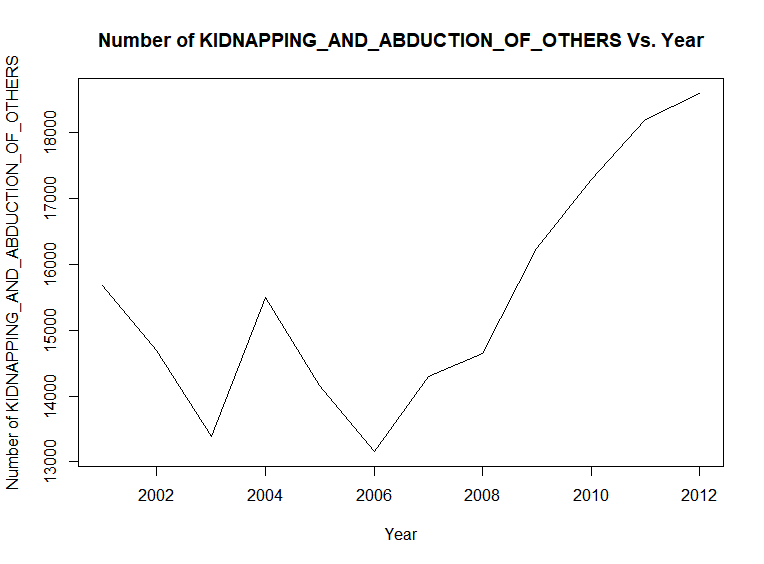
1. A **Trend** can be observed from the above plot, number of registered Kidnapping and Abduction cases was increasing in the period of 2001 to 2012 in India.
2. But we have a **random movement** due to decrease of number of registered Kidnapping and Abduction cases in between year 2002 and 2004 due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

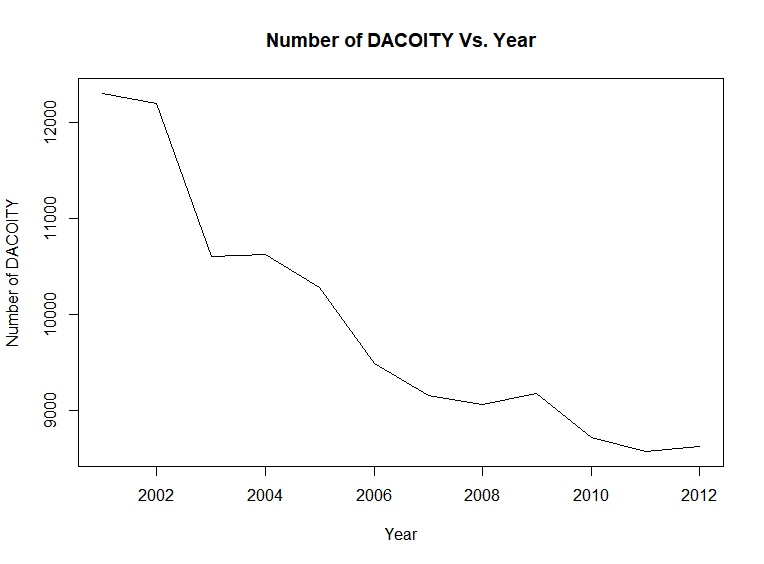
**Analysis:**

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**



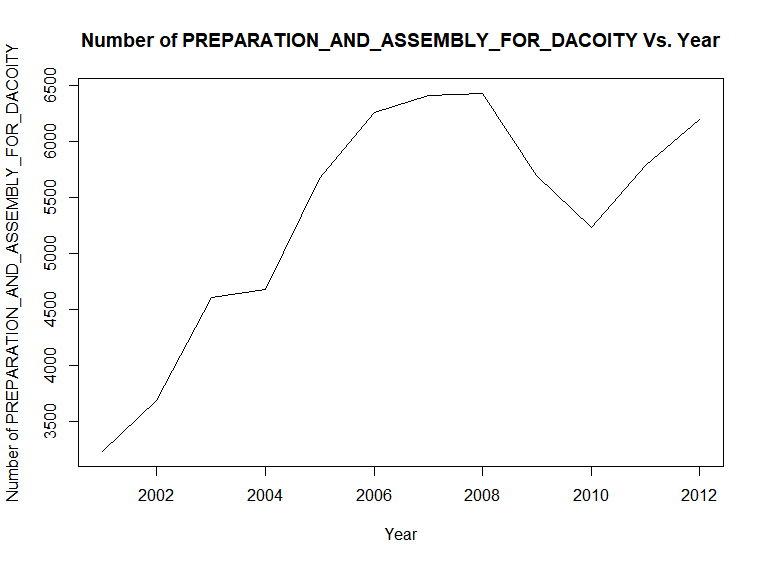
**Analysis:**

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.



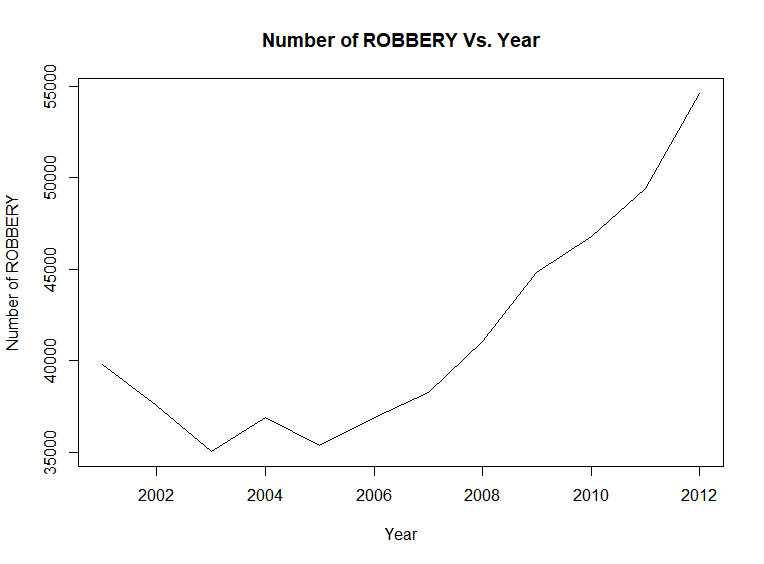
**Analysis:**

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.



**Analysis:**

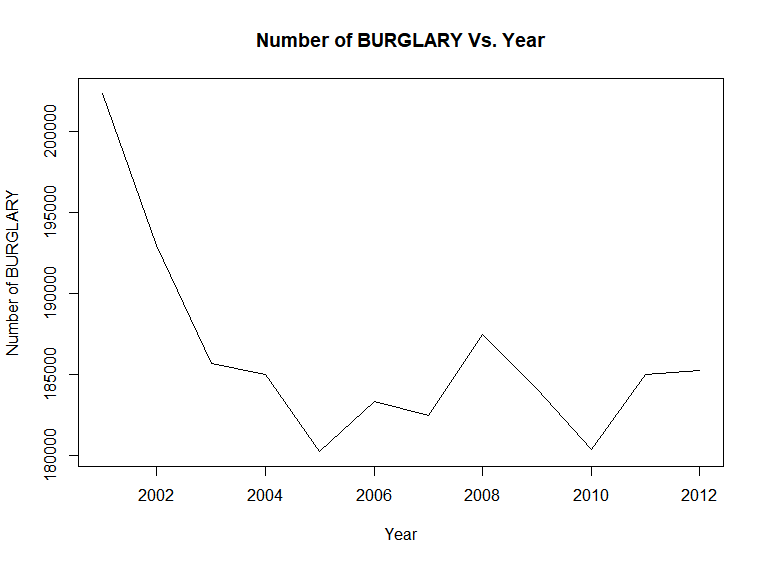
1. A trend can be observed from 2001 to 2008, number of cases registered under preparation and assembly for dacoity was increasing in 2001 to 2008.
2. Lowest number of cases registered in 2001, around 3,550 cases.
3. Highest number of cases registered in 2008, around 6,125 cases.



Random movement

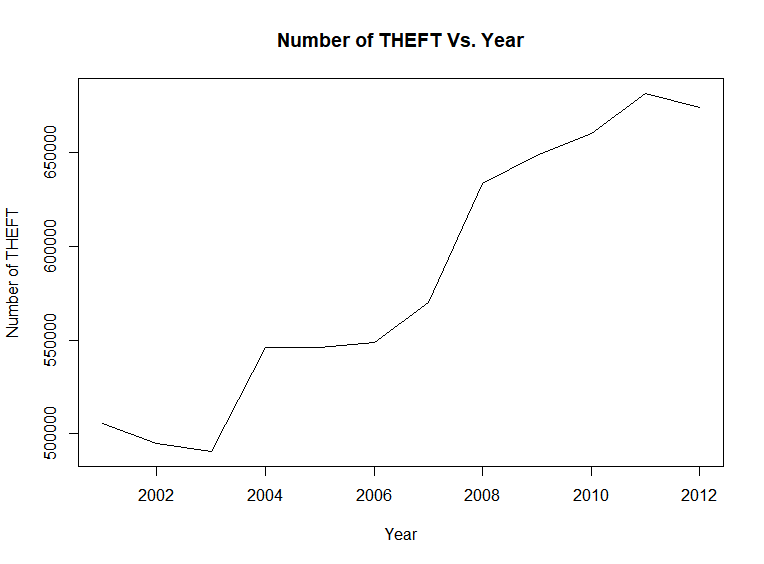
**Analysis:**

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.



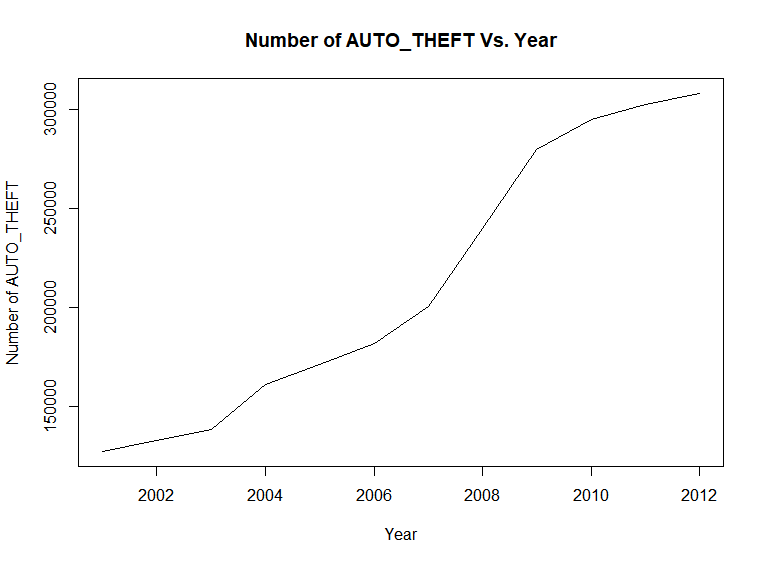
**Analysis:**

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.



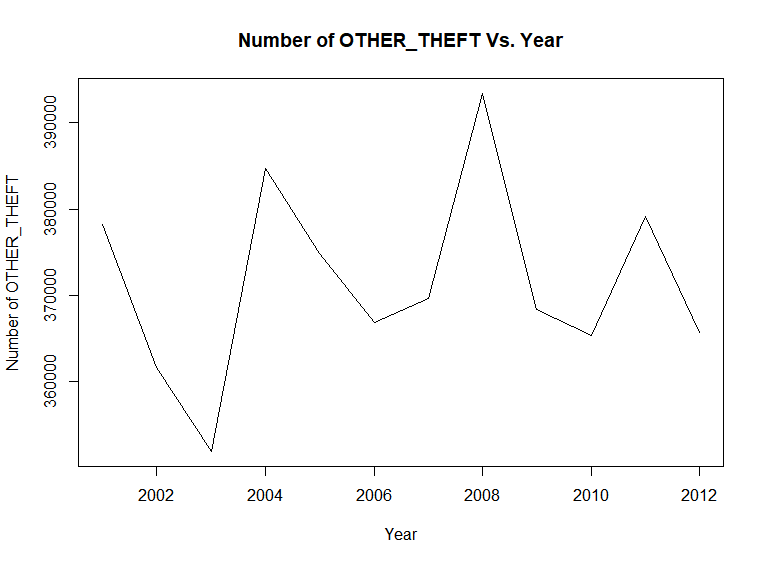
**Analysis:**

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.



**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.

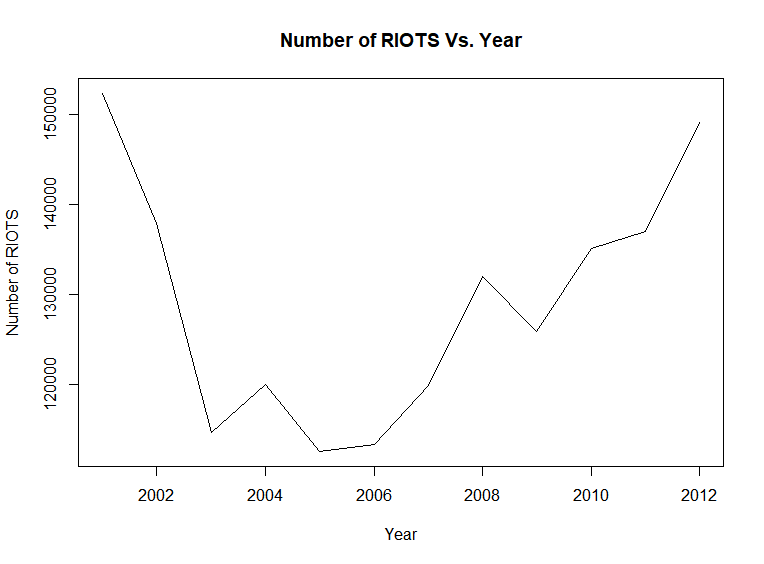


CYCLIC

PATTERN

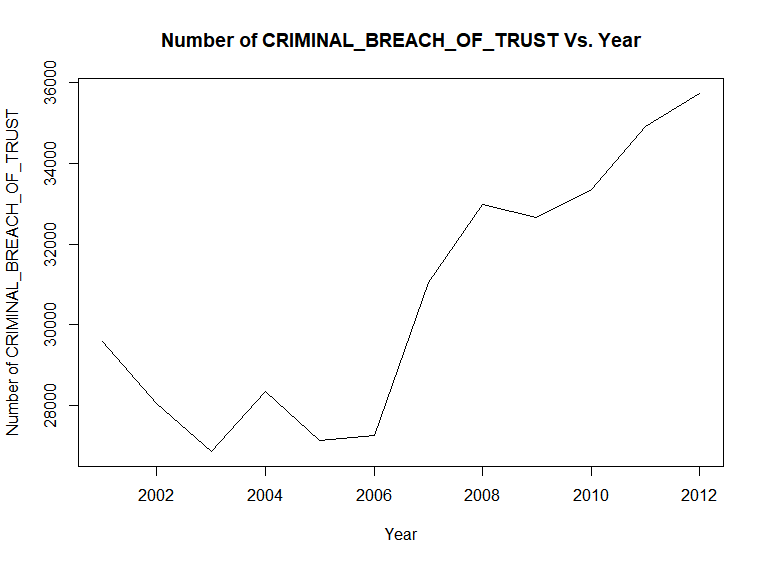
**Analysis:**

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.



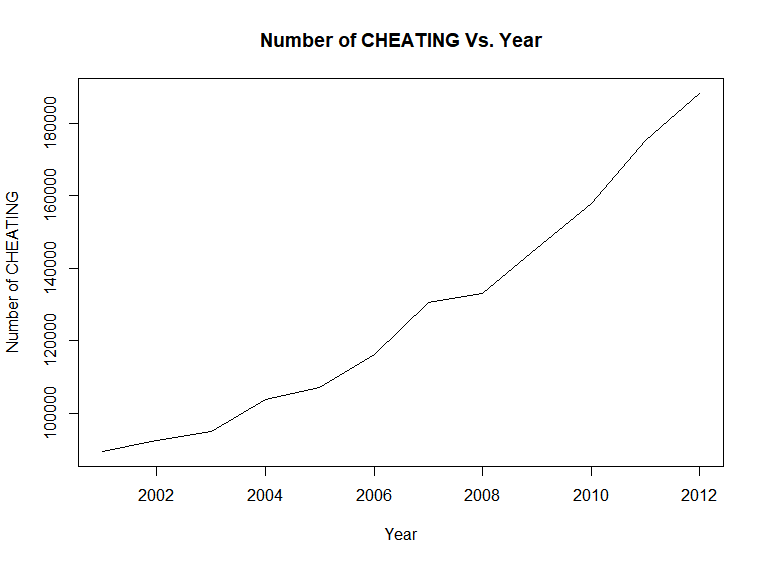
**Analysis:**

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.



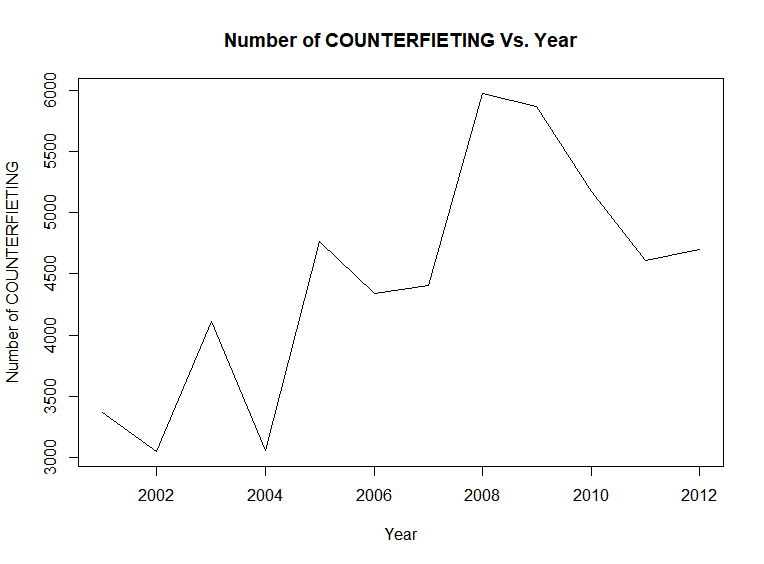
**Analysis:**

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.



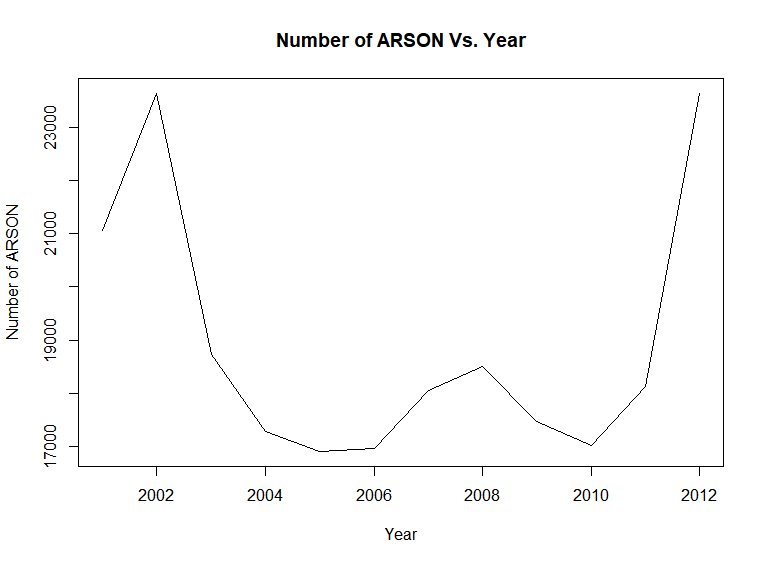
**Analysis:**

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.



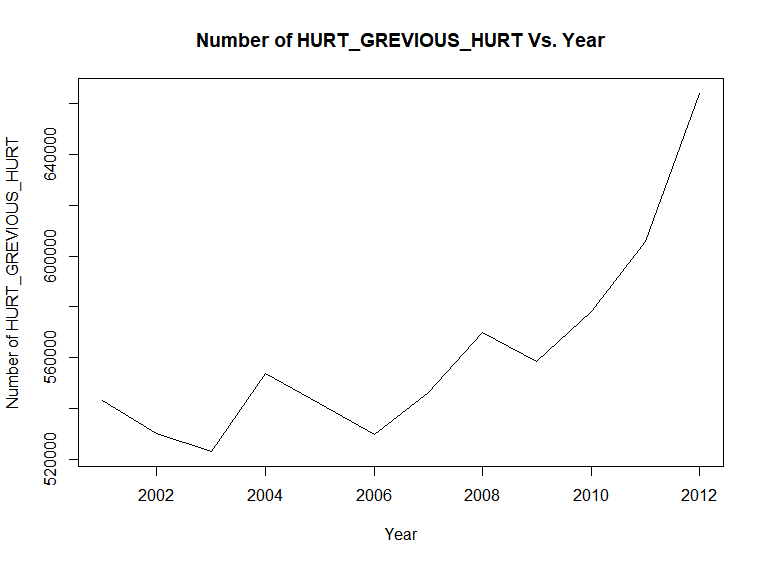
**Analysis:**

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.



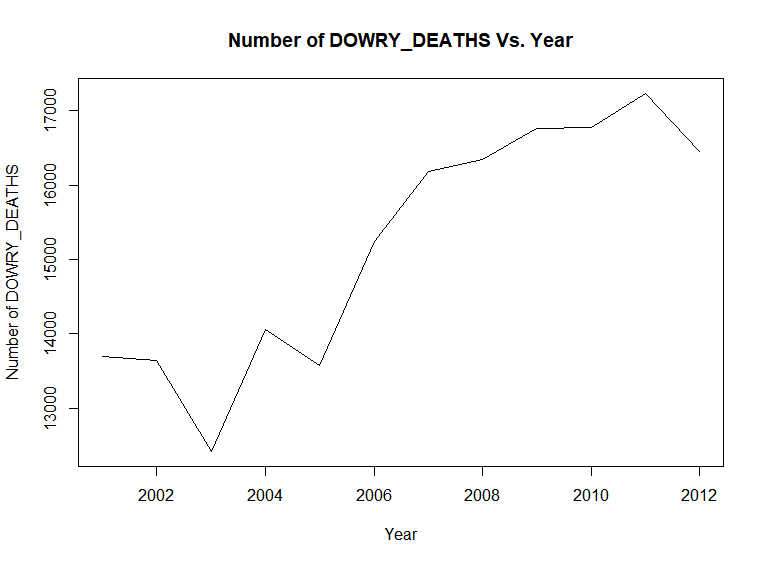
**Analysis:**

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.



**Analysis:**

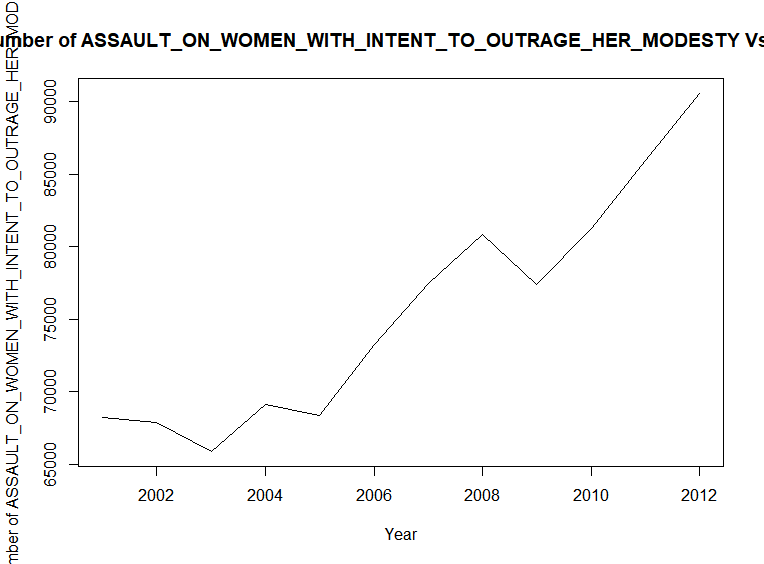
1. A trend with cyclic variation can be observed in the above plot, from 2001 to 2009 number of cases was increasing with a cyclic variation and after that the number of cases monotonically increased.
2. Lowest number of cases registered in 2003 with around 5,20,000 cases.
3. Highest number of cases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



Random Movements

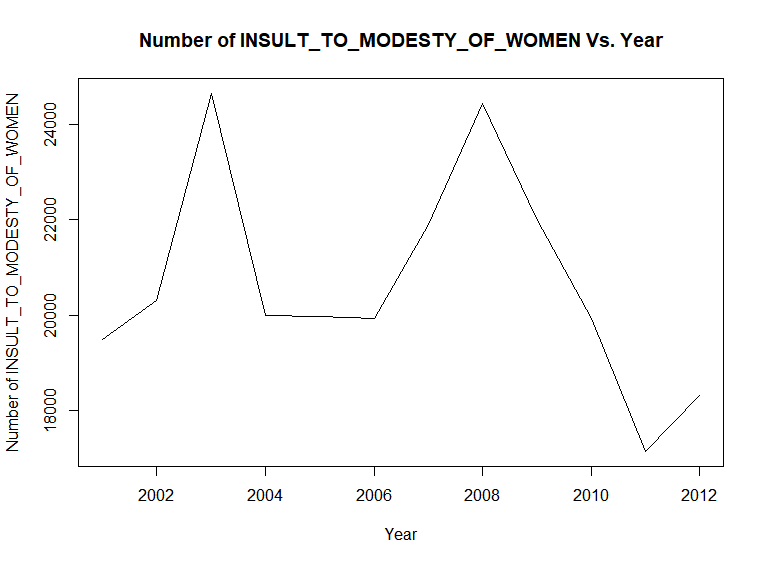
**Analysis:**

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.



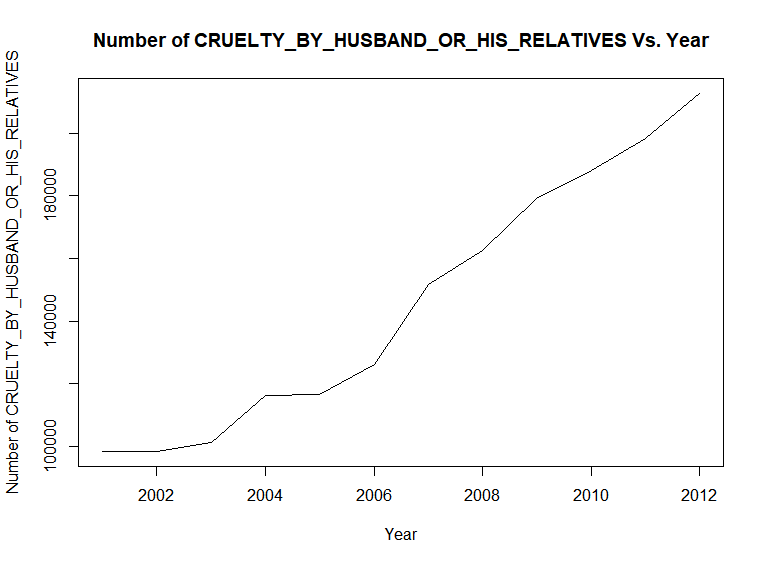
**Analysis:**

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.



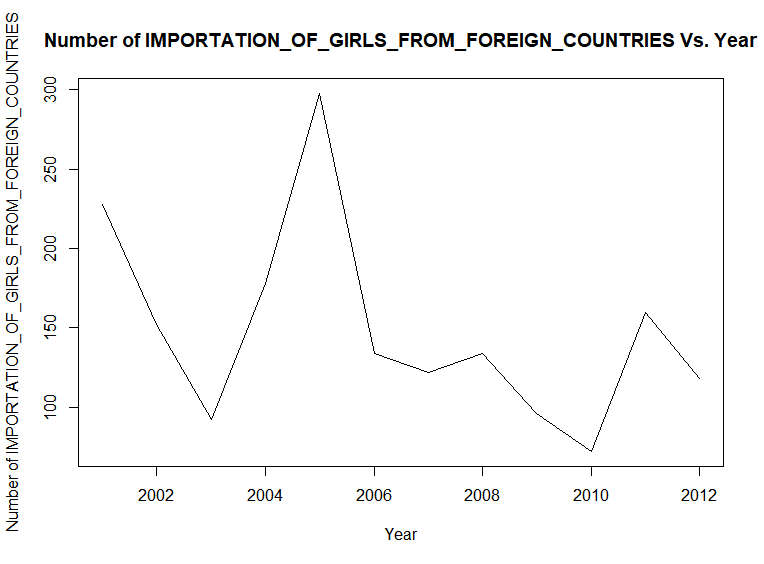
**Analysis:**

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.



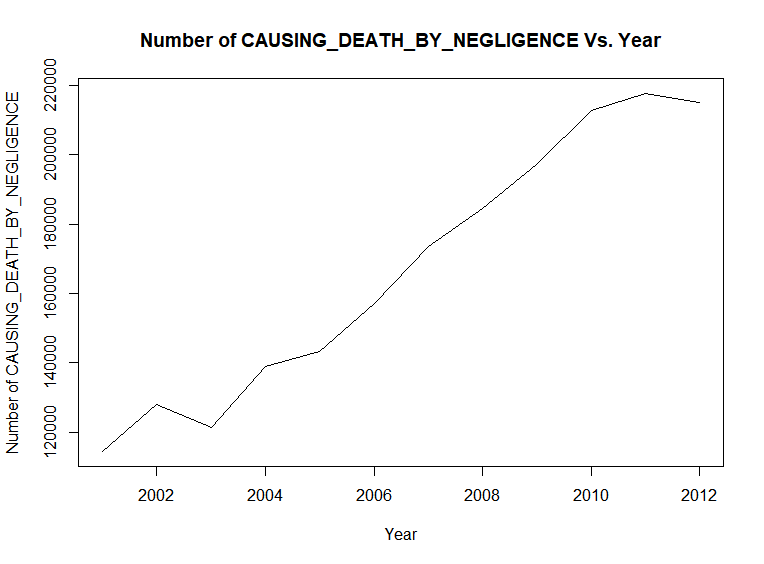
**Analysis:**

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.



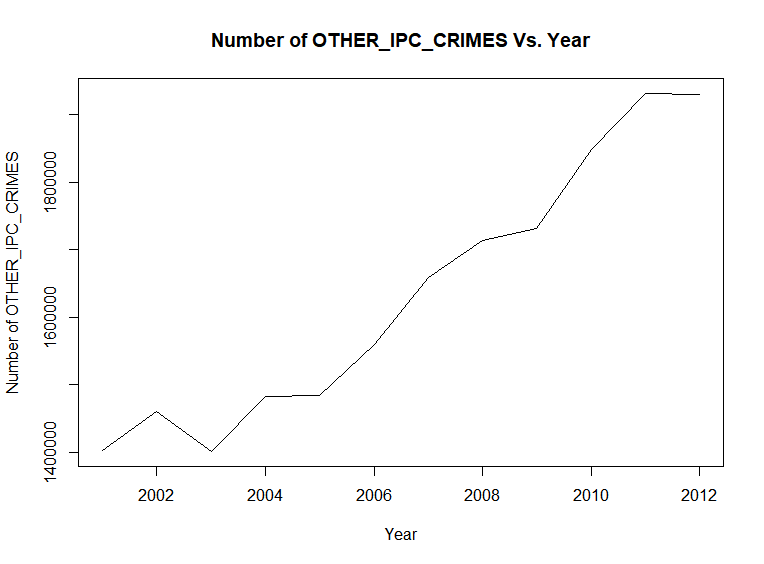
**Analysis:**

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.



**Analysis:**

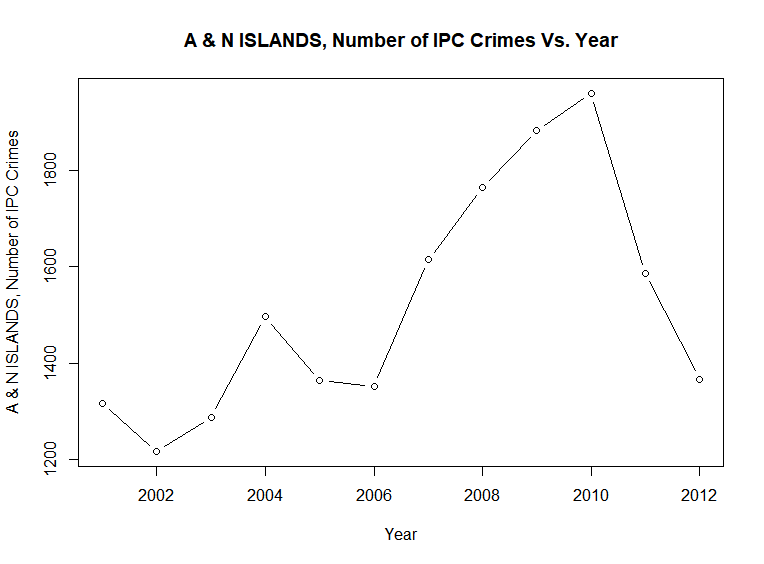
1. A trend can be observed in the above plot, number of cases registered under causing death by negligence was increasing from 2001 to 2012.
2. Lowest number of registered cases in 2001 with around 1,27,000 cases.
3. Highest number of cases registered in 2011 with around 2,17,000 cases.

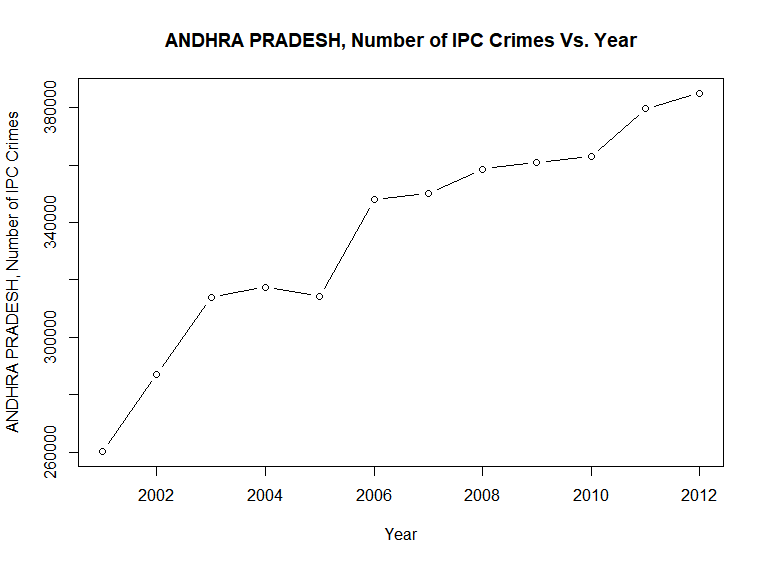


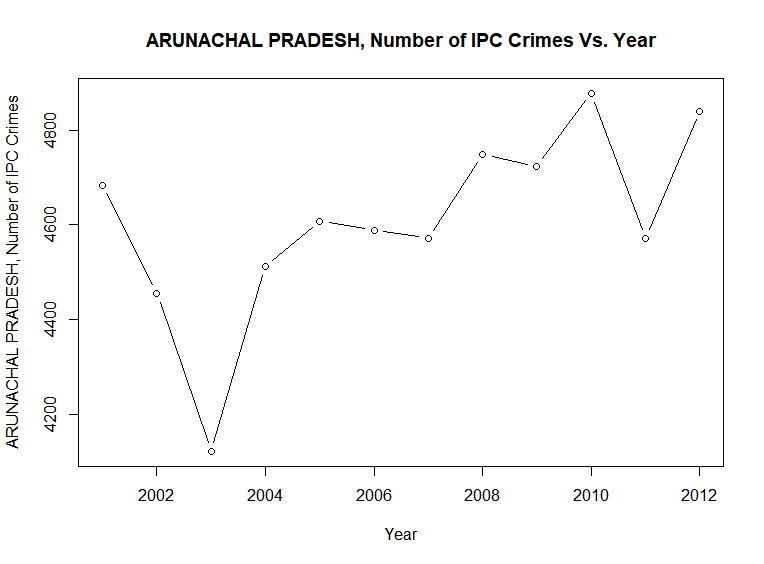
**Analysis:**

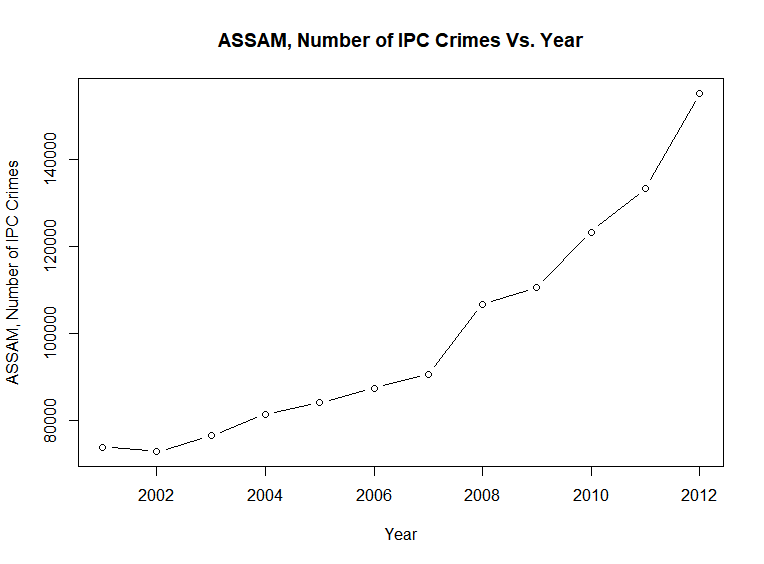
1. A trend can be observed in the above plot, the number of cases under other ipc crimes increased from 2001 to 2012.
2. Lowest number of cases registered in 2003 with around 1,40,000 cases.
3. Highest number of cases registered in 2011 with around 1,90,000 cases.

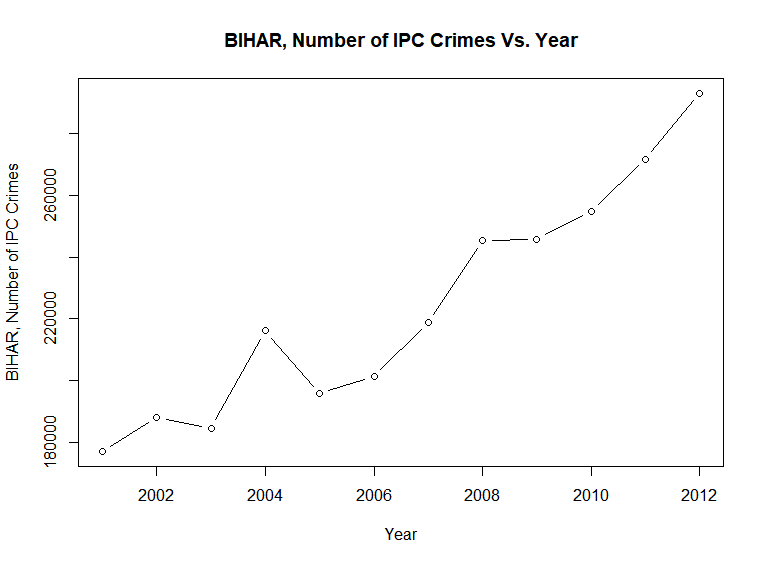
## We also wrote a code in R for state wise Crime vs. Plots, which we have attached with the report but not added the plots in the report because the code produces 35 \* 30 = 1050 plots, r code is also of 1086 lines, so we are not pasting the code in the report, but we are pasting few State wise TOTAL\_IPC\_CRIMES vs. Year plots using DataGroupedBySTATE\_Year\_Names.csv data set :

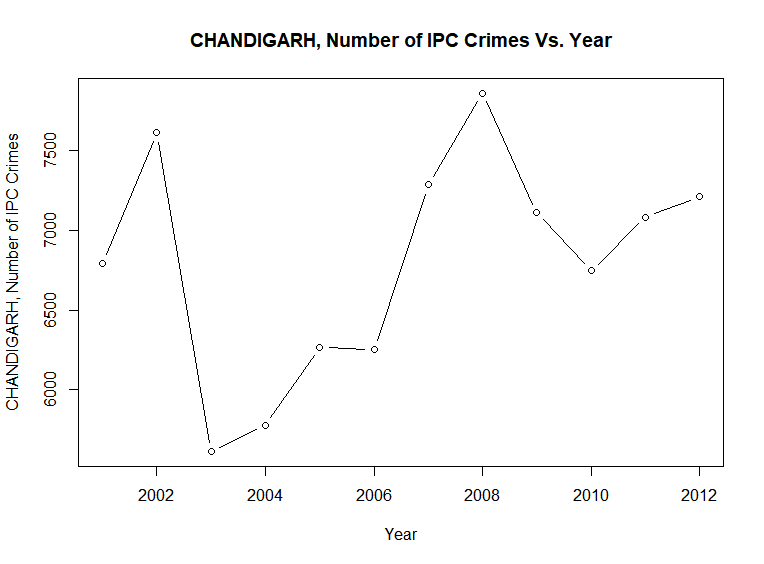


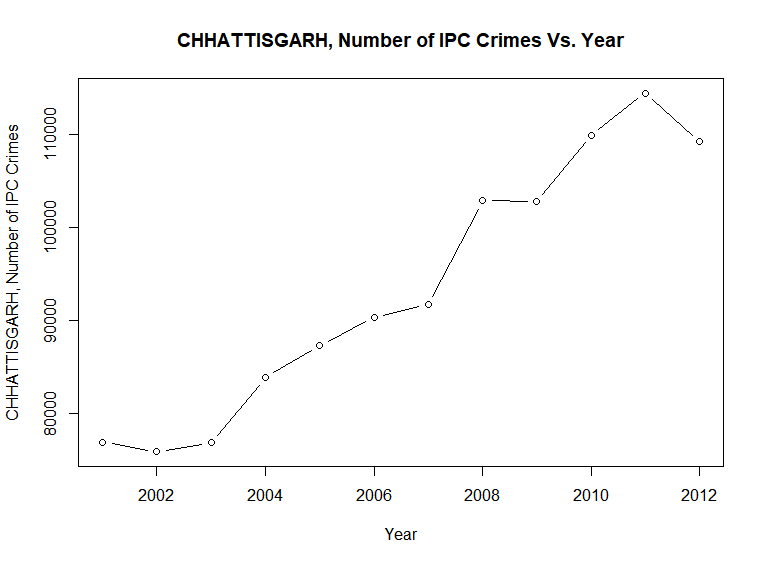


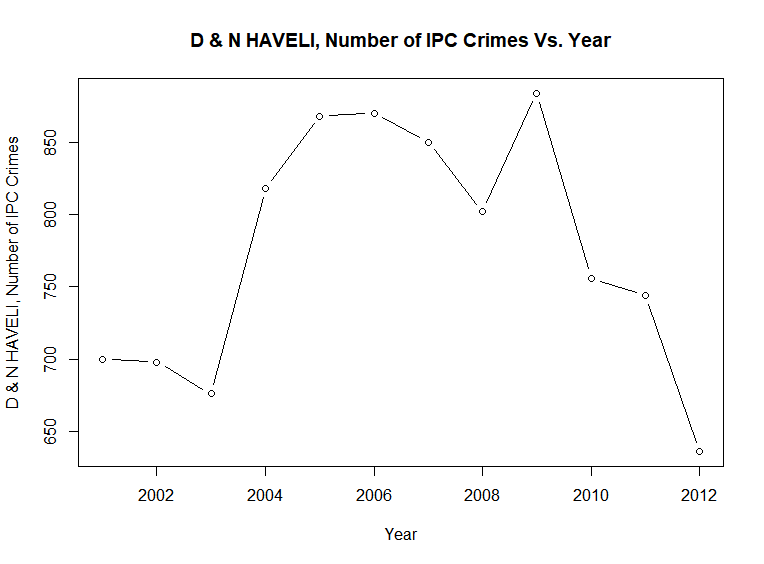


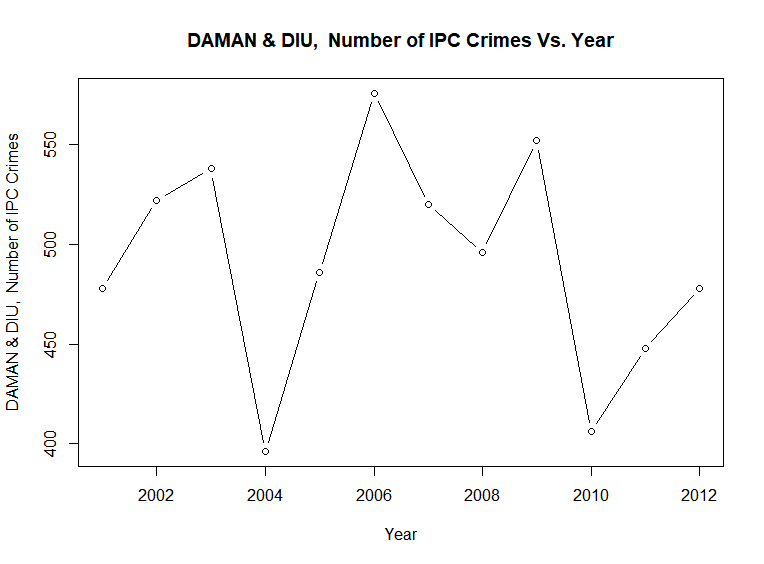


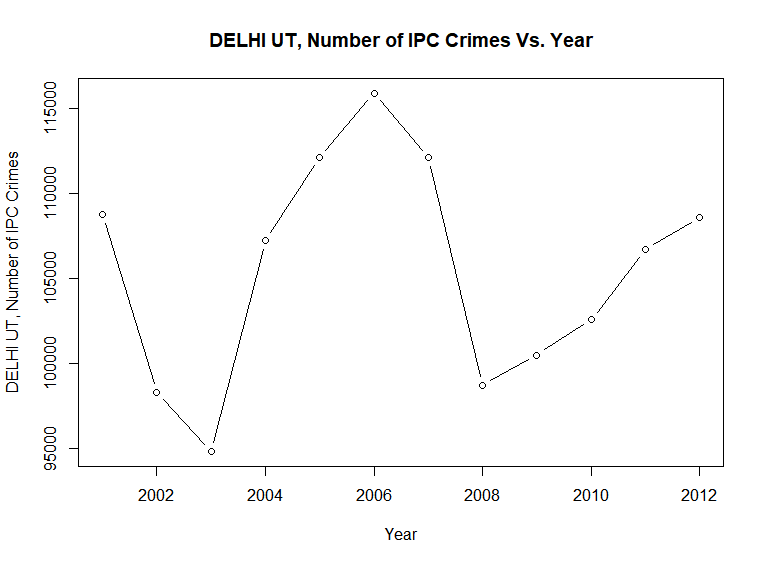


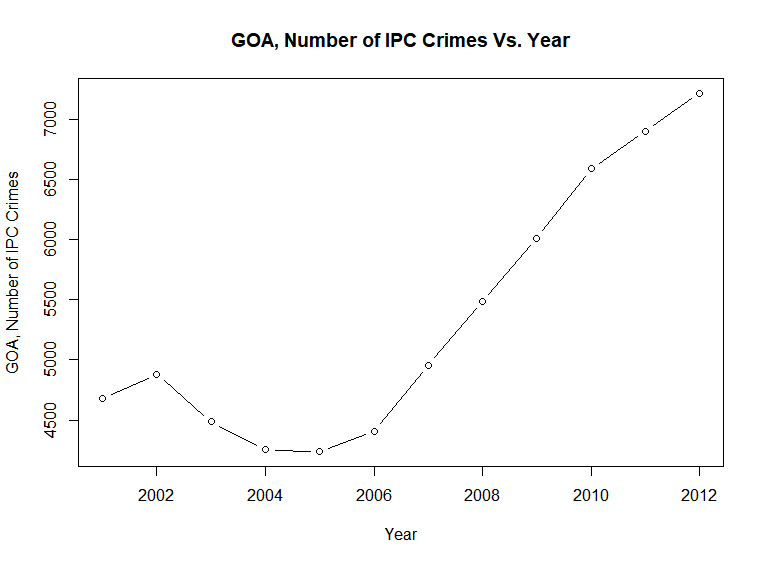


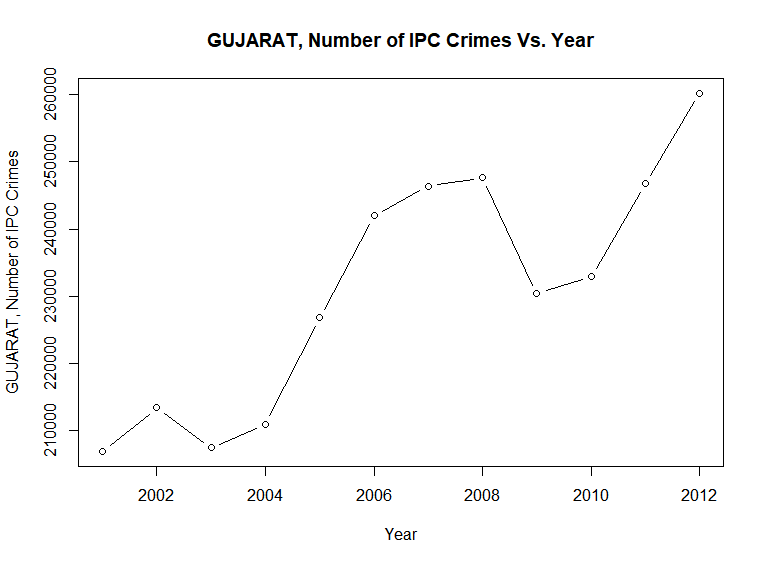


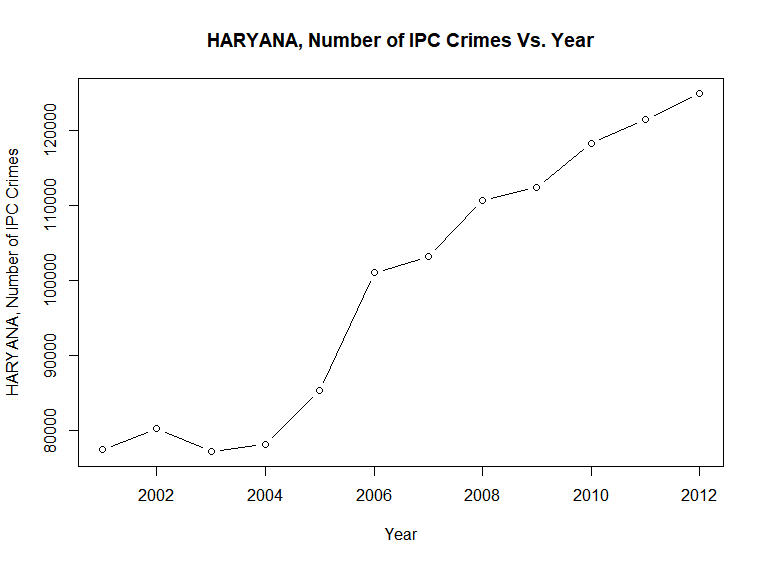


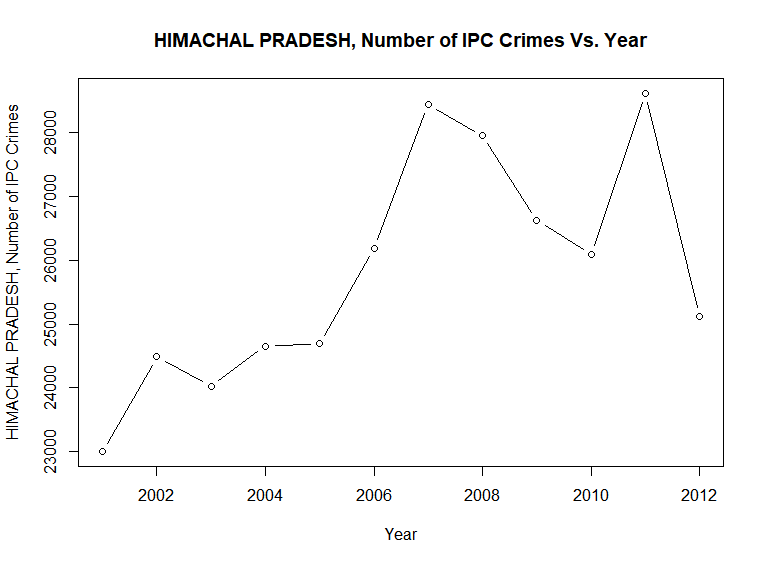


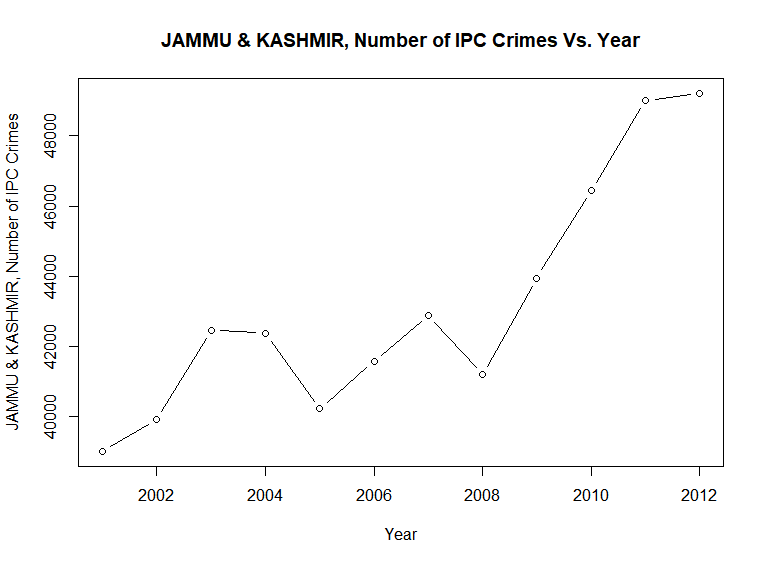


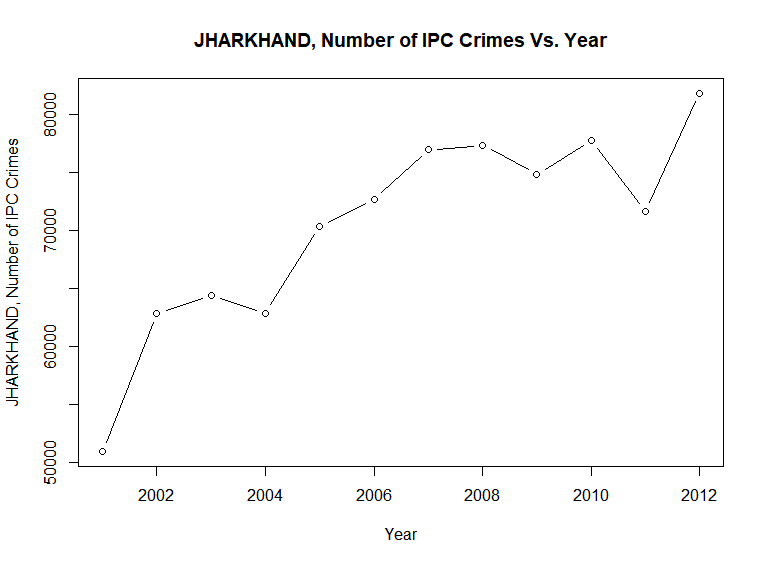


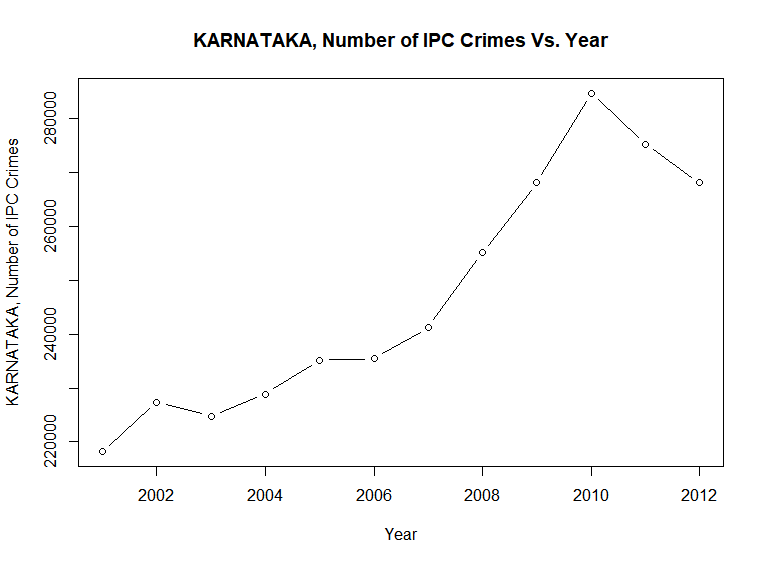


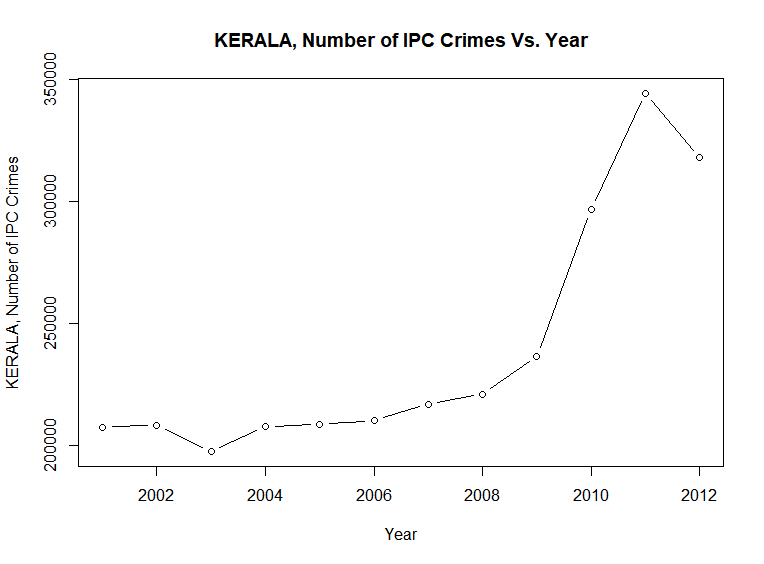


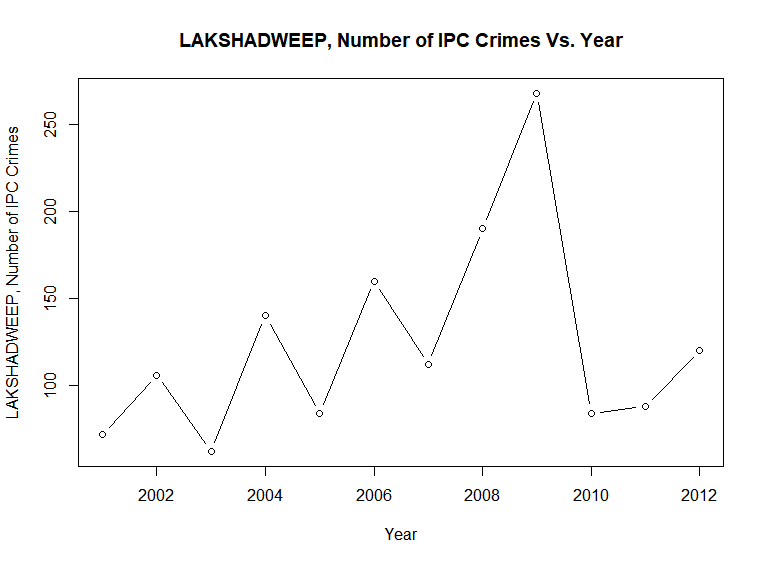


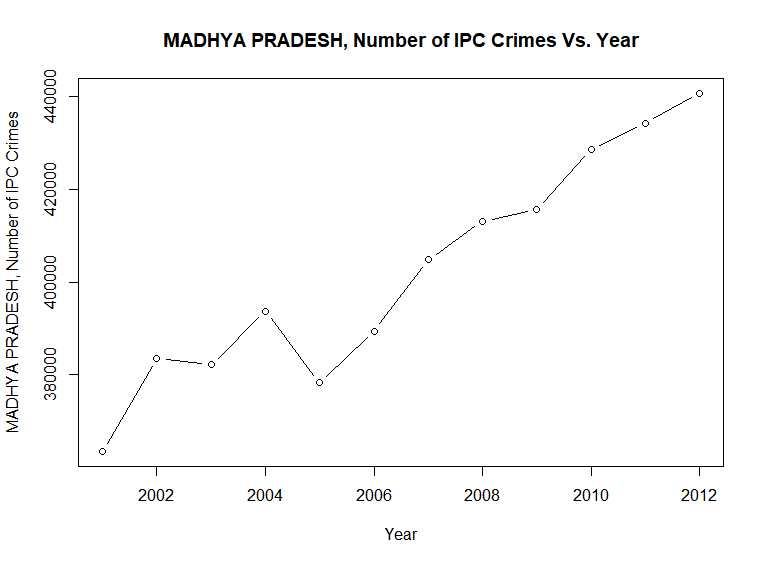


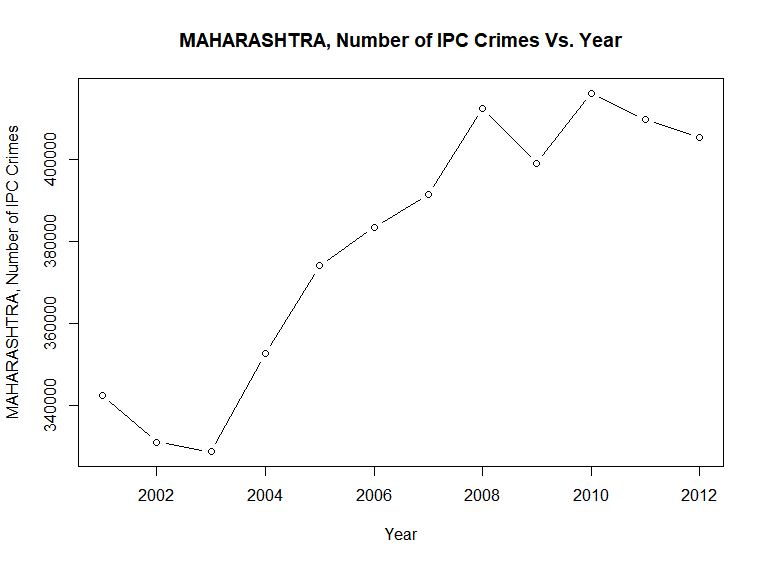


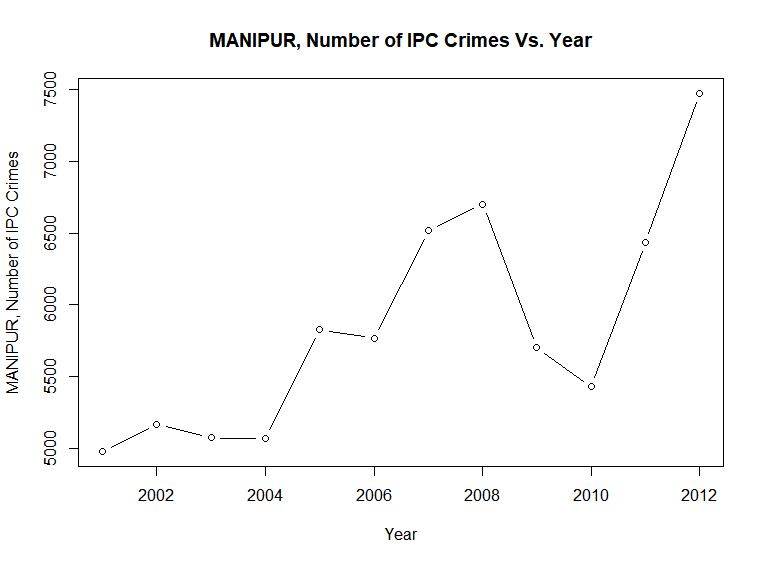


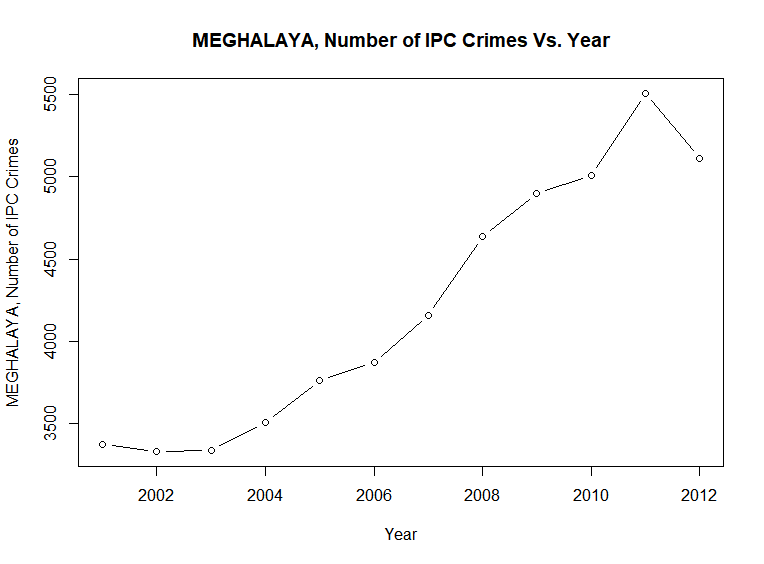


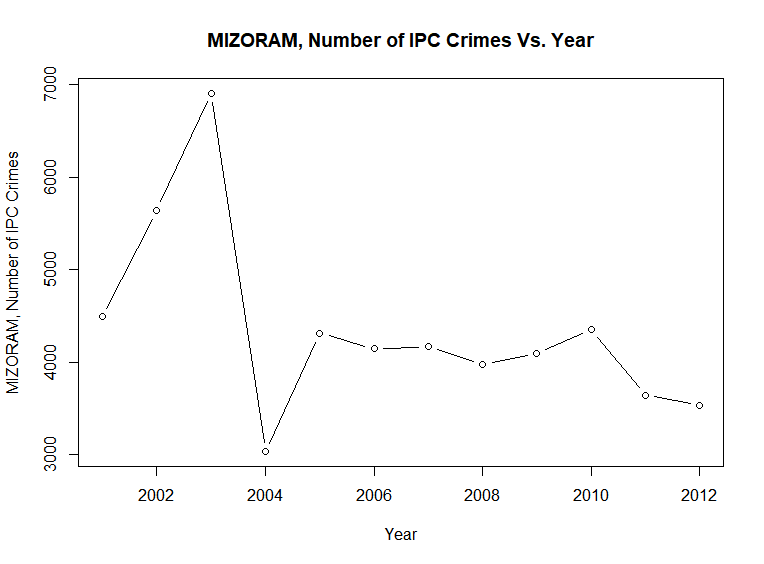


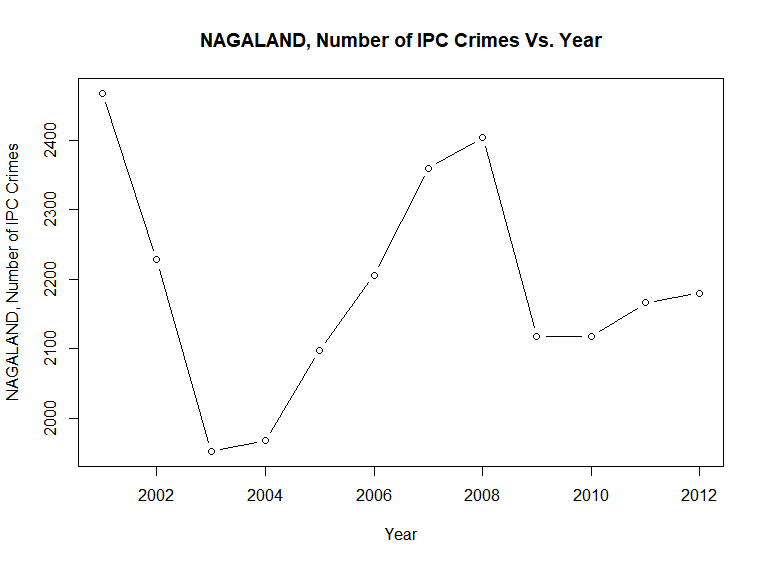


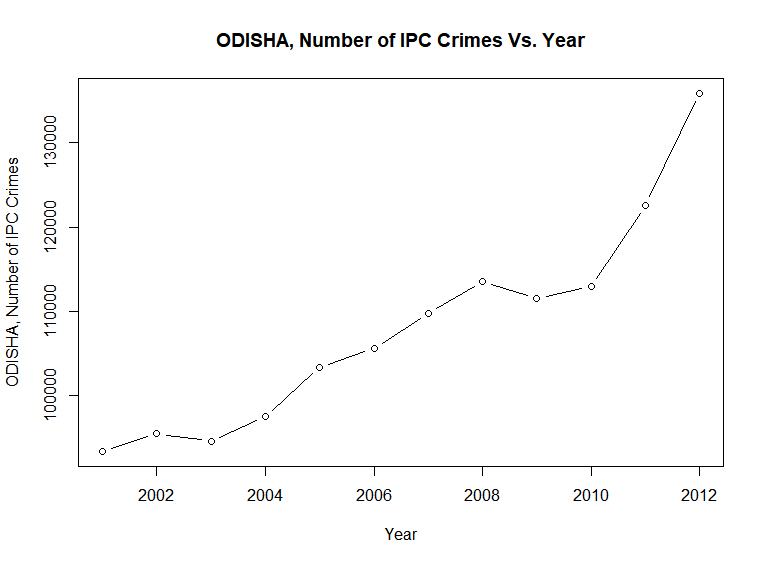


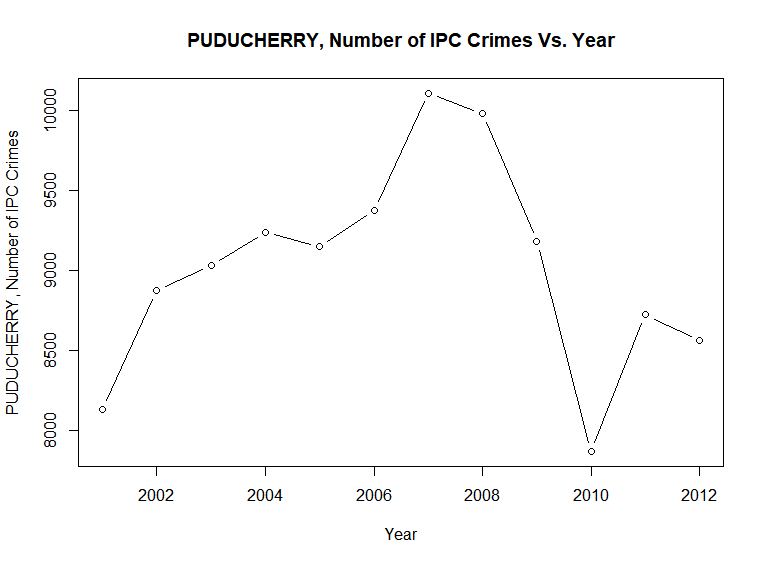


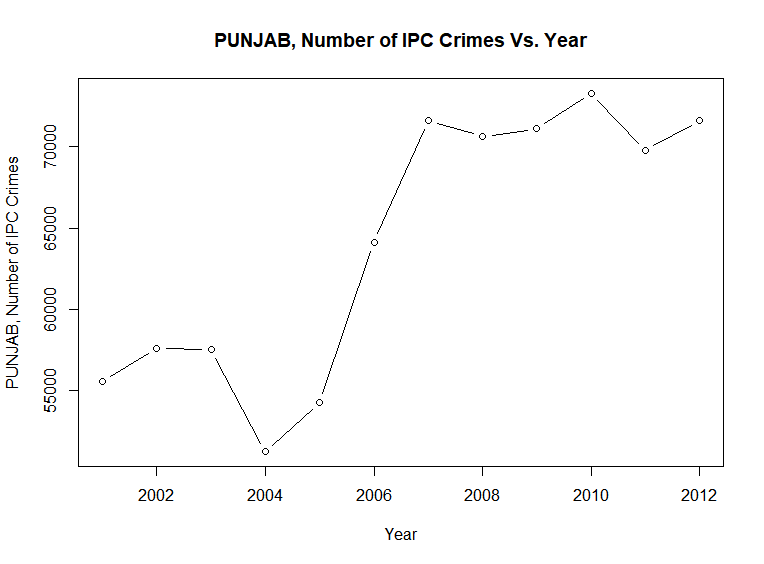


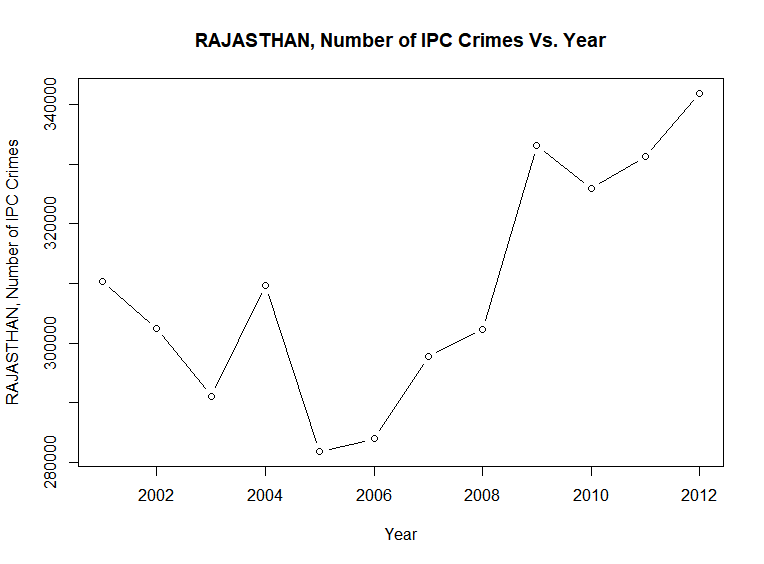


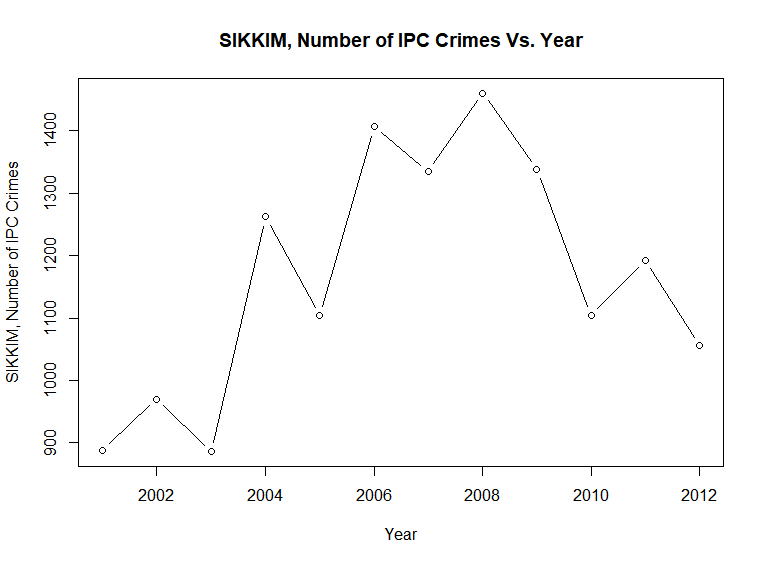


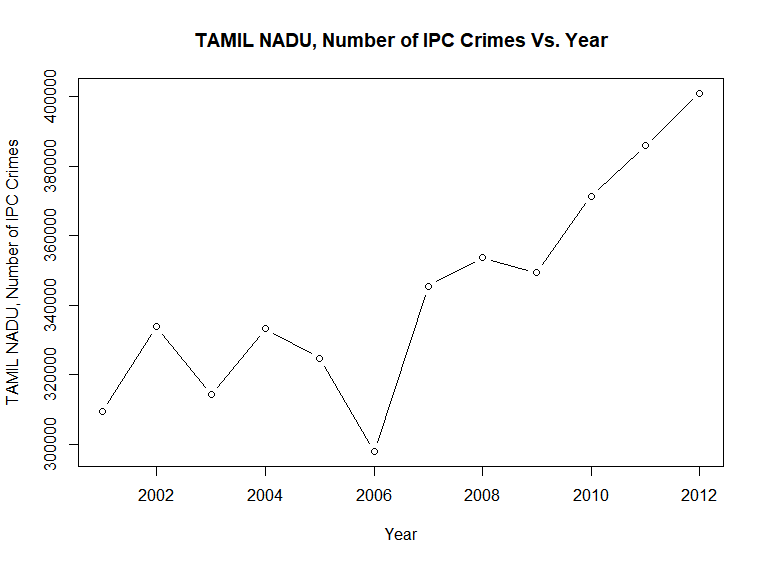


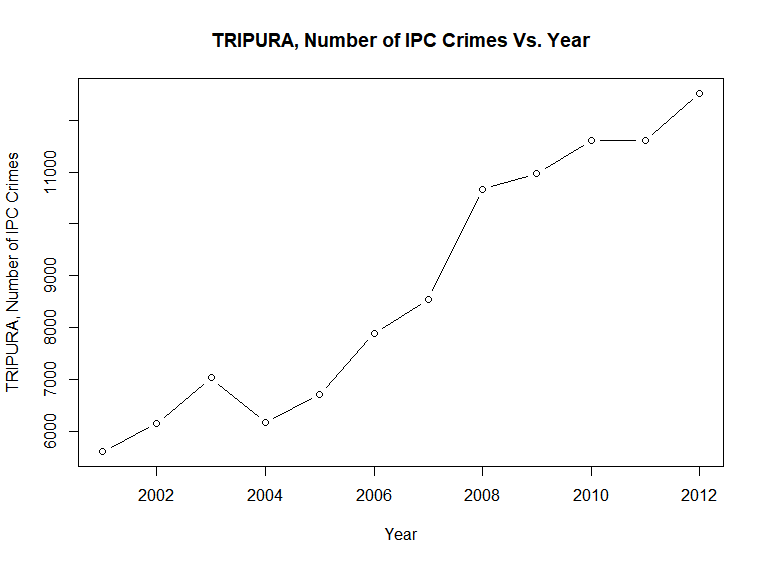


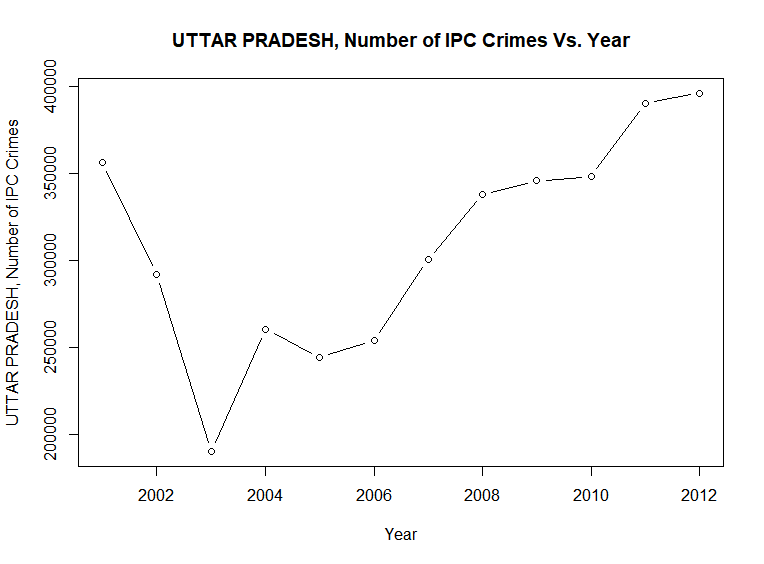


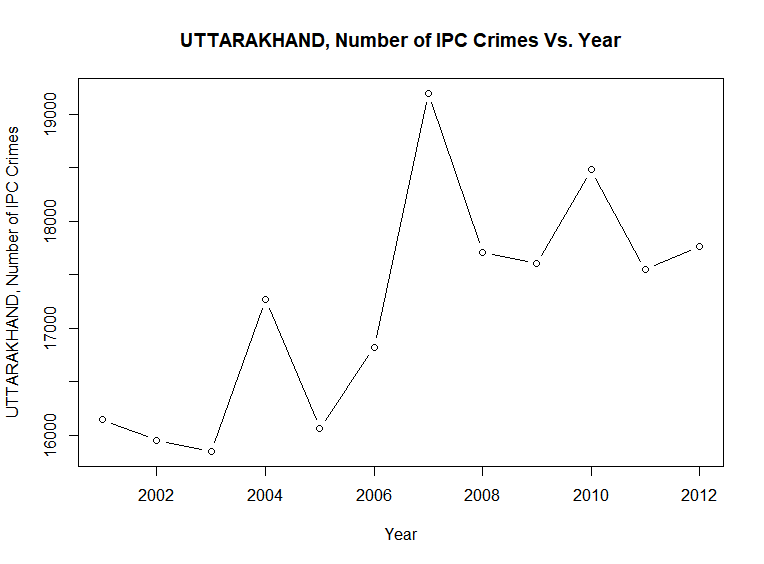


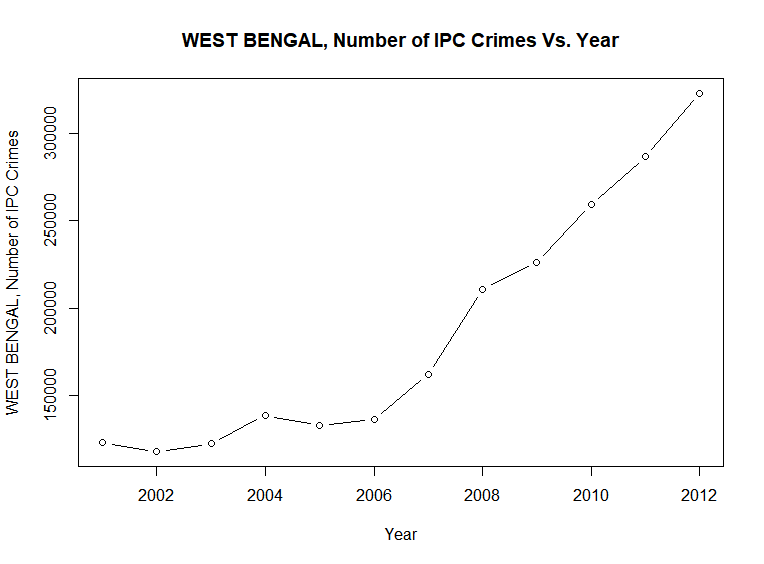












## Bar Plot : Plots of count of specific and total IPC crimes vs. State using DataGroupedBySTATENames.csv derived data set:

**R Code (barplots.R):**

dataset = read.csv('DataGroupedBySTATENames.csv')

library(ggplot2)

barplot(dataset$MURDER,names.arg = dataset$STATE\_UT, main =paste("Murder cases in period 2001-2012 "), xlab = "State/UT", ylab = "Number of Murder cases", cex.names = 0.7, las = 2, col = "Red", border = "Gold")

barplot(dataset$ATTEMPT\_TO\_MURDER,names.arg = dataset$STATE\_UT, main =paste("ATTEMPT\_TO\_MURDER cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of ATTEMPT\_TO\_MURDER cases", cex.names = 0.7, las = 2, col = rainbow(26))

barplot(dataset$CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER,names.arg = dataset$STATE\_UT, main =paste("CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of CULPABLE\_HOMICIDE\_NOT\_AMOUNTING\_TO\_MURDER cases", cex.names = 0.7, las = 2, col = rainbow(20))

barplot(dataset$RAPE,names.arg = dataset$STATE\_UT, main =paste("RAPE cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of RAPE cases", cex.names = 0.7, las = 2, col = rainbow(24))

barplot(dataset$CUSTODIAL\_RAPE,names.arg = dataset$STATE\_UT, main =paste("CUSTODIAL\_RAPE cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of CUSTODIAL\_RAPE cases", cex.names = 0.7, las = 2, col = rainbow(22))

barplot(dataset$OTHER\_RAPE,names.arg = dataset$STATE\_UT, main =paste("OTHER\_RAPE cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of OTHER\_RAPE cases", cex.names = 0.7, las = 2, col = rainbow(15))

barplot(dataset$KIDNAPPING\_\_\_ABDUCTION,names.arg = dataset$STATE\_UT, main =paste("KIDNAPPING\_\_\_ABDUCTION cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number KIDNAPPING\_\_\_ABDUCTION of cases", cex.names = 0.7, las = 2, col = rainbow(25))

barplot(dataset$KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS,names.arg = dataset$STATE\_UT, main =paste("KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of KIDNAPPING\_AND\_ABDUCTION\_OF\_WOMEN\_AND\_GIRLS cases", cex.names = 0.4, las = 2, col = "blue", border = "black")

barplot(dataset$KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS,names.arg = dataset$STATE\_UT, main =paste("KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of KIDNAPPING\_AND\_ABDUCTION\_OF\_OTHERS cases", cex.names = 0.7, las = 2, col = "Purple", border = "Gold")

barplot(dataset$DACOITY,names.arg = dataset$STATE\_UT, main =paste("DACOITY cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of DACOITY cases", cex.names = 0.7, las = 2, col = rainbow(13))

barplot(dataset$PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY,names.arg = dataset$STATE\_UT, main =paste("PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of PREPARATION\_AND\_ASSEMBLY\_FOR\_DACOITY cases", cex.names = 0.7, las = 2, col = "orange")

barplot(dataset$ROBBERY,names.arg = dataset$STATE\_UT, main =paste("ROBBERY cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number ROBBERY of cases", cex.names = 0.7, las = 2, col = "maroon", border = "Gold")

barplot(dataset$BURGLARY,names.arg = dataset$STATE\_UT, main =paste("BURGLARY cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of cases", cex.names = 0.7, las = 2, col = "Dark Green", border = "Gold")

barplot(dataset$THEFT,names.arg = dataset$STATE\_UT, main =paste("THEFT cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of THEFT cases", cex.names = 0.7, las = 2, col = "yellow", border = "black")

barplot(dataset$AUTO\_THEFT,names.arg = dataset$STATE\_UT, main =paste("AUTO\_THEFT cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of AUTO\_THEFT cases", cex.names = 0.7, las = 2, col = rainbow(29))

barplot(dataset$OTHER\_THEFT,names.arg = dataset$STATE\_UT, main =paste("OTHER\_THEFT cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of OTHER\_THEFT cases", cex.names = 0.7, las = 2, col = rainbow(28))

barplot(dataset$RIOTS,names.arg = dataset$STATE\_UT, main =paste("RIOTS cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of RIOTS cases", cex.names = 0.7, las = 2, col = rainbow(27))

barplot(dataset$CRIMINAL\_BREACH\_OF\_TRUST,names.arg = dataset$STATE\_UT, main =paste("CRIMINAL\_BREACH\_OF\_TRUST cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of CRIMINAL\_BREACH\_OF\_TRUST cases", cex.names = 0.7, las = 2, col = rainbow(26))

barplot(dataset$CHEATING,names.arg = dataset$STATE\_UT, main =paste("CHEATING cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of CHEATING cases", cex.names = 0.7, las = 2, col = rainbow(24))

barplot(dataset$COUNTERFIETING,names.arg = dataset$STATE\_UT, main =paste("COUNTERFIETING cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of COUNTERFIETING cases", cex.names = 0.7, las = 2, col = rainbow(25))

barplot(dataset$ARSON,names.arg = dataset$STATE\_UT, main =paste("ARSON cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of ARSON cases", cex.names = 0.7, las = 2, col = rainbow(29))

barplot(dataset$HURT\_GREVIOUS\_HURT,names.arg = dataset$STATE\_UT, main =paste("HURT\_GREVIOUS\_HURT cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of HURT\_GREVIOUS\_HURT cases", cex.names = 0.7, las = 2, col = rainbow(23))

barplot(dataset$DOWRY\_DEATHS,names.arg = dataset$STATE\_UT, main =paste("DOWRY\_DEATHS cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of DOWRY\_DEATHS cases", cex.names = 0.7, las = 2, col = rainbow(22))

barplot(dataset$ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY,names.arg = dataset$STATE\_UT, main =paste("ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of ASSAULT\_ON\_WOMEN\_WITH\_INTENT\_TO\_OUTRAGE\_HER\_MODESTY cases", cex.names = 0.7, las = 2, col = rainbow(21))

barplot(dataset$INSULT\_TO\_MODESTY\_OF\_WOMEN,names.arg = dataset$STATE\_UT, main =paste("INSULT\_TO\_MODESTY\_OF\_WOMEN cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of INSULT\_TO\_MODESTY\_OF\_WOMEN cases", cex.names = 0.7, las = 2, col = rainbow(20))

barplot(dataset$CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES,names.arg = dataset$STATE\_UT, main =paste("CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of CRUELTY\_BY\_HUSBAND\_OR\_HIS\_RELATIVES cases", cex.names = 0.7, las = 2, col = rainbow(19))

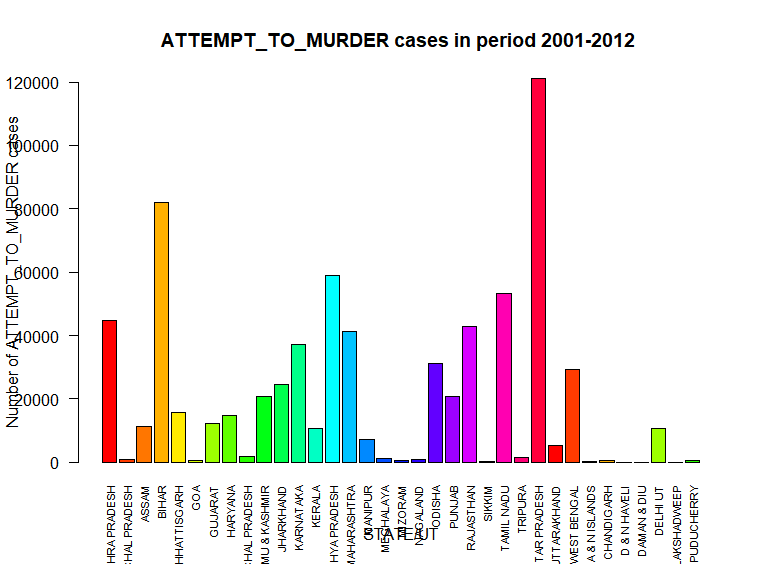
barplot(dataset$IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES,names.arg = dataset$STATE\_UT, main =paste("IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of IMPORTATION\_OF\_GIRLS\_FROM\_FOREIGN\_COUNTRIES cases", cex.names = 0.7, las = 2, col = rainbow(18))

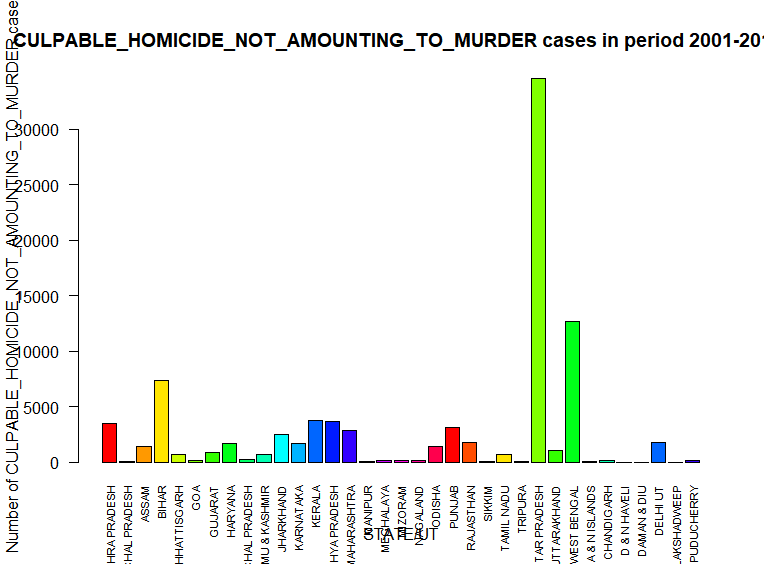
barplot(dataset$CAUSING\_DEATH\_BY\_NEGLIGENCE,names.arg = dataset$STATE\_UT, main =paste("CAUSING\_DEATH\_BY\_NEGLIGENCE cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of CAUSING\_DEATH\_BY\_NEGLIGENCE cases", cex.names = 0.7, las = 2, col = rainbow(17))

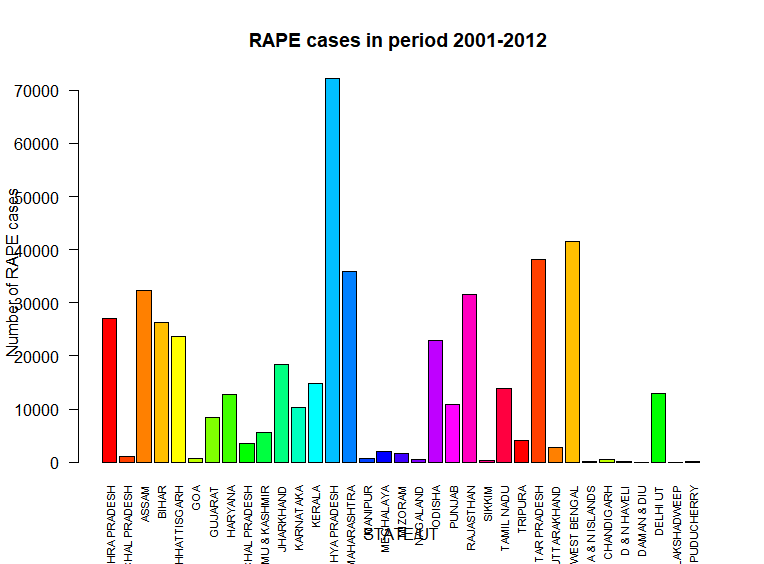
barplot(dataset$OTHER\_IPC\_CRIMES,names.arg = dataset$STATE\_UT, main =paste("OTHER\_IPC\_CRIMES cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of OTHER\_IPC\_CRIMES cases", cex.names = 0.7, las = 2, col = rainbow(16))

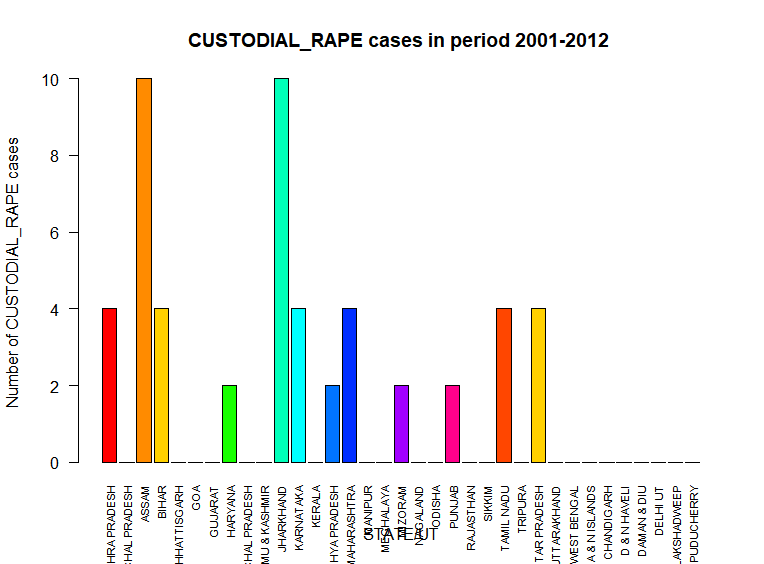
barplot(dataset$TOTAL\_IPC\_CRIMES,names.arg = dataset$STATE\_UT, main =paste("TOTAL\_IPC\_CRIMES cases in period 2001-2012 "), xlab = "STATE/UT", ylab = "Number of TOTAL\_IPC\_CRIMES cases", cex.names = 0.7, las = 2, col = rainbow(15))

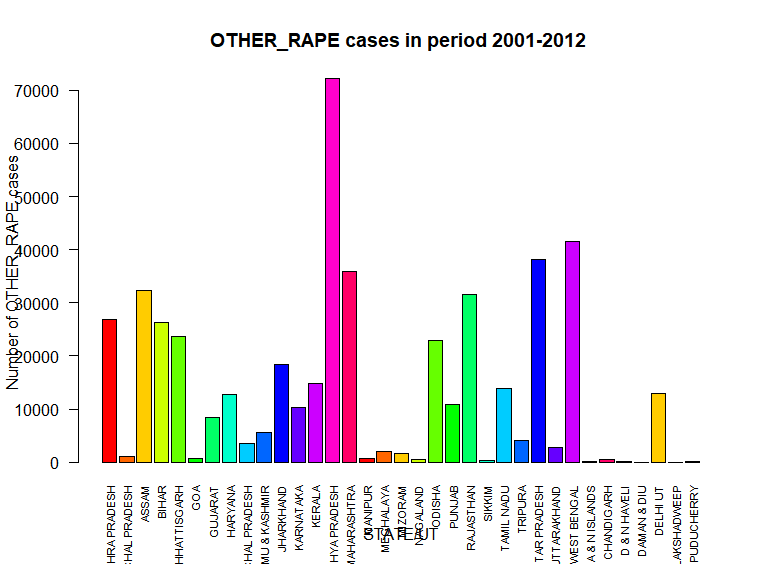


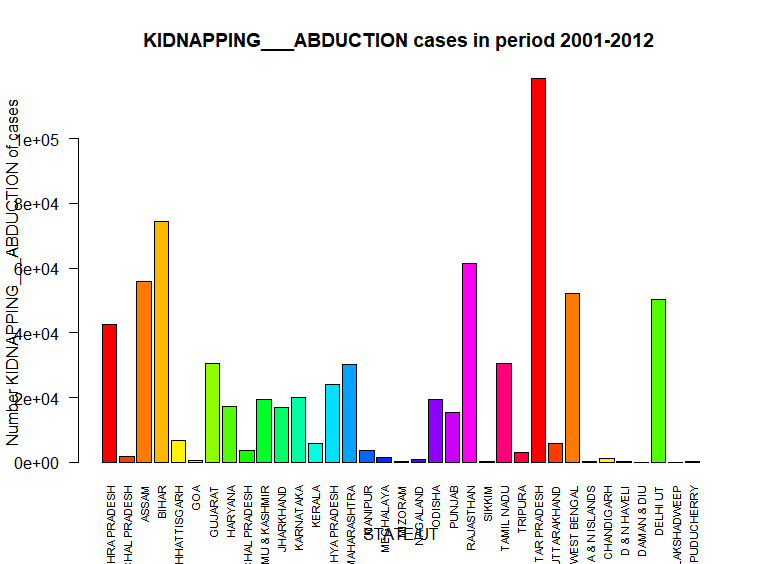


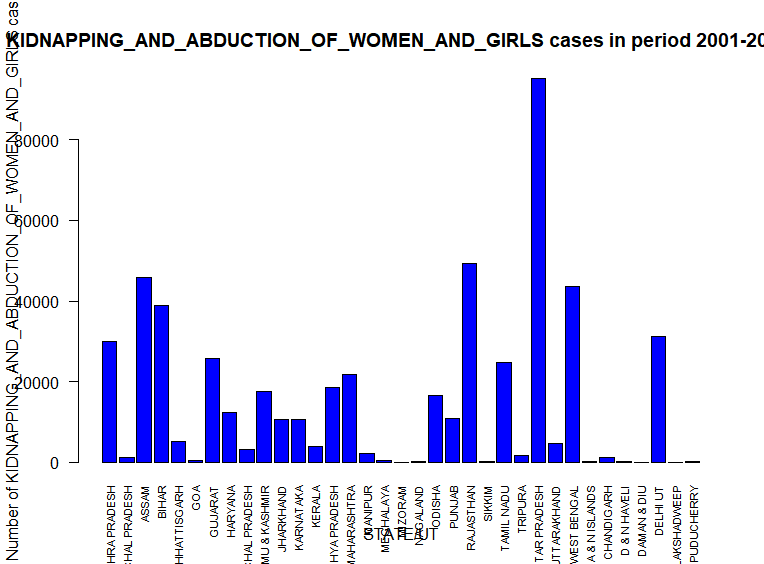


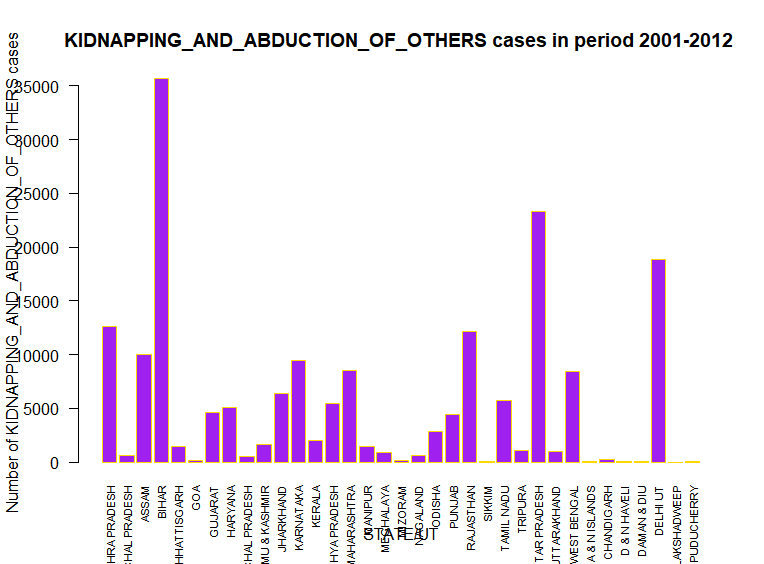


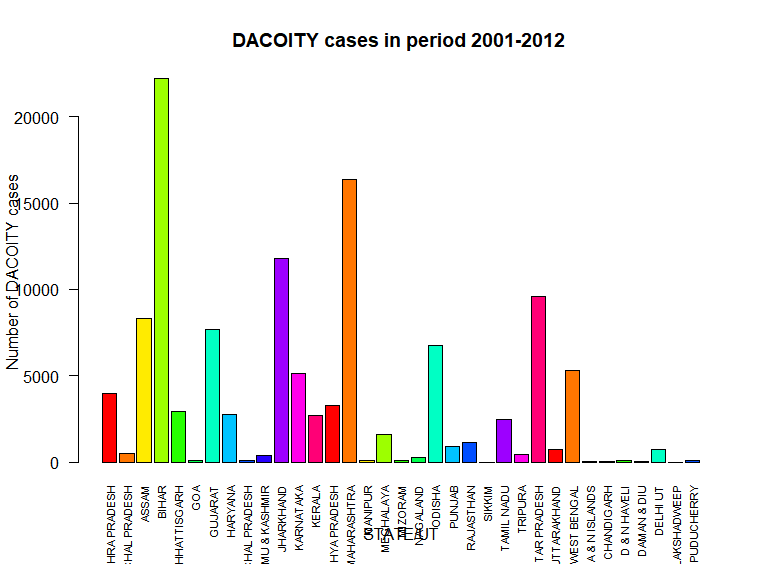


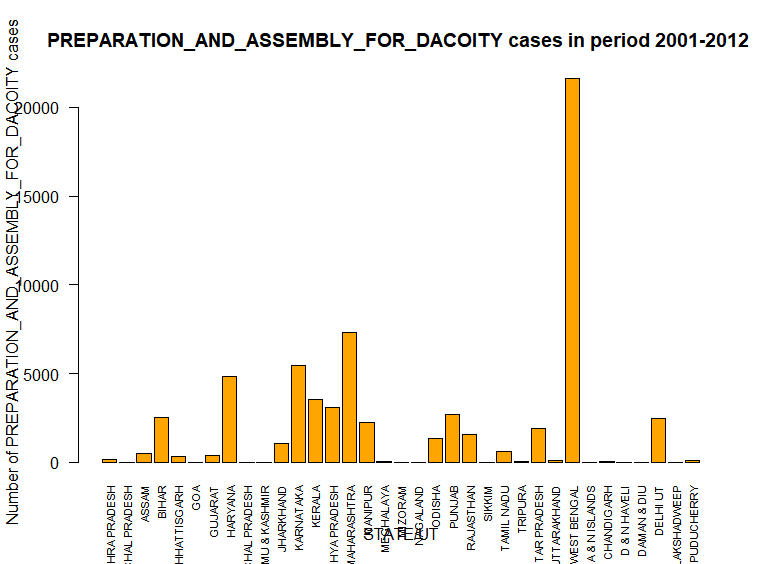


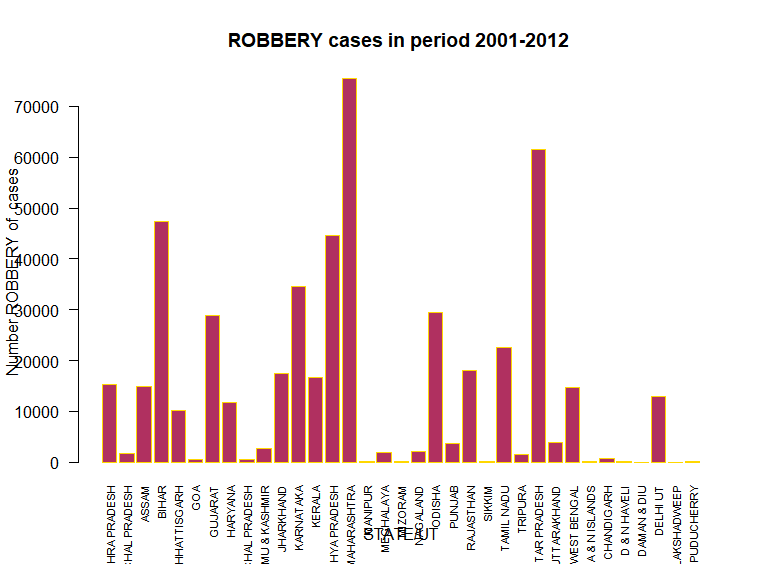


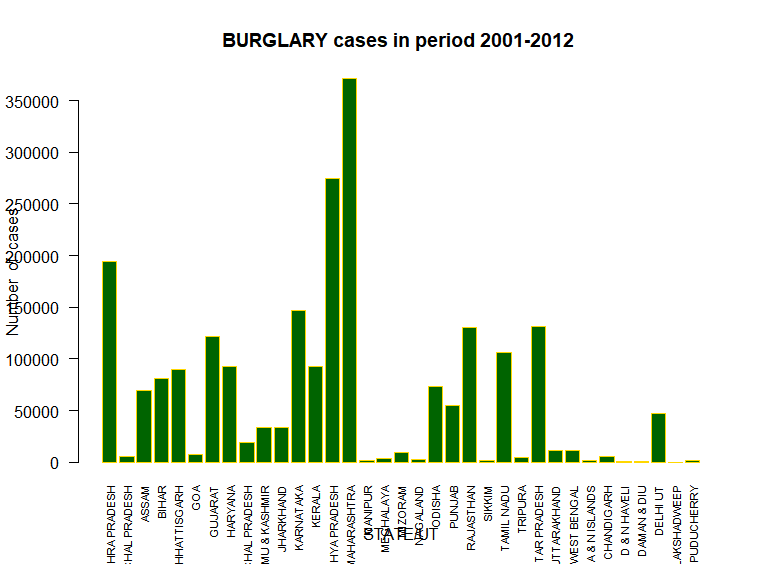


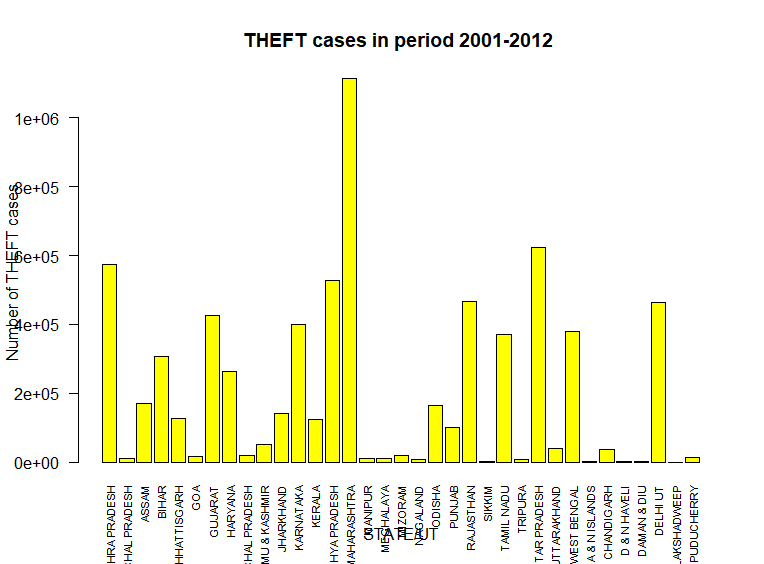


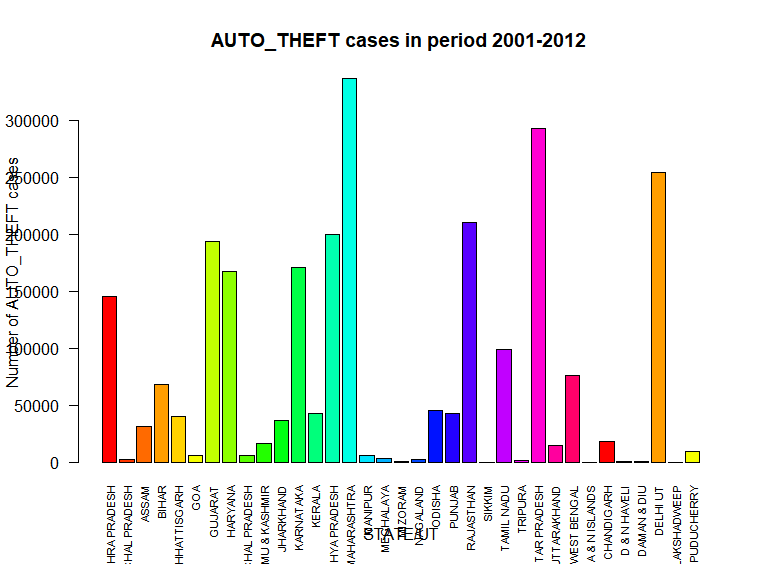


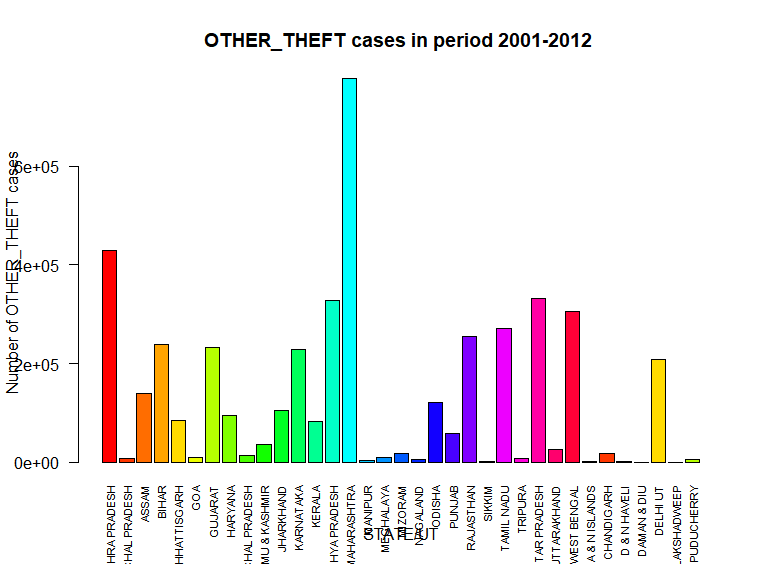


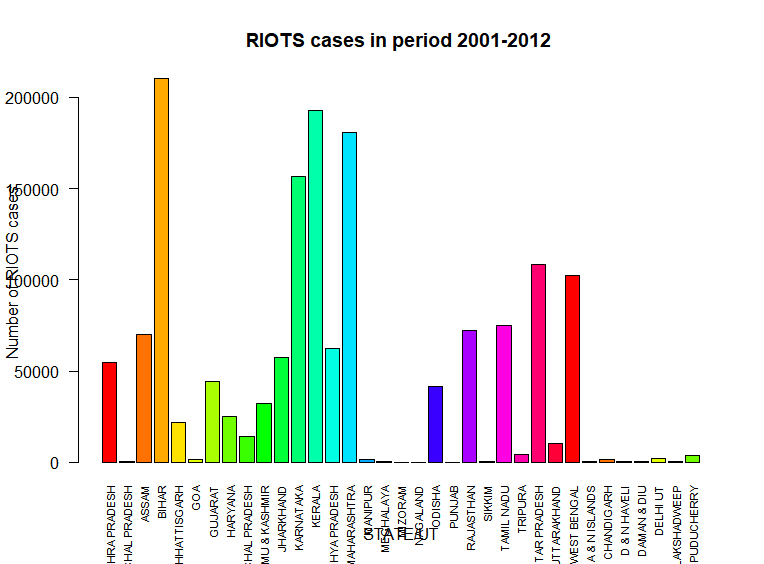


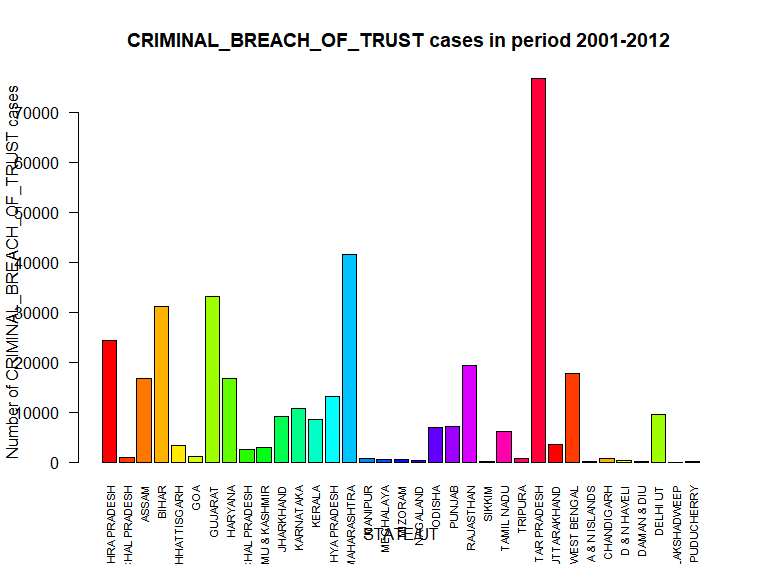


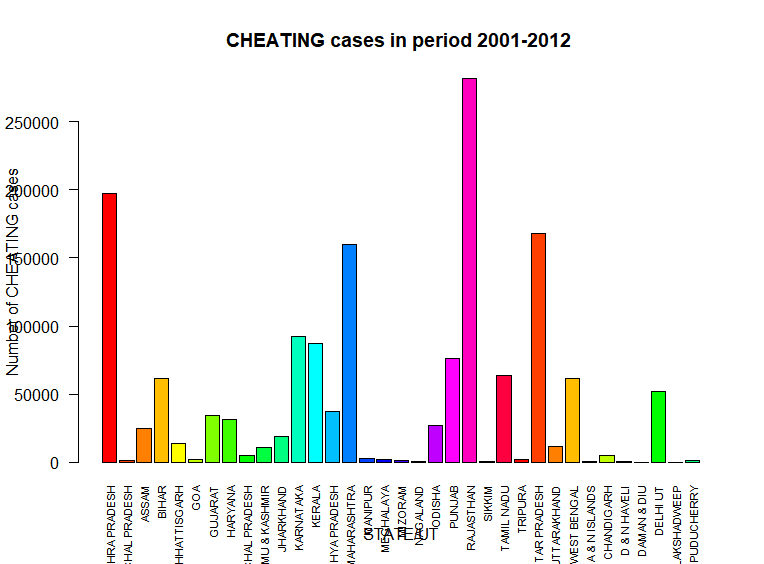


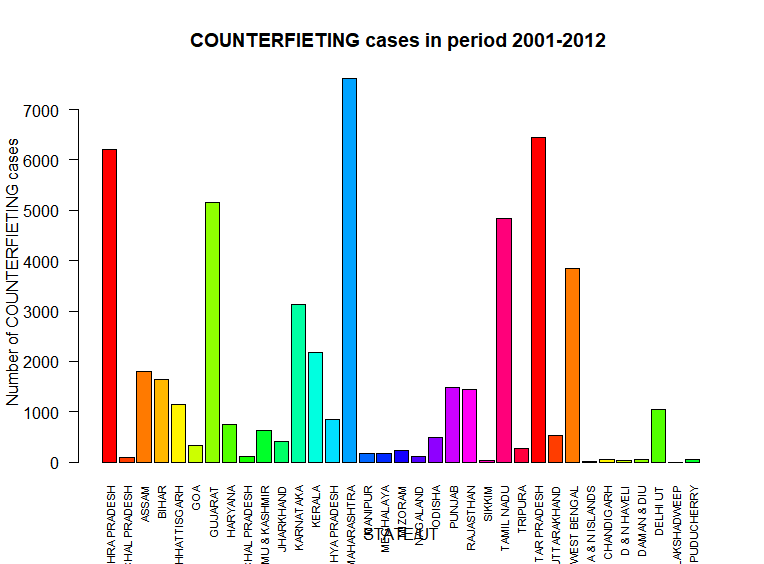


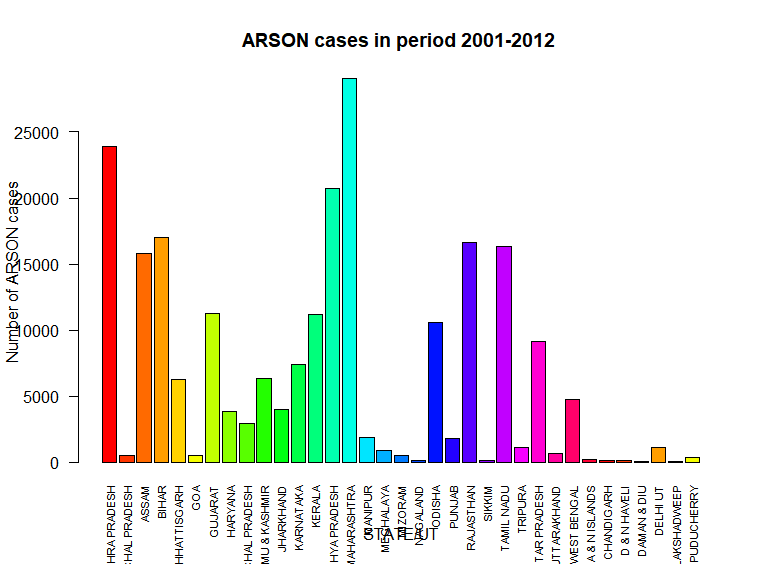


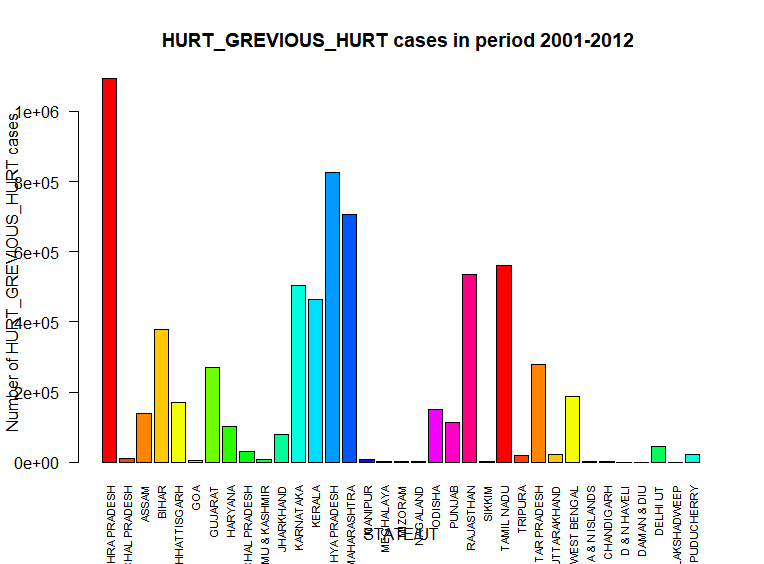


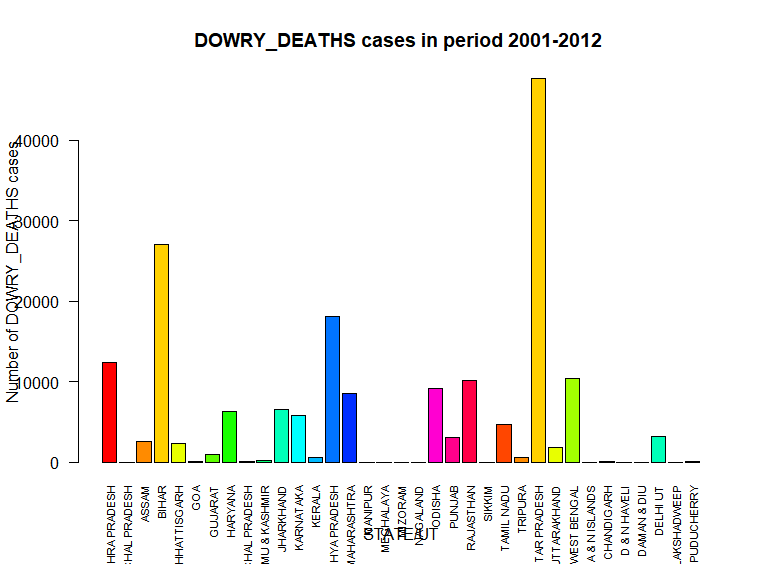


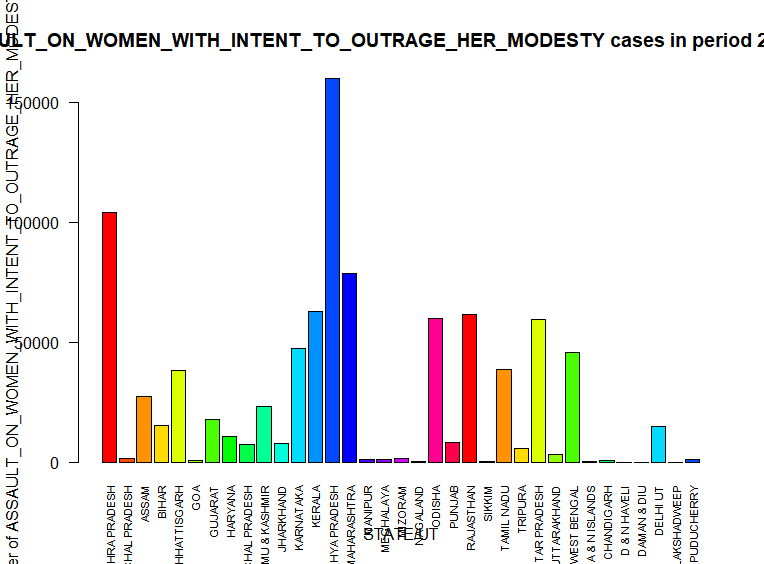


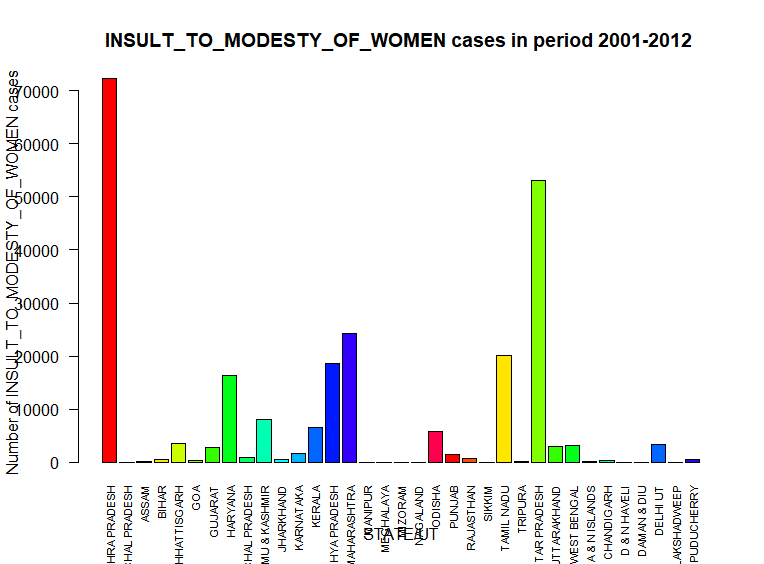


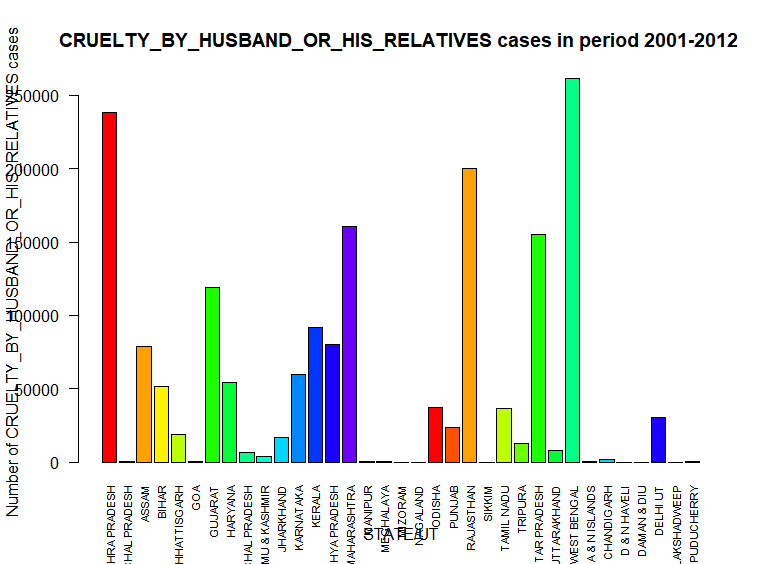


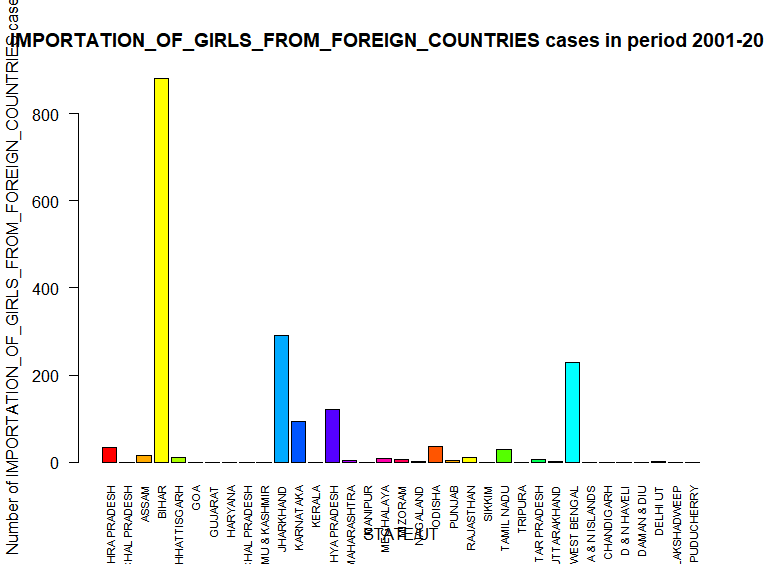


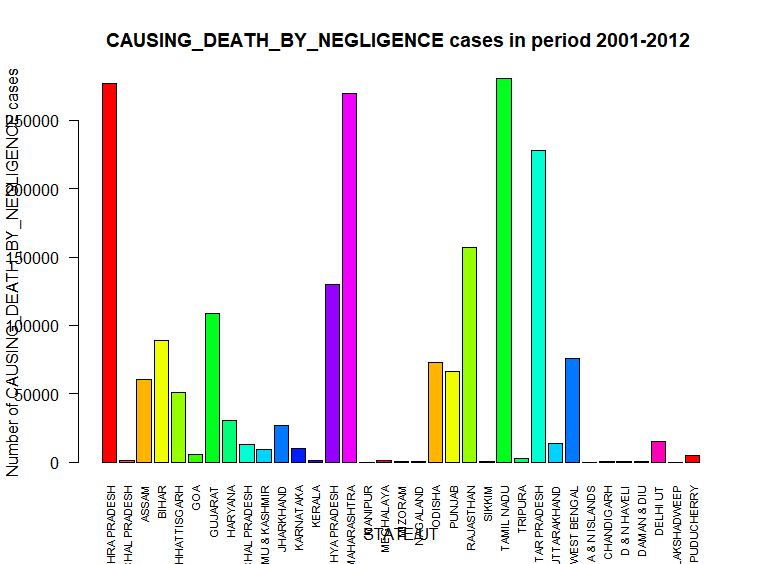


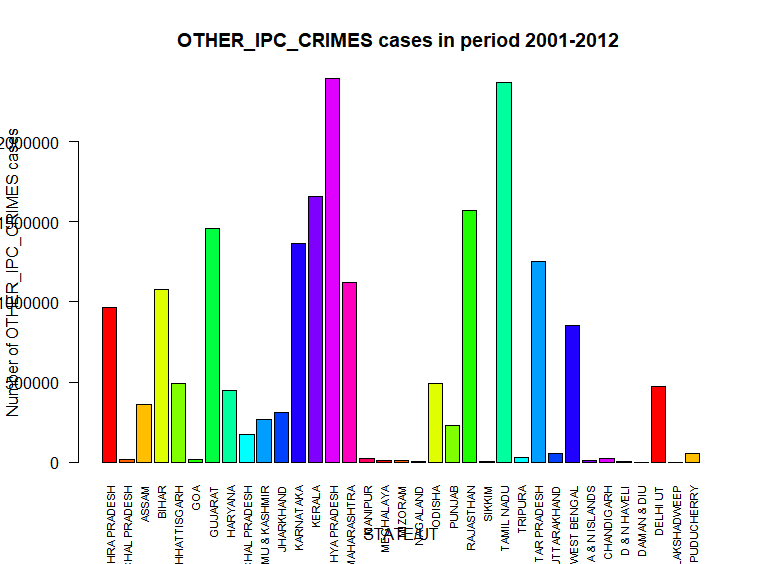


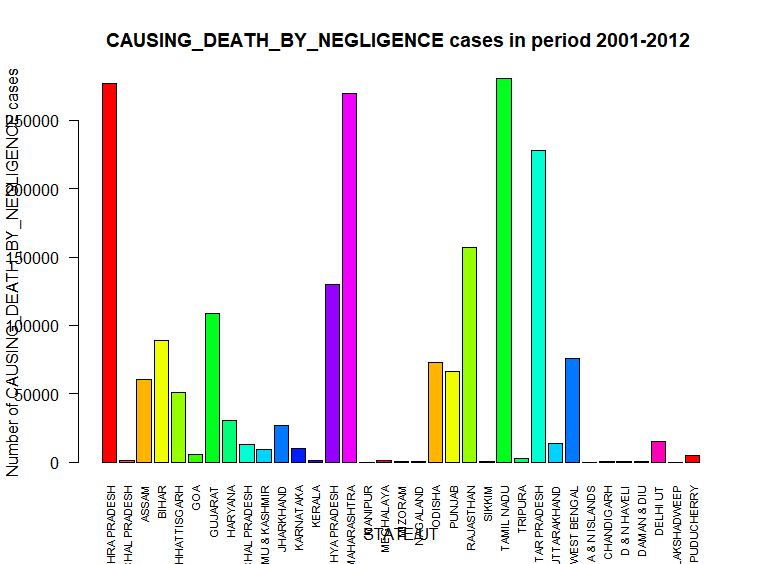


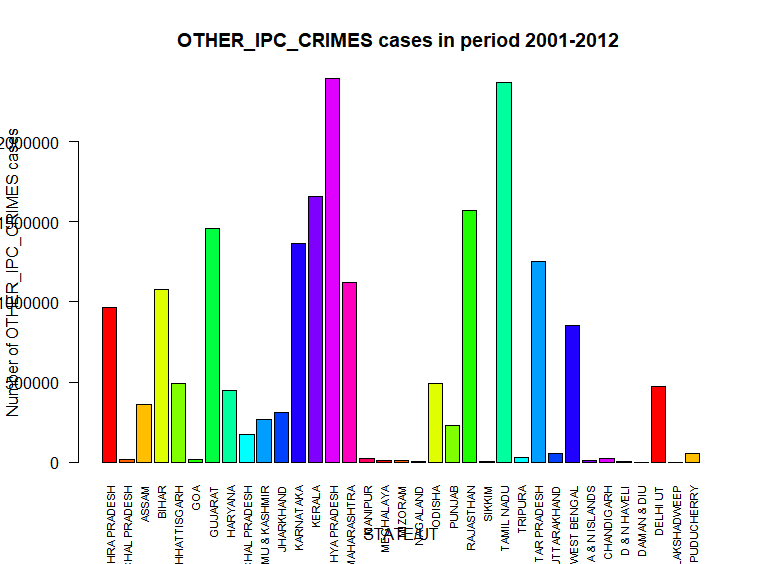


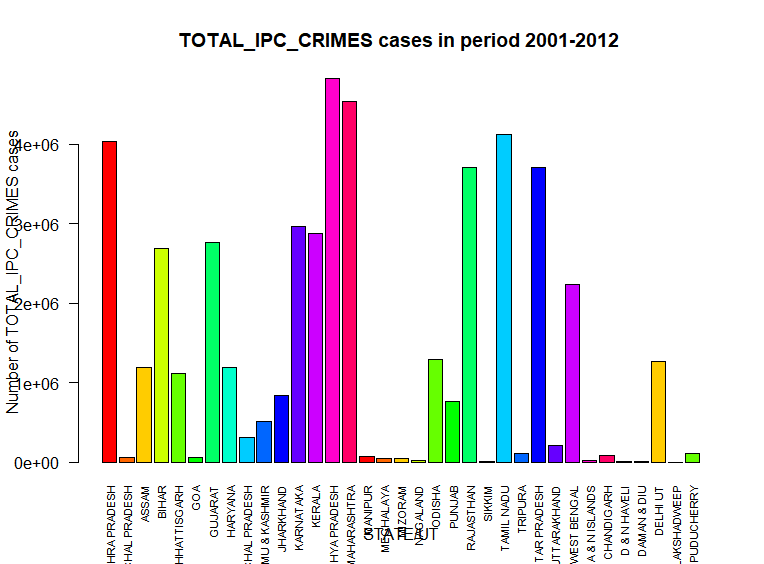












## Applying Clustering Algorithms On The Data sets:

**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**

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.

1. **Data set:** 01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv
2. **Applying k means clustering:**

**Step wise implementation:**

1. **Importing the data set:**

dataset = read.csv("01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv")

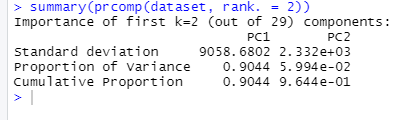
1. **Applying the previously mentioned “Taking care of missing values” template on the dataset**
2. **Choosing the variables on which we want to apply clustering (We excluded the TOTAL\_IPC\_CRIMES variable as it is sum what the sum of other 29 variables):**

dataset = dataset[,c(4:32)]

1. **Checking the % variability explained by first 2 principal components:**

summary(prcomp(dataset, rank. = 2))

Output:

****

**Analysis:**

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

1. **Applying pca to the our data:**

library(caret)

library(e1071)

# Creating preProcess class object

pca = preProcess(x = dataset, method = 'pca', pcaComp = 2)

#Using pca object to apply pca on our data

dataset = predict(pca, dataset)

1. **Using the elbow method to find the optimal number of solutions:**

We plot the with in cluster sum of squares (wcss) vs. Number of clusters in the cluster and for the point in the graph for which there is the highest difference of wcss from the previous point, the corresponding value of number of clusters is our number of cluster.

set.seed(6)

wcss = vector()

# Using the elbow method to find the optimal number of clusters

for (i in 1:100) wcss[i] = sum(kmeans(dataset, i)$withinss)

plot(1:100,

wcss,

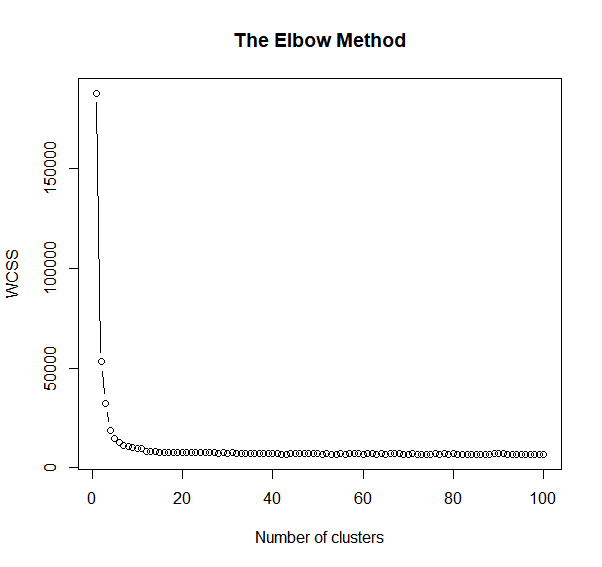
type = 'b',

main = paste('The Elbow Method'),

xlab = 'Number of clusters',

ylab = 'WCSS')

Output:



**Analysis:**

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

1. **Fitting K means to the data set:**

# Fitting K-Means to the dataset

set.seed(29)

kmeans = kmeans(x = dataset, centers = 2)

y\_kmeans = kmeans$cluster #Storing the cluster data for each record in dataset into y\_kmeans

1. **Visualizing the clusters:**

# Visualising the clusters

library(cluster)

clusplot(dataset,

y\_kmeans,

lines = 0,

shade = TRUE,

color = TRUE,

labels = 2,

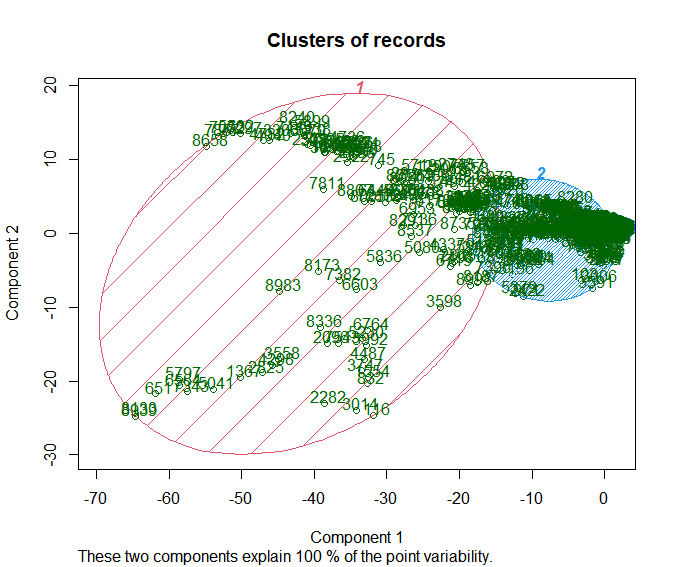
plotchar = FALSE,

span = TRUE,

main = paste('Clusters of records'),

)

Output:



**Analysis:**

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 .
7. **Applying Hierarchical Clustering on the data set:**
8. **Importing the data set:**

dataset = read.csv("01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv")

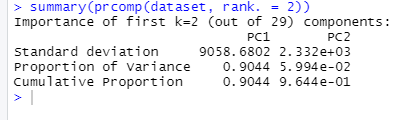
1. **Applying the previously mentioned “Taking care of missing values” template on the dataset**
2. **Choosing the variables on which we want to apply clustering (We excluded the TOTAL\_IPC\_CRIMES variable as it is sum what the sum of other 29 variables):**

dataset = dataset[,c(4:32)]

1. **Checking the % variability explained by first 2 principal components:**

summary(prcomp(dataset, rank. = 2))

Output:

****

**Analysis:**

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

1. **Applying pca to the our data:**

library(caret)

library(e1071)

# Creating preProcess class object

pca = preProcess(x = dataset, method = 'pca', pcaComp = 2)

#Using pca object to apply pca on our data

dataset = predict(pca, dataset)

1. **Using the dendogram to find the optimal level of clusters:**

# Using the dendrogram to find the optimal number of clusters

#Using hclust() method with distance matrix as dissimilarity structure #and agglomerative method ward.D

dendrogram = hclust(d = dist(dataset, method = 'euclidean'), method = 'ward.D')

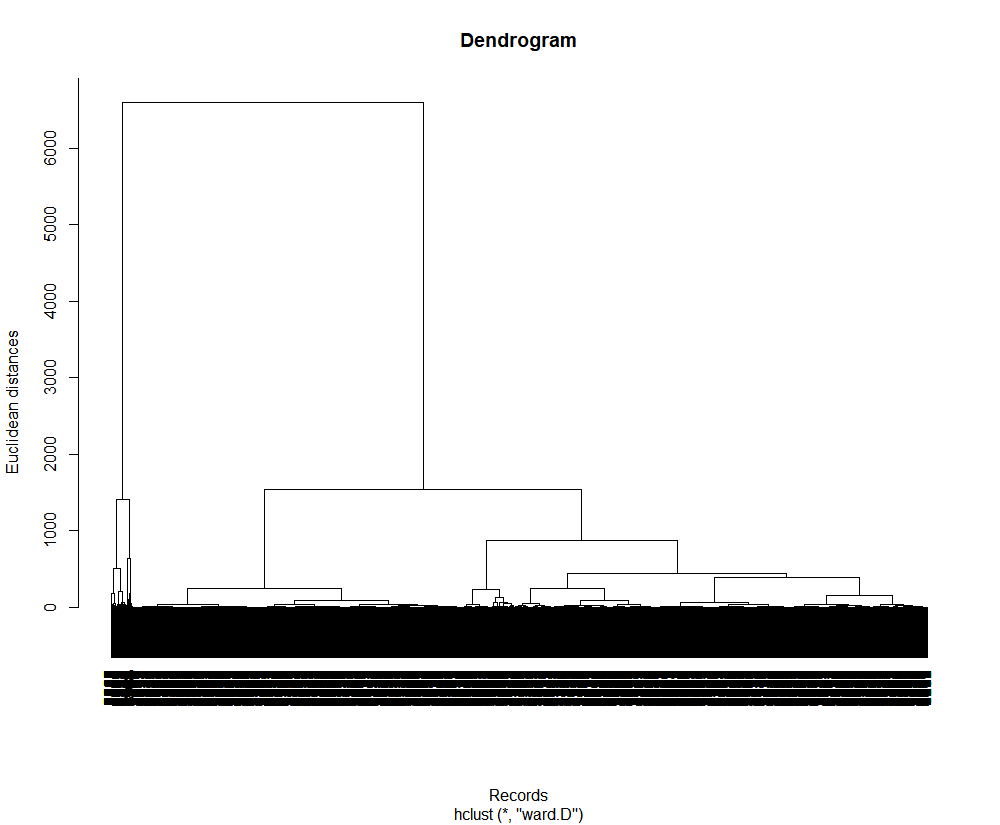
plot(dendrogram,

main = paste('Dendrogram'),

xlab = 'Records',

ylab = 'Euclidean distances')

Output:



**Analysis:**

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 centers will be the optimal number of clusters/centers.

1. **Fitting Hierarchical Clustering to the dataset**

# Fitting Hierarchical Clustering to the dataset

hc = hclust(d = dist(dataset, method = 'euclidean'), method = 'ward.D')

y\_hc = cutree(hc, 2)

1. **Visualizing the clusters**

# Visualising the clusters

library(cluster)

clusplot(dataset,

y\_hc,

lines = 0,

shade = TRUE,

color = TRUE,

labels= 2,

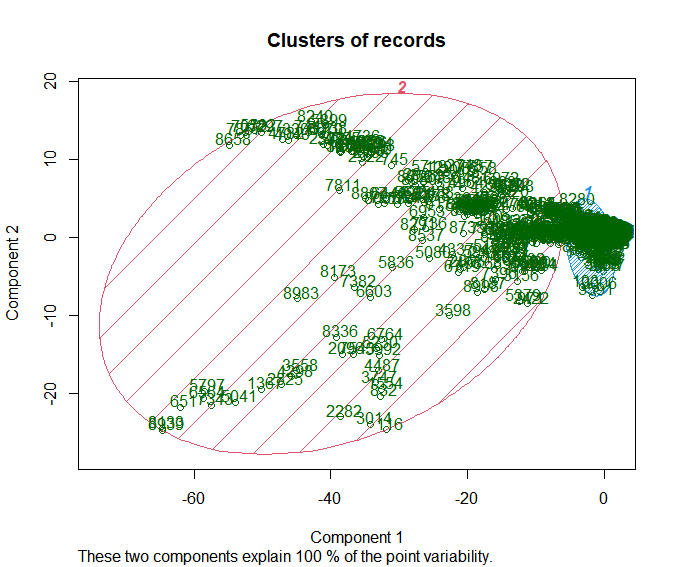
plotchar = FALSE,

span = TRUE,

main = paste('Clusters of records'),

)

Output:



**Analysis:**

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 .
8. **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.

1. **Applying K means clustering:**

**R CODE:**

# K-Means Clustering

# Importing the dataset

dataset = read.csv('DataGroupedBySTATE\_DISTRICT\_Names.csv')

# Feature Scaling

#dataset[,c(3:32)] = scale(dataset[,c(3:32)])

dataset = dataset[,c(3:31)]

summary(prcomp(dataset, rank. = 2))

library(caret)

library(e1071)

pca = preProcess(x = dataset, method = 'pca', pcaComp = 2)

dataset = predict(pca, dataset)

set.seed(6)

wcss = vector()

# Using the elbow method to find the optimal number of clusters

for (i in 1:30) wcss[i] = sum(kmeans(dataset, i)$withinss)

plot(1:30,

wcss,

type = 'b',

main = paste('The Elbow Method'),

xlab = 'Number of clusters',

ylab = 'WCSS')

# Fitting K-Means to the dataset

set.seed(29)

kmeans = kmeans(x = dataset, centers = 2)

y\_kmeans = kmeans$cluster

# Visualising the clusters

library(cluster)

clusplot(dataset,

y\_kmeans,

lines = 0,

shade = TRUE,

color = TRUE,

labels = 2,

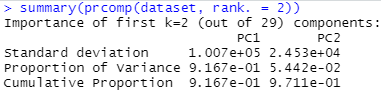
plotchar = FALSE,

span = TRUE,

main = paste('Clusters of records'),

)

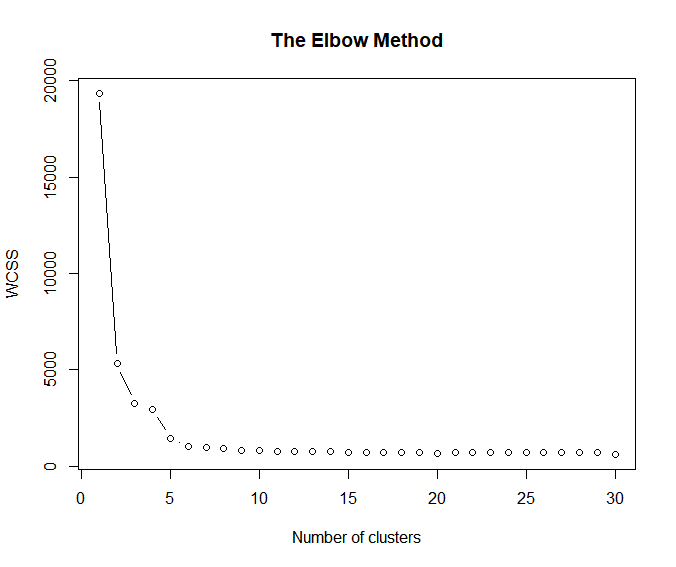
**Ouput:**

****

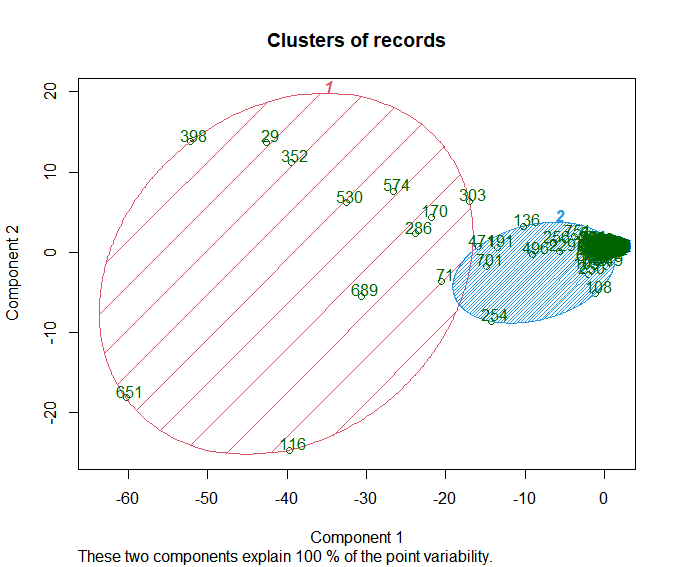
**Analysis**

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.



**Analysis:**

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.
6. **Applying hierarchical clustering:**

**R Code:**

# Hierarchical Clustering

# Importing the dataset

dataset = read.csv('DataGroupedBySTATE\_DISTRICT\_Names.csv')

# Feature Scaling

#dataset[,c(3,31)] = scale(dataset[,c(3,31)])

summary(prcomp(dataset, rank. = 2))

library(caret)

library(e1071)

pca = preProcess(x = dataset, method = 'pca', pcaComp = 2)

dataset = predict(pca, dataset)

dataset = dataset[,c(3:32)]

# Using the dendrogram to find the optimal number of clusters

dendrogram = hclust(d = dist(dataset, method = 'euclidean'), method = 'ward.D')

plot(dendrogram,

main = paste('Dendrogram'),

xlab = 'Records',

ylab = 'Euclidean distances')

# Fitting Hierarchical Clustering to the dataset

hc = hclust(d = dist(dataset, method = 'euclidean'), method = 'ward.D')

y\_hc = cutree(hc, 2)

# Visualising the clusters

library(cluster)

clusplot(dataset,

y\_hc,

lines = 0,

shade = TRUE,

color = TRUE,

labels= 2,

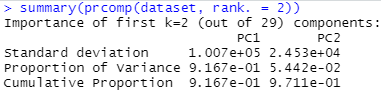
plotchar = FALSE,

span = TRUE,

main = paste('Clusters of records'),

)

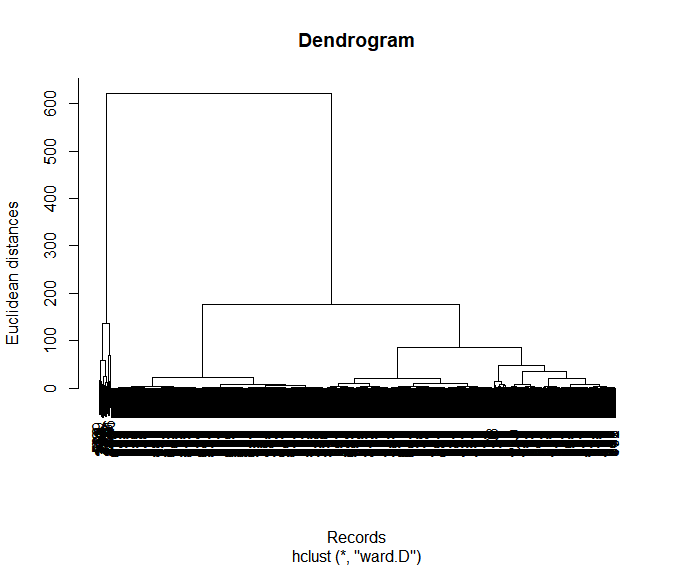
**Output:**

****

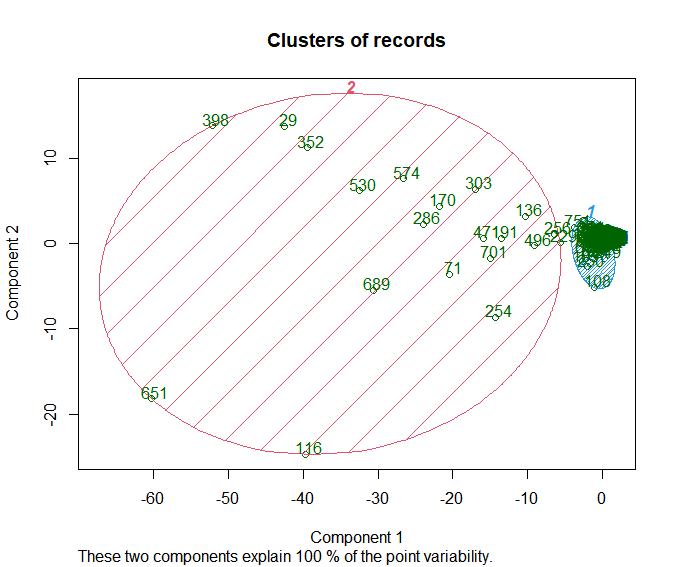
**Analysis**

PC1 explains 91.67 % of variability of data.

PC2 explains 5.442 % of variability of data.



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



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.
6. **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.

1. **Applying k means clustering:**

**R code:**

# K-Means Clustering

# Importing the dataset

dataset = read.csv('DataGroupedBySTATENames.csv')

dataset = dataset[,c(2:30)]

summary(prcomp(dataset, rank. = 2))

library(caret)

library(e1071)

pca = preProcess(x = dataset, method = 'pca', pcaComp = 2)

dataset = predict(pca, dataset)

set.seed(6)

wcss = vector()

# Using the elbow method to find the optimal number of clusters

for (i in 1:30) wcss[i] = sum(kmeans(dataset, i)$withinss)

plot(1:30,

wcss,

type = 'b',

main = paste('The Elbow Method'),

xlab = 'Number of clusters',

ylab = 'WCSS')

# Fitting K-Means to the dataset

set.seed(29)

kmeans = kmeans(x = dataset, centers = 2)

y\_kmeans = kmeans$cluster

# Visualising the clusters

library(cluster)

clusplot(dataset,

y\_kmeans,

lines = 0,

shade = TRUE,

color = TRUE,

labels = 2,

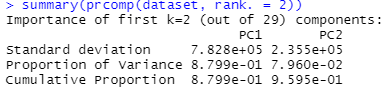
plotchar = FALSE,

span = TRUE,

main = paste('Clusters of records'),

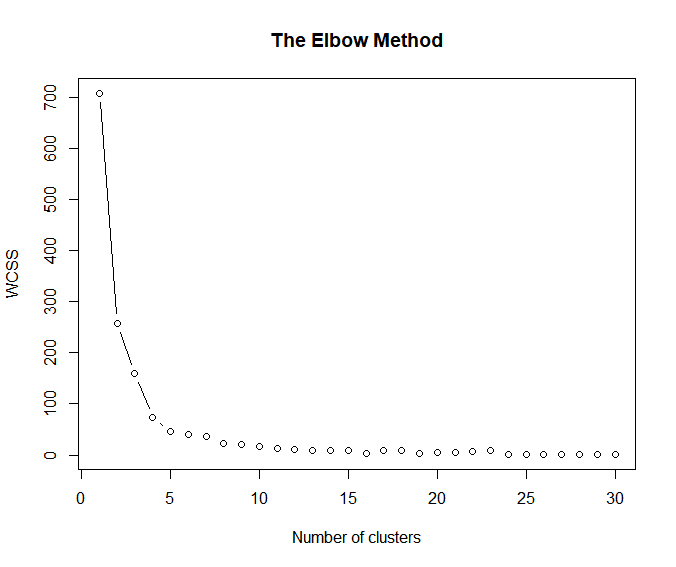
)

**Output:**

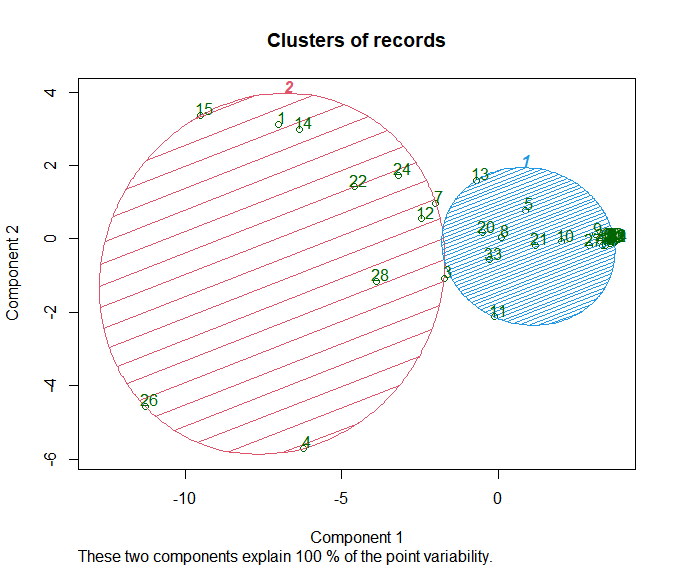
****

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.



**Analysis:**

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.
6. **Applying hierarchical clustering:**

**R code:**

# Hierarchical Clustering

# Importing the dataset

dataset = read.csv('DataGroupedBySTATENames.csv')

# Feature Scaling

#dataset[-1] = scale(dataset[-1])

dataset = dataset[,c(2:30)]

summary(prcomp(dataset, rank. = 2))

library(caret)

library(e1071)

pca = preProcess(x = dataset, method = 'pca', pcaComp = 2)

dataset = predict(pca, dataset)

# Using the dendrogram to find the optimal number of clusters

dendrogram = hclust(d = dist(dataset, method = 'euclidean'), method = 'ward.D')

plot(dendrogram,

main = paste('Dendrogram'),

xlab = 'Records',

ylab = 'Euclidean distances')

# Fitting Hierarchical Clustering to the dataset

hc = hclust(d = dist(dataset, method = 'euclidean'), method = 'ward.D')

y\_hc = cutree(hc, 2)

# Visualising the clusters

library(cluster)

clusplot(dataset,

y\_hc,

lines = 0,

shade = TRUE,

color = TRUE,

labels= 2,

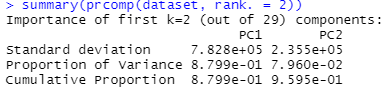
plotchar = FALSE,

span = TRUE,

main = paste('Clusters of records'),

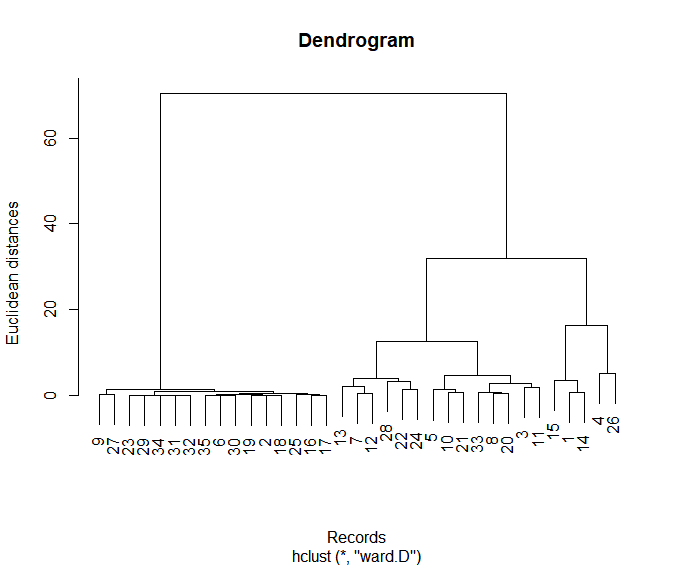
)

**Output:**

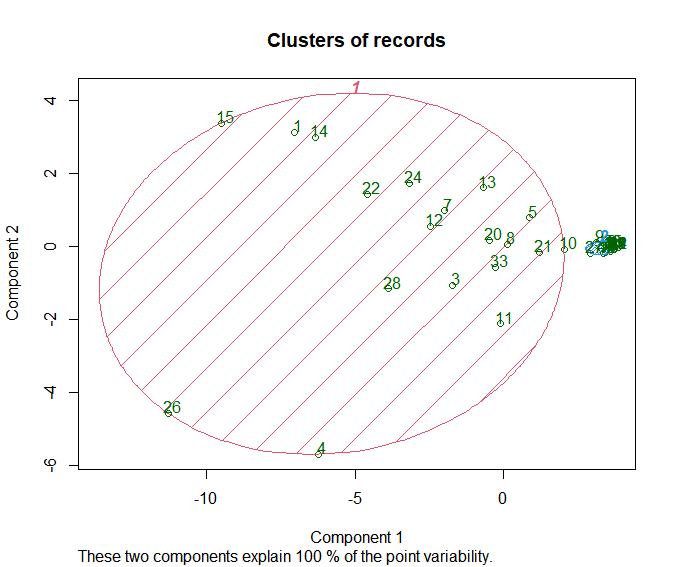
****

PC1explains 87.99 % variability of the data.

PC2 explains 7.96 % variability of the data.



From the above dendogram we can conclude that the optimal number of clusters/centers is 2.



**Analysis:**

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.

## Applying regression algorithms on the data sets:

## 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 variable, 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.**

**NOTE: All the data sets and r code for these 390 models are attached with report. The file names are kept as they end with the index of the dependent variable in the respective data set.**

**We are only going to analyze model performances on the basis of r-squared and adjusted R-squared value of the regression models.**

**LOGIC OF THE R CODE WE ARE GOING TO USE TO EVALUATE R SQAURED AND ADJUSTED R SQUARED VALUE OF NON LINEAR MODEL IS 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.**

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.

**Logic:**

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)

1. **Multiple regression:**

**R code:**

# Multiple Linear Regression

# Importing the dataset

dataset <- read.csv("dataa.csv")

# Splitting the dataset into the Training set and Test set

# install.packages('caTools')

library(caTools)

set.seed(123)

dataset1 <- dataset[, c(1:844,874)]

split = sample.split(dataset1$TOTAL.IPC.CRIMES, SplitRatio = 0.8)

training\_set = subset(dataset1, split == TRUE)

test\_set = subset(dataset1, split == FALSE)

# Feature Scaling

#training\_set[845] = scale(training\_set[845])

#test\_set[845] = scale(test\_set[845])

# Fitting Multiple Linear Regression to the Training set

regressor = lm(formula = TOTAL.IPC.CRIMES ~ .,

data = training\_set)

summary(regressor)

# Predicting the Test set results

y\_pred = predict(regressor,newdata = test\_set)

#Visualising the training set results

library(ggplot2)

ggplot() +

geom\_point(aes(x = training\_set$YEAR, y = training\_set$TOTAL.IPC.CRIMES),

colour = 'red') +

geom\_point(aes(x = training\_set$YEAR, y = predict(regressor, newdata = training\_set)),

colour = 'blue') +

ggtitle('TOTAL.IPC.CRIMES Vs.YEAR (Multiple Regression) - Training set results') +

xlab('YEAR') +

ylab('TOTAL.IPC.CRIMES')

ggplot() +

geom\_point(aes(x = test\_set$YEAR, y = test\_set$TOTAL.IPC.CRIMES),

colour = 'red') +

geom\_point(aes(x = test\_set$YEAR, y = predict(regressor, newdata = test\_set)),

colour = 'blue') +

ggtitle('TOTAL.IPC.CRIMES Vs.YEAR (Multiple Regression) - Test Set Results') +

xlab('YEAR') +

ylab('TOTAL.IPC.CRIMES')

arr1 = array(test\_set$TOTAL.IPC.CRIMES)

ssres1 = 0

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

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

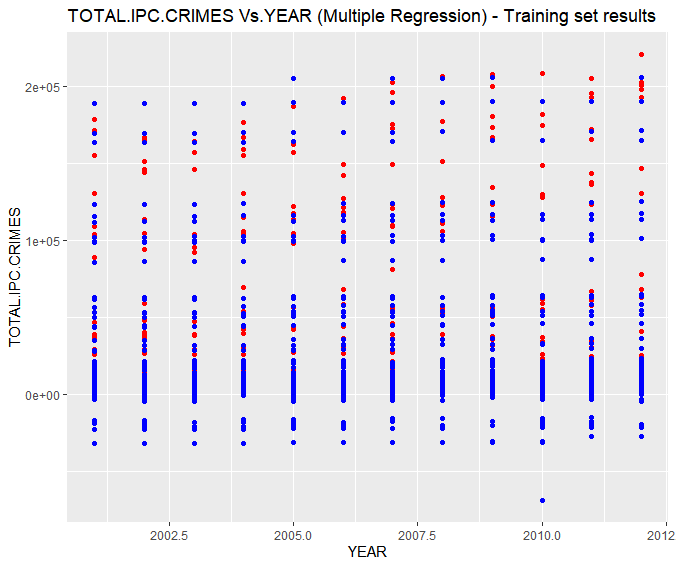
ssres1 = ssres1 + ((abs(arr1[i] - ss[i]))/abs(arr1[i]))

}

m = nrow(test\_set)

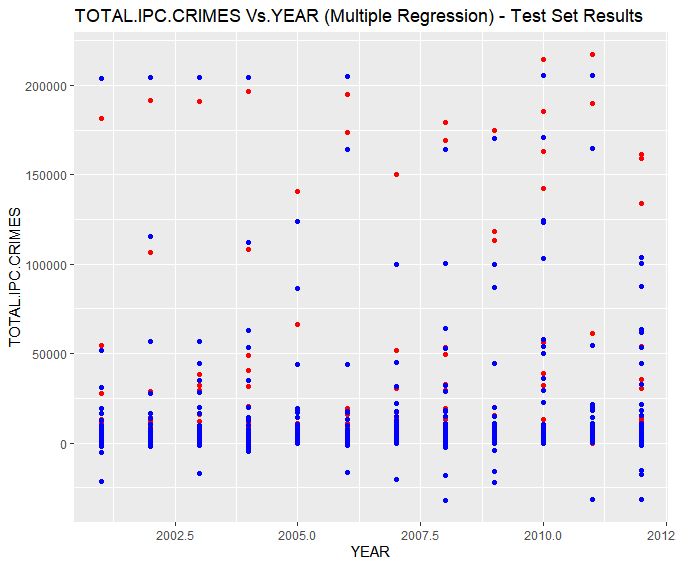
mape = (ssres1/m)

**Output:**



**Analysis:**

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 where as 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.



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 where as 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

Coefficients: (9 not defined because of singularities)

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

(Intercept) -2.378e+05 3.910e+04 -6.082 1.26e-09 \*\*\*

STATE.UT.A...N.ISLANDS -8.466e+04 2.604e+03 -32.508 < 2e-16 \*\*\*

STATE.UT.ANDHRA.PRADESH 7.749e+04 2.587e+03 29.957 < 2e-16 \*\*\*

STATE.UT.ARUNACHAL.PRADESH -8.423e+04 2.521e+03 -33.410 < 2e-16 \*\*\*

STATE.UT.ASSAM -2.301e+04 2.303e+03 -9.992 < 2e-16 \*\*\*

STATE.UT.BIHAR 2.592e+04 2.466e+03 10.509 < 2e-16 \*\*\*

STATE.UT.CHANDIGARH -8.333e+04 2.666e+03 -31.252 < 2e-16 \*\*\*

STATE.UT.CHHATTISGARH -2.430e+04 2.212e+03 -10.985 < 2e-16 \*\*\*

STATE.UT.D...N.HAVELI -8.615e+04 2.466e+03 -34.931 < 2e-16 \*\*\*

STATE.UT.DAMAN...DIU -8.633e+04 2.466e+03 -35.002 < 2e-16 \*\*\*

STATE.UT.DELHI.UT -7.430e+04 2.417e+03 -30.744 < 2e-16 \*\*\*

STATE.UT.GOA -8.396e+04 2.521e+03 -33.297 < 2e-16 \*\*\*

STATE.UT.GUJARAT 2.961e+04 2.466e+03 12.004 < 2e-16 \*\*\*

STATE.UT.HARYANA -4.164e+04 2.441e+03 -17.059 < 2e-16 \*\*\*

STATE.UT.HIMACHAL.PRADESH -2.919e+04 2.165e+03 -13.485 < 2e-16 \*\*\*

STATE.UT.JAMMU...KASHMIR -3.604e+04 2.258e+03 -15.961 < 2e-16 \*\*\*

STATE.UT.JHARKHAND -5.112e+04 2.587e+03 -19.761 < 2e-16 \*\*\*

STATE.UT.KARNATAKA 1.608e+04 2.317e+03 6.942 4.26e-12 \*\*\*

STATE.UT.KERALA 1.269e+04 2.310e+03 5.494 4.08e-08 \*\*\*

STATE.UT.LAKSHADWEEP -8.667e+04 2.766e+03 -31.338 < 2e-16 \*\*\*

STATE.UT.MADHYA.PRADESH 1.184e+05 3.061e+03 38.666 < 2e-16 \*\*\*

STATE.UT.MAHARASHTRA 1.028e+05 2.420e+03 42.502 < 2e-16 \*\*\*

STATE.UT.MANIPUR -7.656e+04 2.506e+03 -30.545 < 2e-16 \*\*\*

STATE.UT.MEGHALAYA -8.449e+04 2.521e+03 -33.509 < 2e-16 \*\*\*

STATE.UT.MIZORAM -8.453e+04 2.587e+03 -32.676 < 2e-16 \*\*\*

STATE.UT.NAGALAND -8.551e+04 2.521e+03 -33.915 < 2e-16 \*\*\*

STATE.UT.ODISHA -3.301e+04 2.587e+03 -12.763 < 2e-16 \*\*\*

STATE.UT.PUDUCHERRY -8.212e+04 2.521e+03 -32.570 < 2e-16 \*\*\*

STATE.UT.PUNJAB -5.455e+04 2.521e+03 -21.637 < 2e-16 \*\*\*

STATE.UT.RAJASTHAN 3.742e+04 2.285e+03 16.371 < 2e-16 \*\*\*

STATE.UT.SIKKIM -8.117e+04 2.244e+03 -36.170 < 2e-16 \*\*\*

STATE.UT.TAMIL.NADU 8.347e+04 2.521e+03 33.106 < 2e-16 \*\*\*

STATE.UT.TRIPURA -7.633e+04 2.198e+03 -34.720 < 2e-16 \*\*\*

STATE.UT.UTTAR.PRADESH 1.339e+04 2.152e+03 6.222 5.23e-10 \*\*\*

STATE.UT.UTTARAKHAND -7.787e+04 2.587e+03 -30.102 < 2e-16 \*\*\*

STATE.UT.WEST.BENGAL NA NA NA NA

DISTRICT.24.PARGANAS.NORTH -7.522e+04 3.624e+03 -20.758 < 2e-16 \*\*\*

DISTRICT.24.PARGANAS.SOUTH -7.463e+04 3.726e+03 -20.029 < 2e-16 \*\*\*

DISTRICT.A.and.N.ISLANDS -1.000e+02 6.350e+03 -0.016 0.98743

DISTRICT.ADILABAD -1.580e+05 3.670e+03 -43.048 < 2e-16 \*\*\*

DISTRICT.AGRA -9.225e+04 3.286e+03 -28.074 < 2e-16 \*\*\*

DISTRICT.AHMEDABAD.COMMR. -9.720e+04 3.644e+03 -26.675 < 2e-16 \*\*\*

DISTRICT.AHMEDABAD.RURAL -1.120e+05 3.539e+03 -31.647 < 2e-16 \*\*\*

DISTRICT.AHMEDNAGAR -1.827e+05 3.612e+03 -50.571 < 2e-16 \*\*\*

DISTRICT.AHWA.DANG -1.149e+05 3.539e+03 -32.468 < 2e-16 \*\*\*

DISTRICT.AIZAWL 1.122e+01 3.586e+03 0.003 0.99750

DISTRICT.AJMER -1.162e+05 3.524e+03 -32.989 < 2e-16 \*\*\*

DISTRICT.AKOLA -1.842e+05 3.467e+03 -53.129 < 2e-16 \*\*\*

DISTRICT.ALAPUZHA -9.042e+04 3.615e+03 -25.012 < 2e-16 \*\*\*

DISTRICT.ALIGARH -9.448e+04 3.439e+03 -27.476 < 2e-16 \*\*\*

DISTRICT.ALIRAJPUR -2.032e+05 4.508e+03 -45.068 < 2e-16 \*\*\*

DISTRICT.ALLAHABAD -9.412e+04 3.516e+03 -26.772 < 2e-16 \*\*\*

DISTRICT.ALMORA -7.540e+03 3.624e+03 -2.081 0.03750 \*

DISTRICT.ALWAR -1.146e+05 3.375e+03 -33.954 < 2e-16 \*\*\*

DISTRICT.AMBALA -4.136e+04 3.568e+03 -11.592 < 2e-16 \*\*\*

DISTRICT.AMBALA.RURAL -4.366e+04 6.286e+03 -6.946 4.13e-12 \*\*\*

DISTRICT.AMBALA.URBAN NA NA NA NA

DISTRICT.AMBEDKAR.NAGAR -9.804e+04 3.439e+03 -28.513 < 2e-16 \*\*\*

DISTRICT.AMRAVATI.COMMR. -1.856e+05 3.506e+03 -52.934 < 2e-16 \*\*\*

DISTRICT.AMRAVATI.RURAL -1.845e+05 3.686e+03 -50.064 < 2e-16 \*\*\*

DISTRICT.AMRELI -1.128e+05 3.539e+03 -31.862 < 2e-16 \*\*\*

DISTRICT.AMRITSAR -2.909e+04 3.753e+03 -7.751 1.05e-14 \*\*\*

DISTRICT.AMRITSAR.RURAL -3.046e+04 3.976e+03 -7.661 2.13e-14 \*\*\*

DISTRICT.ANAND -1.117e+05 3.539e+03 -31.573 < 2e-16 \*\*\*

DISTRICT.ANANTAPUR -1.580e+05 3.624e+03 -43.609 < 2e-16 \*\*\*

DISTRICT.ANANTNAG -4.805e+04 3.506e+03 -13.704 < 2e-16 \*\*\*

DISTRICT.ANDAMAN 4.762e+01 3.739e+03 0.013 0.98984

DISTRICT.ANGUL -5.019e+04 3.726e+03 -13.470 < 2e-16 \*\*\*

DISTRICT.ANJAW -1.852e+03 4.453e+03 -0.416 0.67746

DISTRICT.ANUPPUR -2.026e+05 4.070e+03 -49.780 < 2e-16 \*\*\*

DISTRICT.ARARIA -1.092e+05 3.644e+03 -29.957 < 2e-16 \*\*\*

DISTRICT.ARIYALUR -1.668e+05 4.984e+03 -33.475 < 2e-16 \*\*\*

DISTRICT.ARWAL -1.107e+05 3.539e+03 -31.274 < 2e-16 \*\*\*

DISTRICT.ASANSOL -8.381e+04 3.624e+03 -23.128 < 2e-16 \*\*\*

DISTRICT.ASHOK.NAGAR -2.017e+05 4.070e+03 -49.566 < 2e-16 \*\*\*

DISTRICT.AURAIYA -9.783e+04 3.377e+03 -28.967 < 2e-16 \*\*\*

DISTRICT.AURANGABAD -1.088e+05 3.717e+03 -29.277 < 2e-16 \*\*\*

DISTRICT.AURANGABAD.COMMR. -1.853e+05 3.506e+03 -52.841 < 2e-16 \*\*\*

DISTRICT.AURANGABAD.RURAL -1.853e+05 3.554e+03 -52.151 < 2e-16 \*\*\*

DISTRICT.AWANTIPORA -4.906e+04 3.397e+03 -14.442 < 2e-16 \*\*\*

DISTRICT.AZAMGARH -9.713e+04 3.327e+03 -29.192 < 2e-16 \*\*\*

DISTRICT.BADAUN -9.644e+04 3.327e+03 -28.985 < 2e-16 \*\*\*

DISTRICT.BADDIPOLICEDIST -5.647e+04 4.261e+03 -13.253 < 2e-16 \*\*\*

DISTRICT.BAGAHA -1.102e+05 3.586e+03 -30.719 < 2e-16 \*\*\*

DISTRICT.BAGALKOT -9.955e+04 3.396e+03 -29.314 < 2e-16 \*\*\*

DISTRICT.BAGESHWAR -7.530e+03 3.586e+03 -2.100 0.03578 \*

DISTRICT.BAGHPAT -9.764e+04 3.327e+03 -29.344 < 2e-16 \*\*\*

DISTRICT.BAHRAICH -9.733e+04 3.286e+03 -29.619 < 2e-16 \*\*\*

DISTRICT.BAKSA NA NA NA NA

DISTRICT.BALAGHAT -2.011e+05 4.018e+03 -50.054 < 2e-16 \*\*\*

DISTRICT.BALASORE -4.949e+04 3.554e+03 -13.927 < 2e-16 \*\*\*

DISTRICT.BALLIA -9.795e+04 3.286e+03 -29.808 < 2e-16 \*\*\*

DISTRICT.BALOD -6.096e+04 6.200e+03 -9.832 < 2e-16 \*\*\*

DISTRICT.BALODA.BAZAR -5.971e+04 6.200e+03 -9.630 < 2e-16 \*\*\*

DISTRICT.BALRAMPUR -7.943e+04 3.032e+03 -26.197 < 2e-16 \*\*\*

DISTRICT.BANDA -9.765e+04 3.327e+03 -29.348 < 2e-16 \*\*\*

DISTRICT.BANDIPORA -4.941e+04 4.007e+03 -12.330 < 2e-16 \*\*\*

DISTRICT.BANGALORE.COMMR. -7.287e+04 3.544e+03 -20.561 < 2e-16 \*\*\*

DISTRICT.BANGALORE.RURAL -9.547e+04 3.436e+03 -27.784 < 2e-16 \*\*\*

DISTRICT.BANKA -1.096e+05 3.586e+03 -30.568 < 2e-16 \*\*\*

DISTRICT.BANKURA -8.381e+04 3.798e+03 -22.070 < 2e-16 \*\*\*

DISTRICT.BANSWARA -1.204e+05 3.524e+03 -34.160 < 2e-16 \*\*\*

DISTRICT.BARABANKI -9.684e+04 3.286e+03 -29.470 < 2e-16 \*\*\*

DISTRICT.BARAGARH -5.095e+04 3.624e+03 -14.060 < 2e-16 \*\*\*

DISTRICT.BARAMULLA -4.814e+04 3.446e+03 -13.969 < 2e-16 \*\*\*

DISTRICT.BARAN -1.197e+05 3.415e+03 -35.035 < 2e-16 \*\*\*

DISTRICT.BAREILLY -9.434e+04 3.439e+03 -27.437 < 2e-16 \*\*\*

DISTRICT.BARMER -1.201e+05 3.464e+03 -34.674 < 2e-16 \*\*\*

DISTRICT.BARNALA -3.034e+04 3.753e+03 -8.083 7.48e-16 \*\*\*

DISTRICT.BARPETA -6.009e+04 3.475e+03 -17.291 < 2e-16 \*\*\*

DISTRICT.BARWANI -2.015e+05 4.070e+03 -49.519 < 2e-16 \*\*\*

DISTRICT.BASKA -6.226e+04 3.610e+03 -17.246 < 2e-16 \*\*\*

DISTRICT.BASTI -9.780e+04 3.327e+03 -29.394 < 2e-16 \*\*\*

DISTRICT.BATALA -3.003e+04 3.539e+03 -8.485 < 2e-16 \*\*\*

DISTRICT.BDN.CP -8.293e+04 6.344e+03 -13.073 < 2e-16 \*\*\*

DISTRICT.BEED -1.843e+05 3.467e+03 -53.150 < 2e-16 \*\*\*

DISTRICT.BEGUSARAI -1.082e+05 3.500e+03 -30.922 < 2e-16 \*\*\*

DISTRICT.BELGAUM -9.556e+04 3.436e+03 -27.811 < 2e-16 \*\*\*

DISTRICT.BELLARY -9.803e+04 3.396e+03 -28.864 < 2e-16 \*\*\*

DISTRICT.BEMETARA NA NA NA NA

DISTRICT.BERHAMPUR -5.119e+04 3.624e+03 -14.126 < 2e-16 \*\*\*

DISTRICT.BETTIAH -1.081e+05 3.500e+03 -30.877 < 2e-16 \*\*\*

DISTRICT.BETUL -2.009e+05 4.221e+03 -47.596 < 2e-16 \*\*\*

DISTRICT.BHABHUA -1.098e+05 3.644e+03 -30.140 < 2e-16 \*\*\*

DISTRICT.BHADRAK -5.067e+04 3.726e+03 -13.597 < 2e-16 \*\*\*

DISTRICT.BHAGALPUR -1.082e+05 3.539e+03 -30.577 < 2e-16 \*\*\*

DISTRICT.BHANDARA -1.861e+05 3.506e+03 -53.067 < 2e-16 \*\*\*

DISTRICT.BHARATPUR -1.153e+05 3.341e+03 -34.518 < 2e-16 \*\*\*

DISTRICT.BHARUCH -1.123e+05 3.500e+03 -32.095 < 2e-16 \*\*\*

DISTRICT.BHATINDA -2.931e+04 3.506e+03 -8.360 < 2e-16 \*\*\*

DISTRICT.BHAVNAGAR -1.094e+05 3.539e+03 -30.918 < 2e-16 \*\*\*

DISTRICT.BHILWARA -1.178e+05 3.464e+03 -34.014 < 2e-16 \*\*\*

DISTRICT.BHIM.NAGAR -9.786e+04 4.809e+03 -20.351 < 2e-16 \*\*\*

DISTRICT.BHIND -2.004e+05 3.976e+03 -50.395 < 2e-16 \*\*\*

DISTRICT.BHIWANI -4.136e+04 3.568e+03 -11.590 < 2e-16 \*\*\*

DISTRICT.BHOJPUR -1.080e+05 3.539e+03 -30.519 < 2e-16 \*\*\*

DISTRICT.BHOPAL -1.895e+05 4.070e+03 -46.554 < 2e-16 \*\*\*

DISTRICT.BHOPAL.RLY. -2.024e+05 3.942e+03 -51.361 < 2e-16 \*\*\*

DISTRICT.BIDAR -9.818e+04 3.485e+03 -28.174 < 2e-16 \*\*\*

DISTRICT.BIEO -6.335e+04 6.233e+03 -10.164 < 2e-16 \*\*\*

DISTRICT.BIJAPUR -9.831e+04 3.619e+03 -27.165 < 2e-16 \*\*\*

DISTRICT.BIJNOR -9.659e+04 3.377e+03 -28.599 < 2e-16 \*\*\*

DISTRICT.BIKANER -1.191e+05 3.524e+03 -33.795 < 2e-16 \*\*\*

DISTRICT.BILASPUR -5.549e+04 3.025e+03 -18.344 < 2e-16 \*\*\*

DISTRICT.BIRBHUM -8.331e+04 3.586e+03 -23.232 < 2e-16 \*\*\*

DISTRICT.BISHNUPUR -8.633e+03 3.496e+03 -2.469 0.01356 \*

DISTRICT.BIZAPUR -6.088e+04 3.366e+03 -18.084 < 2e-16 \*\*\*

DISTRICT.BKP.CP -7.977e+04 6.344e+03 -12.574 < 2e-16 \*\*\*

DISTRICT.BOKARO -3.186e+04 3.586e+03 -8.885 < 2e-16 \*\*\*

DISTRICT.BOLANGIR -5.068e+04 3.670e+03 -13.811 < 2e-16 \*\*\*

DISTRICT.BONGAIGAON -6.138e+04 3.475e+03 -17.661 < 2e-16 \*\*\*

DISTRICT.BORDER -4.821e+04 4.008e+03 -12.029 < 2e-16 \*\*\*

DISTRICT.BORDER.DISTRICT -4.881e+04 6.216e+03 -7.852 4.76e-15 \*\*\*

DISTRICT.BOUDH -5.199e+04 3.624e+03 -14.346 < 2e-16 \*\*\*

DISTRICT.BUDGAM -4.834e+04 3.397e+03 -14.231 < 2e-16 \*\*\*

DISTRICT.BULANDSHAHAR -9.588e+04 3.377e+03 -28.390 < 2e-16 \*\*\*

DISTRICT.BULDHANA -1.845e+05 3.506e+03 -52.607 < 2e-16 \*\*\*

DISTRICT.BUNDI -1.194e+05 3.375e+03 -35.381 < 2e-16 \*\*\*

DISTRICT.BURDWAN -8.113e+04 3.554e+03 -22.829 < 2e-16 \*\*\*

DISTRICT.BURHANPUR -2.028e+05 4.221e+03 -48.055 < 2e-16 \*\*\*

DISTRICT.BUXAR -1.094e+05 3.644e+03 -30.032 < 2e-16 \*\*\*

DISTRICT.C.I.D. -6.247e+04 3.535e+03 -17.673 < 2e-16 \*\*\*

DISTRICT.CACHAR -5.838e+04 3.475e+03 -16.800 < 2e-16 \*\*\*

DISTRICT.CAR -1.689e+03 6.350e+03 -0.266 0.79031

DISTRICT.CAW -1.160e+04 4.099e+03 -2.830 0.00467 \*\*

DISTRICT.CBCID -9.767e+04 3.846e+03 -25.397 < 2e-16 \*\*\*

DISTRICT.CBPURA -9.987e+04 3.850e+03 -25.939 < 2e-16 \*\*\*

DISTRICT.CENTRAL -7.975e+03 3.610e+03 -2.209 0.02721 \*

DISTRICT.CHAIBASA -3.355e+04 3.586e+03 -9.357 < 2e-16 \*\*\*

DISTRICT.CHAMARAJNAGAR -1.001e+05 3.485e+03 -28.716 < 2e-16 \*\*\*

DISTRICT.CHAMBA -5.576e+04 3.447e+03 -16.179 < 2e-16 \*\*\*

DISTRICT.CHAMOLI -7.390e+03 3.624e+03 -2.039 0.04145 \*

DISTRICT.CHAMPAWAT -7.475e+03 3.624e+03 -2.063 0.03918 \*

DISTRICT.CHAMPHAI -6.624e+02 3.798e+03 -0.174 0.86155

DISTRICT.CHANDEL -8.922e+03 3.567e+03 -2.501 0.01240 \*

DISTRICT.CHANDIGARH 1.226e+03 3.644e+03 0.336 0.73656

DISTRICT.CHANDOLI -9.812e+04 3.377e+03 -29.053 < 2e-16 \*\*\*

DISTRICT.CHANDRAPUR -1.846e+05 3.467e+03 -53.251 < 2e-16 \*\*\*

DISTRICT.CHANGLANG -1.163e+03 3.577e+03 -0.325 0.74508

DISTRICT.CHATRA -3.352e+04 3.624e+03 -9.249 < 2e-16 \*\*\*

DISTRICT.CHENGAI -1.574e+05 4.162e+03 -37.827 < 2e-16 \*\*\*

DISTRICT.CHENNAI -1.556e+05 3.681e+03 -42.280 < 2e-16 \*\*\*

DISTRICT.CHENNAI.RLY. -1.684e+05 3.539e+03 -47.590 < 2e-16 \*\*\*

DISTRICT.CHENNAISUBURBAN -1.639e+05 4.984e+03 -32.886 < 2e-16 \*\*\*

DISTRICT.CHHATARPUR -1.996e+05 3.942e+03 -50.642 < 2e-16 \*\*\*

DISTRICT.CHHINDWARA -1.995e+05 4.070e+03 -49.011 < 2e-16 \*\*\*

DISTRICT.CHICKMAGALUR -9.874e+04 3.485e+03 -28.336 < 2e-16 \*\*\*

DISTRICT.CHIRANG -6.231e+04 3.610e+03 -17.259 < 2e-16 \*\*\*

DISTRICT.CHITRADURGA -9.775e+04 3.485e+03 -28.053 < 2e-16 \*\*\*

DISTRICT.CHITRAKOOT.DHAM -9.842e+04 3.286e+03 -29.952 < 2e-16 \*\*\*

DISTRICT.CHITTOOR -1.570e+05 3.670e+03 -42.781 < 2e-16 \*\*\*

DISTRICT.CHITTORGARH -1.173e+05 3.375e+03 -34.757 < 2e-16 \*\*\*

DISTRICT.CHURACHANDPUR -8.975e+03 3.671e+03 -2.445 0.01452 \*

DISTRICT.CHURU -1.201e+05 3.415e+03 -35.172 < 2e-16 \*\*\*

DISTRICT.CID -4.110e+04 3.577e+03 -11.489 < 2e-16 \*\*\*

DISTRICT.CID.CRIME -1.158e+05 4.422e+03 -26.188 < 2e-16 \*\*\*

DISTRICT.COIMBATORE.RURAL -1.604e+05 3.577e+03 -44.848 < 2e-16 \*\*\*

DISTRICT.COIMBATORE.URBAN -1.646e+05 3.539e+03 -46.506 < 2e-16 \*\*\*

DISTRICT.COOCHBEHAR -8.327e+04 3.670e+03 -22.692 < 2e-16 \*\*\*

DISTRICT.CP.AMRITSAR -2.976e+04 4.985e+03 -5.971 2.49e-09 \*\*\*

[ reached getOption("max.print") -- omitted 645 rows ]

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

**Analysis:**

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
2. **Support vector regression:**

**R Code:**

# Importing the dataset

dataset = read.csv('dataa.csv')

# Splitting the dataset into the Training set and Test set

# install.packages('caTools')

library(caTools)

set.seed(123)

dataset <- dataset[, c(1:844,874)]

split = sample.split(dataset1$TOTAL.IPC.CRIMES, SplitRatio = 0.8)

training\_set = subset(dataset1, split == TRUE)

test\_set = subset(dataset1, split == FALSE)

# Feature Scaling

training\_set[845] = scale(training\_set[845])

test\_set[845] = scale(test\_set[845])

# Fitting the SVR to the dataset

# Create your regressor here

library(e1071)

regressor = svm(formula = TOTAL.IPC.CRIMES ~ ., data = training\_set, type = 'eps-regression')

# Predicting a new result

#y\_pred = predict(regressor, data.frame(YEAR = 2013))

# Visualising the SVR results

# install.packages('ggplot2')

library(ggplot2)

ggplot() +

geom\_point(aes(x = training\_set$YEAR, y = training\_set$TOTAL.IPC.CRIMES),

colour = 'red') +

geom\_point(aes(x = training\_set$YEAR, y = predict(regressor, newdata = training\_set)),

colour = 'blue') +

ggtitle('TOTAL.IPC.CRIMES Vs.YEAR (SVR Model) - Training set results') +

xlab('YEAR') +

ylab('TOTAL.IPC.CRIMES')

ggplot() +

geom\_point(aes(x = test\_set$YEAR, y = test\_set$TOTAL.IPC.CRIMES),

colour = 'red') +

geom\_point(aes(x = test\_set$YEAR, y = predict(regressor, newdata = test\_set)),

colour = 'blue') +

ggtitle('TOTAL.IPC.CRIMES Vs.YEAR (SVR Model) - Test set results') +

xlab('YEAR') +

ylab('TOTAL.IPC.CRIMES')

arr = array(training\_set$TOTAL.IPC.CRIMES)

ssres = 0

sst = 0

yavg = mean(training\_set$TOTAL.IPC.CRIMES)

ss2 <- array(predict(regressor, newdata = training\_set))

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(training\_set)

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

arr1 = array(test\_set$TOTAL.IPC.CRIMES)

ssres1 = 0

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

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

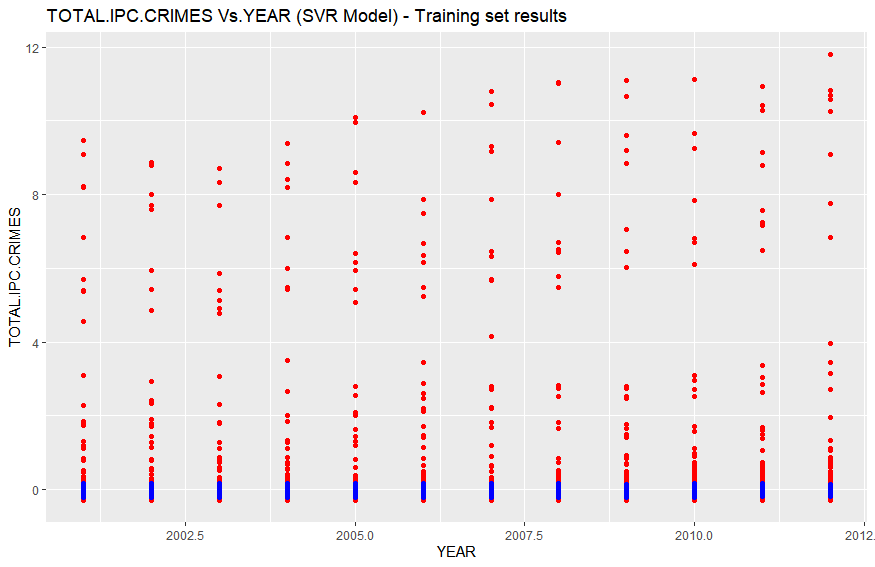
ssres1 = ssres1 + ((abs(arr1[i] - ss[i]))/abs(arr1[i]))

}

m = nrow(test\_set)

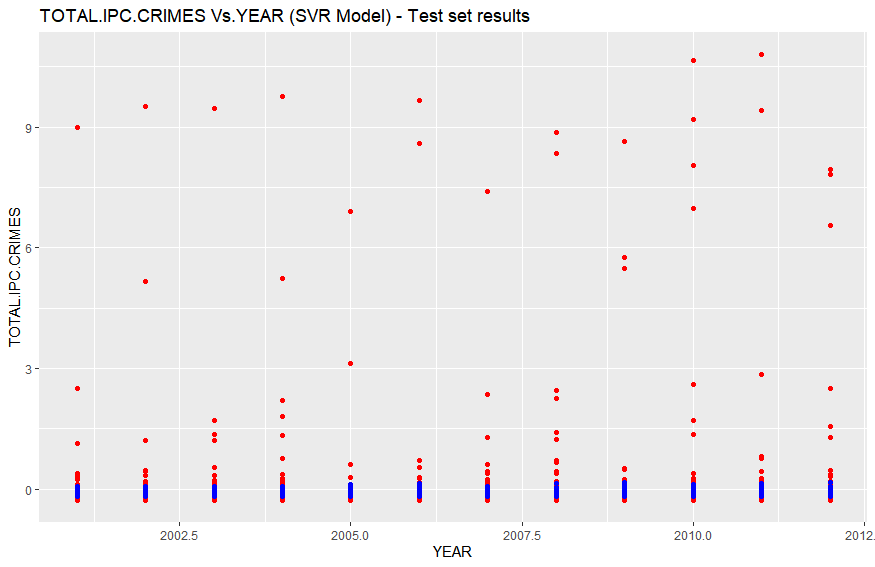
mape = (ssres1/m)

**Output:**



**Analysis:**

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 where as 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 where as 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
2. **Decision Tree:**

**R Code:**

# Decision Tree Regression

# Importing the dataset

dataset = read.csv('dataa.csv')

dataset = dataset[,c(1:844,874)]

# install.packages('caTools')

library(caTools)

set.seed(123)

split = sample.split(dataset1$TOTAL.IPC.CRIMES, SplitRatio = 0.8)

training\_set = subset(dataset1, split == TRUE)

test\_set = subset(dataset1, split == FALSE)

# Fitting Decision Tree Regression to the dataset

# install.packages('rpart')

library(rpart)

regressor = rpart(formula = TOTAL.IPC.CRIMES ~ .,

data = training\_set,

control = rpart.control(minsplit = 2))

# Predicting a new result with Decision Tree Regression

#y\_pred = predict(regressor, data.frame(YEAR = 2013))

# Visualising the Decision Tree Regression results

# install.packages('ggplot2')

library(ggplot2)

ggplot() +

geom\_point(aes(x = training\_set$YEAR, y = training\_set$TOTAL.IPC.CRIMES),

colour = 'red') +

geom\_point(aes(x = training\_set$YEAR, y = predict(regressor, newdata = training\_set)),

colour = 'blue') +

ggtitle('TOTAL.IPC.CRIMES Vs.YEAR (Decision Tree Regression) - Training set') +

xlab('YEAR') +

ylab('TOTAL.IPC.CRIMES')

ggplot() +

geom\_point(aes(x = test\_set$YEAR, y = test\_set$TOTAL.IPC.CRIMES),

colour = 'red') +

geom\_point(aes(x = test\_set$YEAR, y = predict(regressor, newdata = test\_set)),

colour = 'blue') +

ggtitle('TOTAL.IPC.CRIMES Vs.YEAR (Decision Tree Regression) - Test set') +

xlab('YEAR') +

ylab('TOTAL.IPC.CRIMES')

# Plotting the tree

plot(regressor)

text(regressor)

arr = array(training\_set$TOTAL.IPC.CRIMES)

ssres = 0

sst = 0

yavg = mean(training\_set$TOTAL.IPC.CRIMES)

ss2 <- array(predict(regressor, newdata = training\_set))

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(training\_set)

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

arr1 = array(test\_set$TOTAL.IPC.CRIMES)

ssres1 = 0

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

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

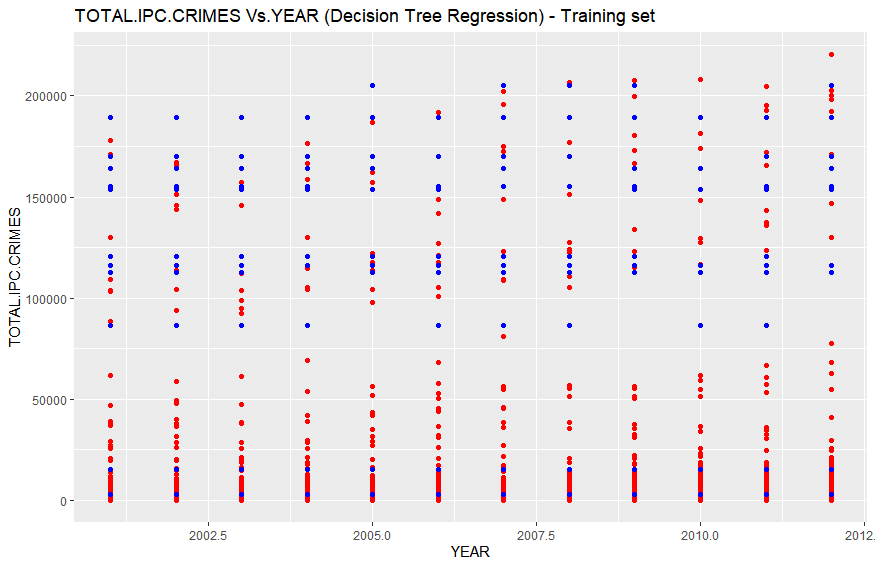
ssres1 = ssres1 + ((abs(arr1[i] - ss[i]))/abs(arr1[i]))

}

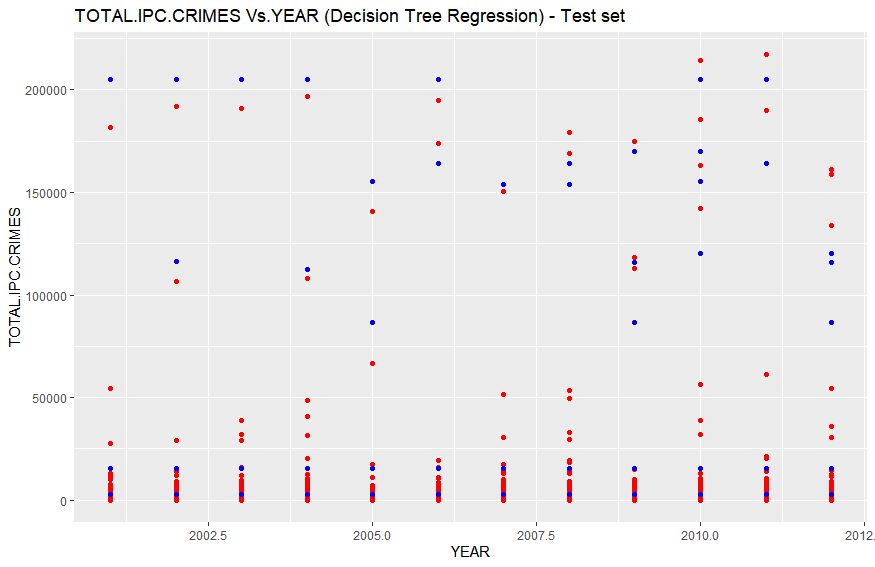
m = nrow(test\_set)

mape = (ssres1/m)

**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 where as 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.



**Analysis:**

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 where as 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
2. **Random Forest Regression:**

**R Code:**

# Random Forest Regression

# Importing the dataset

dataset = read.csv('dataa.csv')

dataset = dataset[,c(1:844,874)]

# Fitting Random Forest Regression to the dataset

# install.packages('randomForest')

library(randomForest)

set.seed(1234)

regressor = randomForest(x = dataset[1:844], y = dataset$TOTAL.IPC.CRIMES, ntree = 50)

# Predicting a new result with Random Forest Regression

#y\_pred = predict(regressor, data.frame(YEAR = 2013))

# Visualising the Random Forest Regression results

# install.packages('ggplot2')

library(ggplot2)

ggplot() +

geom\_point(aes(x = training\_set$YEAR, y = training\_set$TOTAL.IPC.CRIMES),

colour = 'red') +

geom\_point(aes(x = training\_set$YEAR, y = predict(regressor, newdata = training\_set)),

colour = 'blue') +

ggtitle('TOTAL.IPC.CRIMES Vs.YEAR (Random Forest Regression - Training set)') +

xlab('YEAR') +

ylab('TOTAL.IPC.CRIMES')

ggplot() +

geom\_point(aes(x = test\_set$YEAR, y = test\_set$TOTAL.IPC.CRIMES),

colour = 'red') +

geom\_point(aes(x = test\_set$YEAR, y = predict(regressor, newdata = test\_set)),

colour = 'blue') +

ggtitle('TOTAL.IPC.CRIMES Vs.YEAR (Random Forest Regression) - Test set') +

xlab('YEAR') +

ylab('TOTAL.IPC.CRIMES')

arr = array(training\_set$TOTAL.IPC.CRIMES)

ssres = 0

sst = 0

yavg = mean(training\_set$TOTAL.IPC.CRIMES)

ss2 <- array(predict(regressor, newdata = training\_set))

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(training\_set)

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

arr1 = array(test\_set$TOTAL.IPC.CRIMES)

ssres1 = 0

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

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

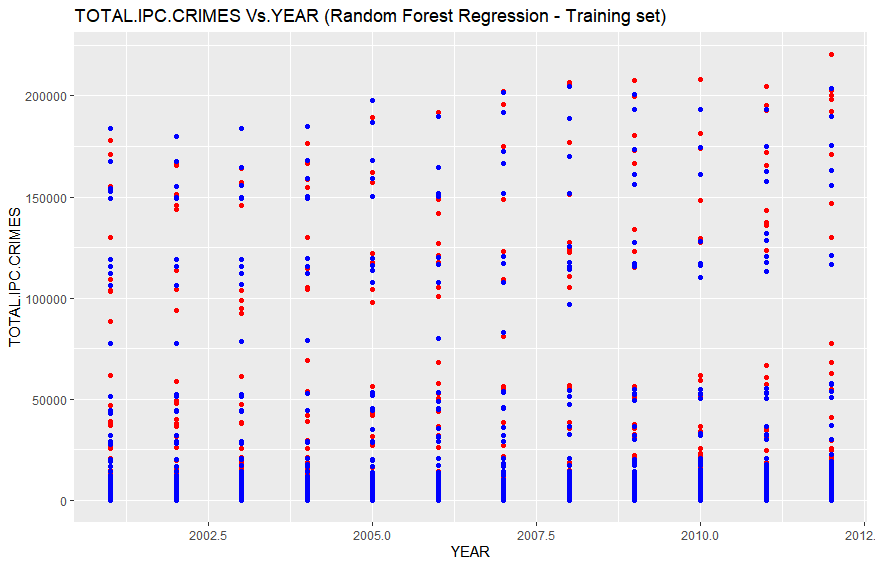
ssres1 = ssres1 + ((abs(arr1[i] - ss[i]))/abs(arr1[i]))

}

m = nrow(test\_set)

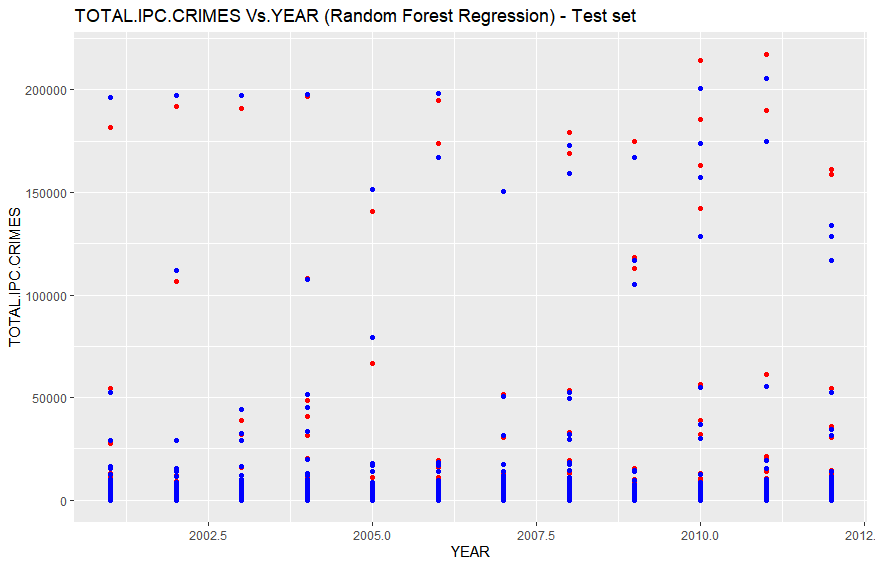
mape = (ssres1/m)

**Output:**



**Analysis:**

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 where as 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.



**Analysis:**

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 where as 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**

1. **Implementation of 5 different regression models on DataGroupedByYEAR.csv data set, with dependent variable RIOTS :**
2. **Simple Linear Regression:**

**R Code:**

# Simple Linear Regression

# Importing the dataset

dataset = read.csv('DataGroupedByYEAR.csv')

# install.packages('caTools')

dataset = dataset[,c(1,18)]

# Feature Scaling

# training\_set = scale(training\_set)

# test\_set = scale(test\_set)

# Fitting Simple Linear Regression to the Training set

regressor = lm(formula = RIOTS ~ YEAR,

data = dataset)

# Predicting a new result

y\_pred = predict(regressor, data.frame(YEAR = 2013))

# Visualising the results

library(ggplot2)

ggplot() +

geom\_point(aes(x = dataset$YEAR, y = dataset$RIOTS),

colour = 'red') +

geom\_line(aes(x = dataset$YEAR, y = predict(regressor, newdata = dataset)),

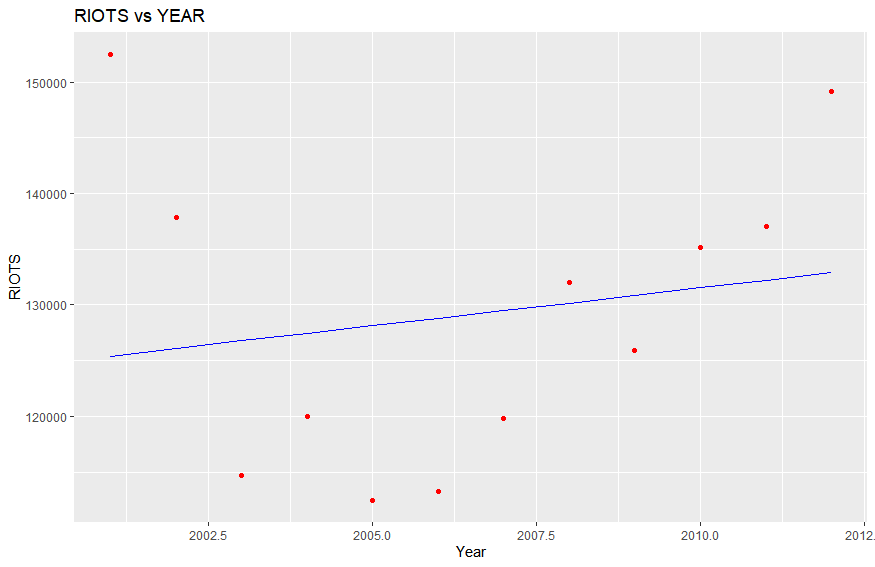
colour = 'blue') +

ggtitle('RIOTS vs YEAR') +

xlab('Year') +

ylab('RIOTS')

**Output:**



**Analysis:**

1. We have only one plot this time, as we had only 12 rows, so we did’nt 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

**Analysis:**

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

**Analysis:**

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

1. **Polynomial Regression:**

**R Code:**

# Polynomial Regression

# Importing the dataset

dataset = read.csv('DataGroupedByYEAR.csv')

dataset = dataset[,c(1,18)]

# Fitting Polynomial Regression to the dataset

dataset$YEAR1 = dataset$YEAR^2

dataset$YEAR2 = dataset$YEAR^3

dataset$YEAR3 = dataset$YEAR^4

poly\_reg = lm(formula = RIOTS ~ YEAR,

data = dataset)

# Visualising the Polynomial Regression results

# install.packages('ggplot2')

library(ggplot2)

ggplot() +

geom\_point(aes(x = dataset$YEAR, y = dataset$RIOTS),

colour = 'red') +

geom\_line(aes(x = dataset$YEAR, y = predict(poly\_reg, newdata = dataset)),

colour = 'blue') +

ggtitle('RIOTS Vs.YEAR (Polynomial Regression)') +

xlab('YEAR') +

ylab('RIOTS')

# Predicting a new result with Polynomial Regression

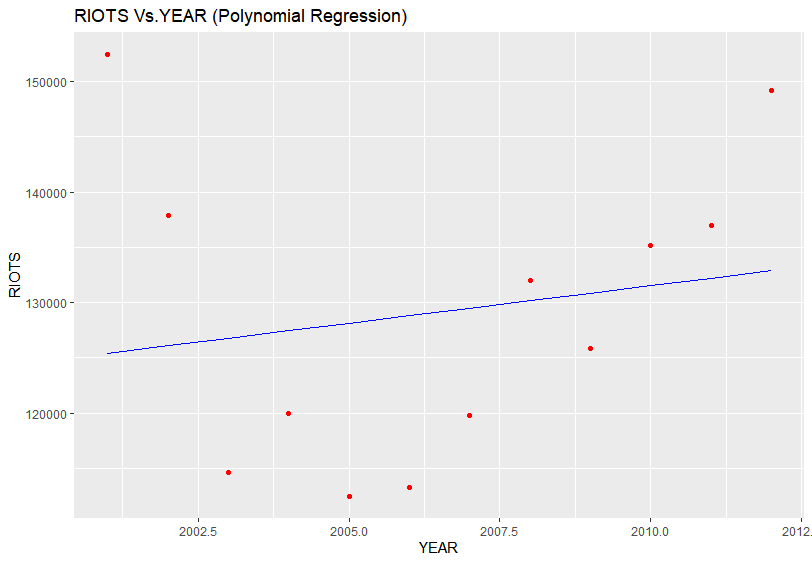
predict(poly\_reg, data.frame(YEAR = 2013,

YEAR1 = 2013^2,

YEAR2 = 2013^3,

YEAR3 = 2013^4))

**Output:**



**Analysis:**

1. We have only one plot this time, as we had only 12 rows, so we did’nt 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

**Analysis**:

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. **Support Vector Regression:**

**R Code:**

# Regression Template

# Importing the dataset

dataset = read.csv('DataGroupedByYEAR.csv')

dataset = dataset[,c(1,18)]

# Fitting the SVR to the dataset

# Create your regressor here

library(e1071)

regressor = svm(formula = RIOTS ~ YEAR, data = dataset, type = 'eps-regression')

# Predicting a new result

y\_pred = predict(regressor, data.frame(YEAR = 2013))

# Visualising the Regression Model results (for higher resolution and smoother curve)

# install.packages('ggplot2')

library(ggplot2)

x\_grid = seq(min(dataset$YEAR), max(dataset$YEAR), 0.1)

ggplot() +

geom\_point(aes(x = dataset$YEAR, y = dataset$RIOTS),

colour = 'red') +

geom\_line(aes(x = x\_grid, y = predict(regressor, newdata = data.frame(YEAR = x\_grid))),

colour = 'blue') +

ggtitle('RIOTS Vs.YEAR (SVR Model)') +

xlab('YEAR') +

ylab('RIOTS')

arr = array(dataset$RIOTS)

ssres = 0

sst = 0

yavg = mean(dataset$RIOTS)

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

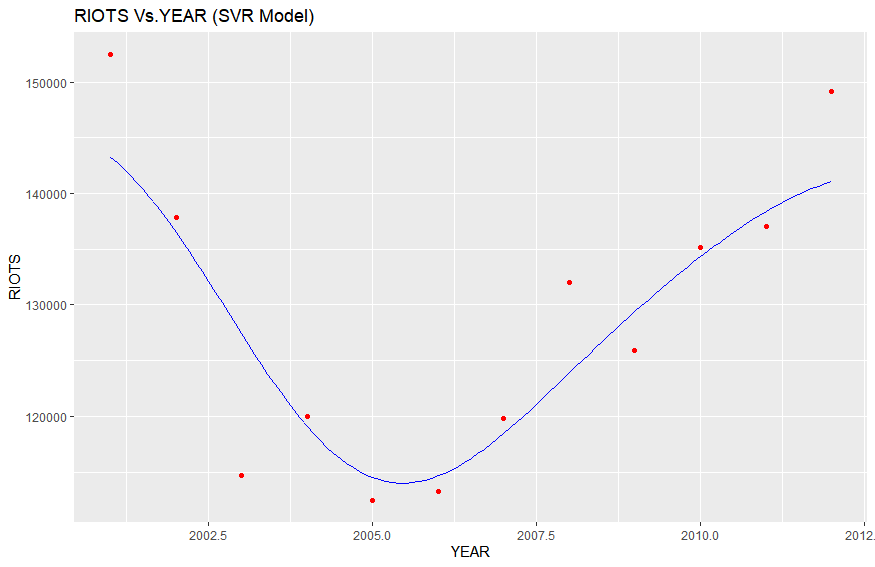
}

R2 = 1 - (ssres/sst)

n = nrow(dataset)

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

**Output:**



1. We have only one plot this time, as we had only 12 rows, so we did’nt 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

1. **Decision Tree Regression:**

**R Code:**

# Decision Tree Regression

# Importing the dataset

dataset = read.csv('DataGroupedByYEAR.csv')

dataset = dataset[,c(1,18)]

# Fitting Decision Tree Regression to the dataset

# install.packages('rpart')

library(rpart)

regressor = rpart(formula = RIOTS ~ YEAR,

data = dataset,

control = rpart.control(minsplit = 2))

# Predicting a new result with Decision Tree Regression

y\_pred = predict(regressor, data.frame(YEAR = 2013))

# Visualising the Decision Tree Regression results (higher resolution)

# install.packages('ggplot2')

library(ggplot2)

x\_grid = seq(min(dataset$YEAR), max(dataset$YEAR), 0.01)

ggplot() +

geom\_point(aes(x = dataset$YEAR, y = dataset$RIOTS),

colour = 'red') +

geom\_line(aes(x = x\_grid, y = predict(regressor, newdata = data.frame(YEAR = x\_grid))),

colour = 'blue') +

ggtitle('RIOTS Vs. YEAR (Decision Tree Regression)') +

xlab('YEAR') +

ylab('RIOTS')

arr = array(dataset$RIOTS)

ssres = 0

sst = 0

yavg = mean(dataset$RIOTS)

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)

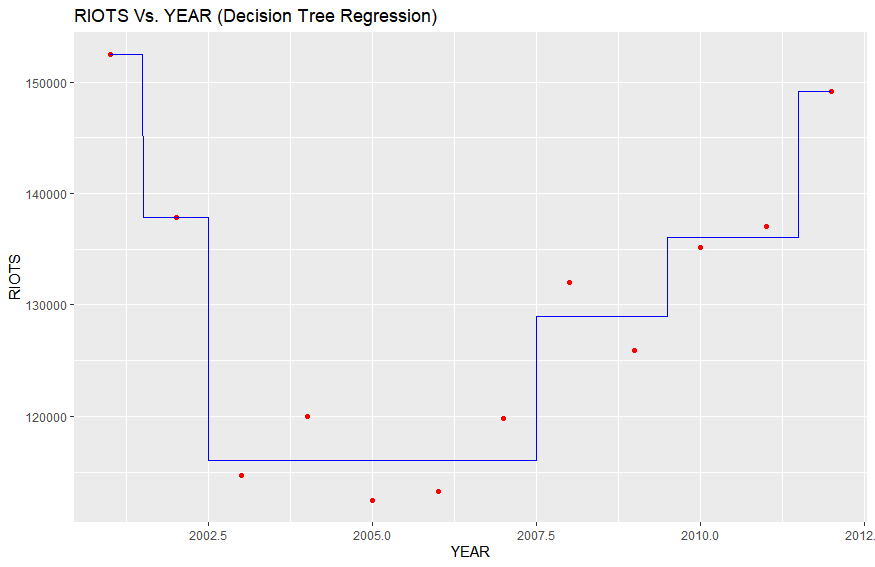
AdjR2 = 1 - (1 - R2)\*((n - 1)/(n - 1 - 1))

# Plotting the tree

plot(regressor)

text(regressor)

**Output:**



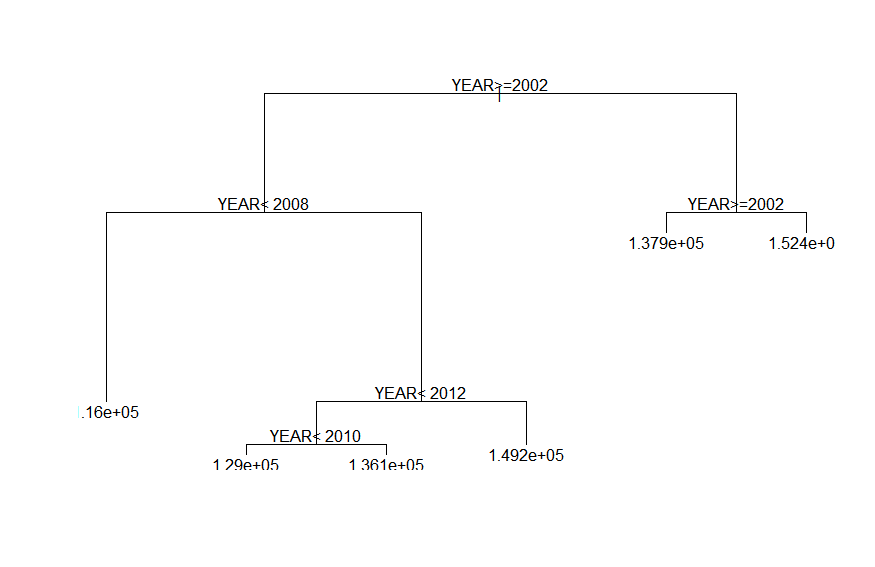
1. We have only one plot this time, as we had only 12 rows, so we did’nt split the data set into training\_set and test\_set.
2. In the above plot we have red points as the actual observed points from the data set and blue line is the regression line that is plot on the basis of predicted values of number 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.

1. **Random Forest regression:**

**R Code:**

# Random Forest Regression

# Importing the dataset

dataset = read.csv('DataGroupedByYEAR.csv')

dataset = dataset[,c(1,18)]

# Fitting Random Forest Regression to the dataset

# install.packages('randomForest')

library(randomForest)

set.seed(1234)

regressor = randomForest(x = dataset[1], y = dataset$RIOTS, ntree = 500)

# Predicting a new result with Random Forest Regression

y\_pred = predict(regressor, data.frame(YEAR = 2013))

# Visualising the Random Forest Regression results (higher resolution)

# install.packages('ggplot2')

library(ggplot2)

x\_grid = seq(min(dataset$YEAR), max(dataset$YEAR), 0.01)

ggplot() +

geom\_point(aes(x = dataset$YEAR, y = dataset$RIOTS),

colour = 'red') +

geom\_line(aes(x = x\_grid, y = predict(regressor, newdata = data.frame(YEAR = x\_grid))),

colour = 'blue') +

ggtitle('RIOTS Vs. YEAR (Random Forest Regression)') +

xlab('YEAR') +

ylab('RIOTS')

arr = array(dataset$RIOTS)

ssres = 0

sst = 0

yavg = mean(dataset$RIOTS)

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)

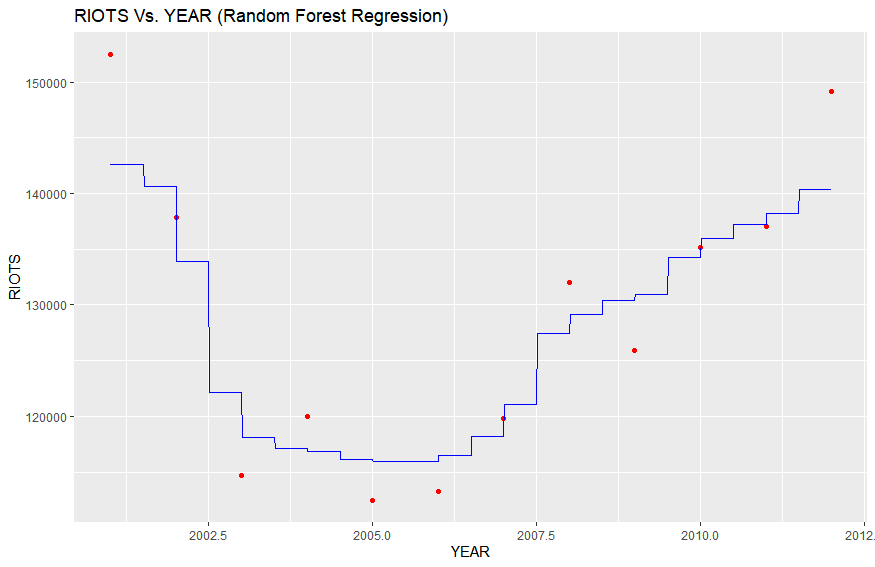
n = nrow(dataset)

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

#Error vs.Regresion tree plot

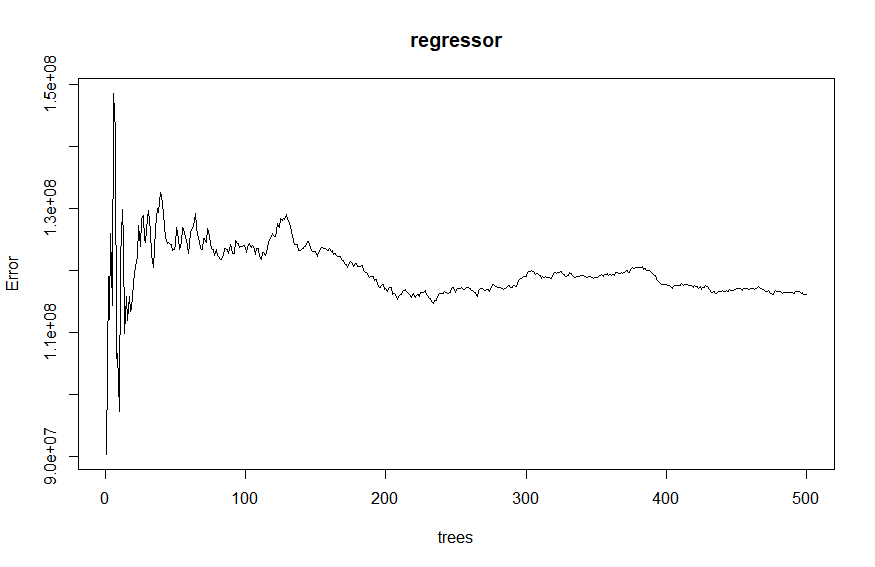
plot(regressor)

**Output:**



**Analysis:**

1. We have only one plot this time, as we had only 12 rows, so we did’nt 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 MODELS 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**

1. **Decision tree classification to classify states on the basis of the total\_ipc\_crimes:**

A decision tree is a machine learning algorithm that partitions the data into subsets. The partitioning process starts with a binary split and continues until no further splits can be made. Various branches of variable length are formed.

By applying modern technology Decision techniques to these cities crime data, Most Crime States can be predicted in different categories . This project analyzes crime data and gives visualizations for easy understanding of the results. It also uses past 12 years’ crime data from data.gov.in to Crime States can be predicted in different categories .

Technologies Used:-

The technologies used for implementing models and visualizations. Technical programming is performed in ‘R’ using its Machine Learning libraries.

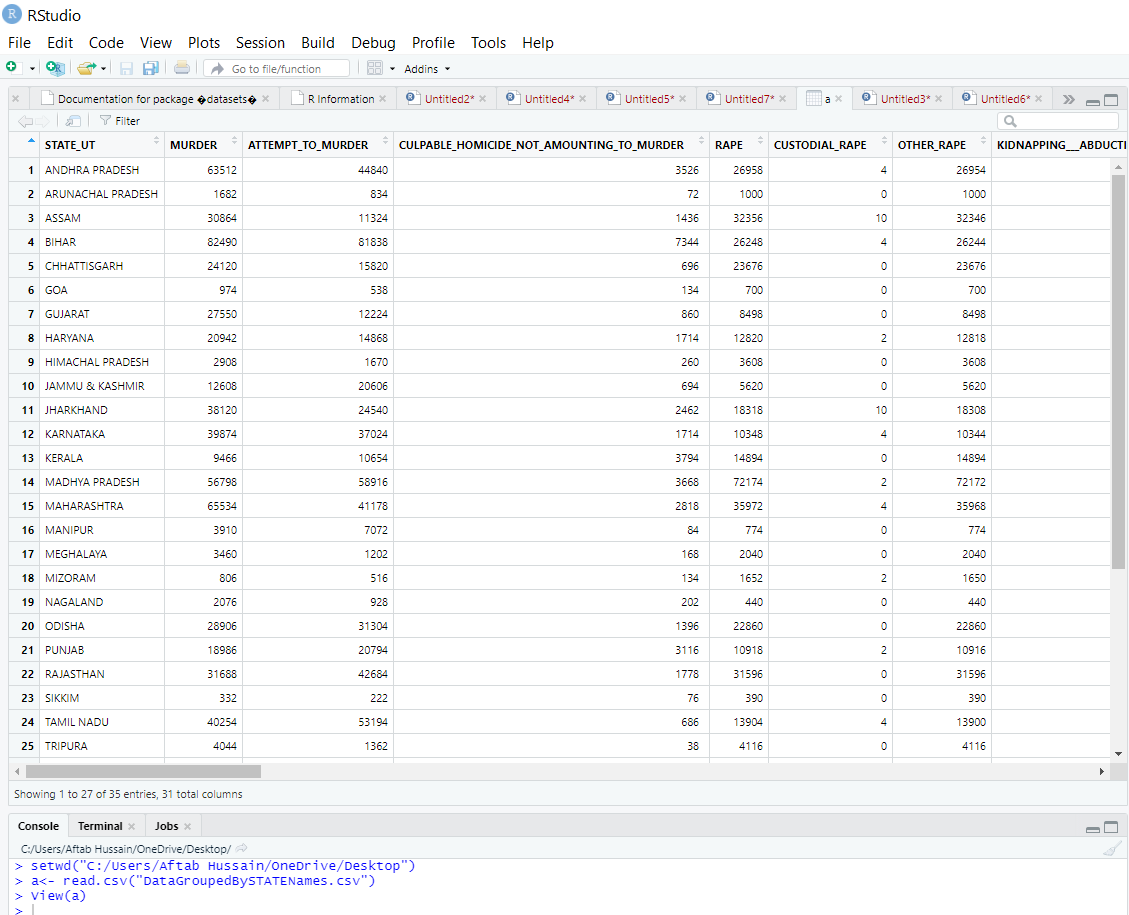
Data Processing :-

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

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

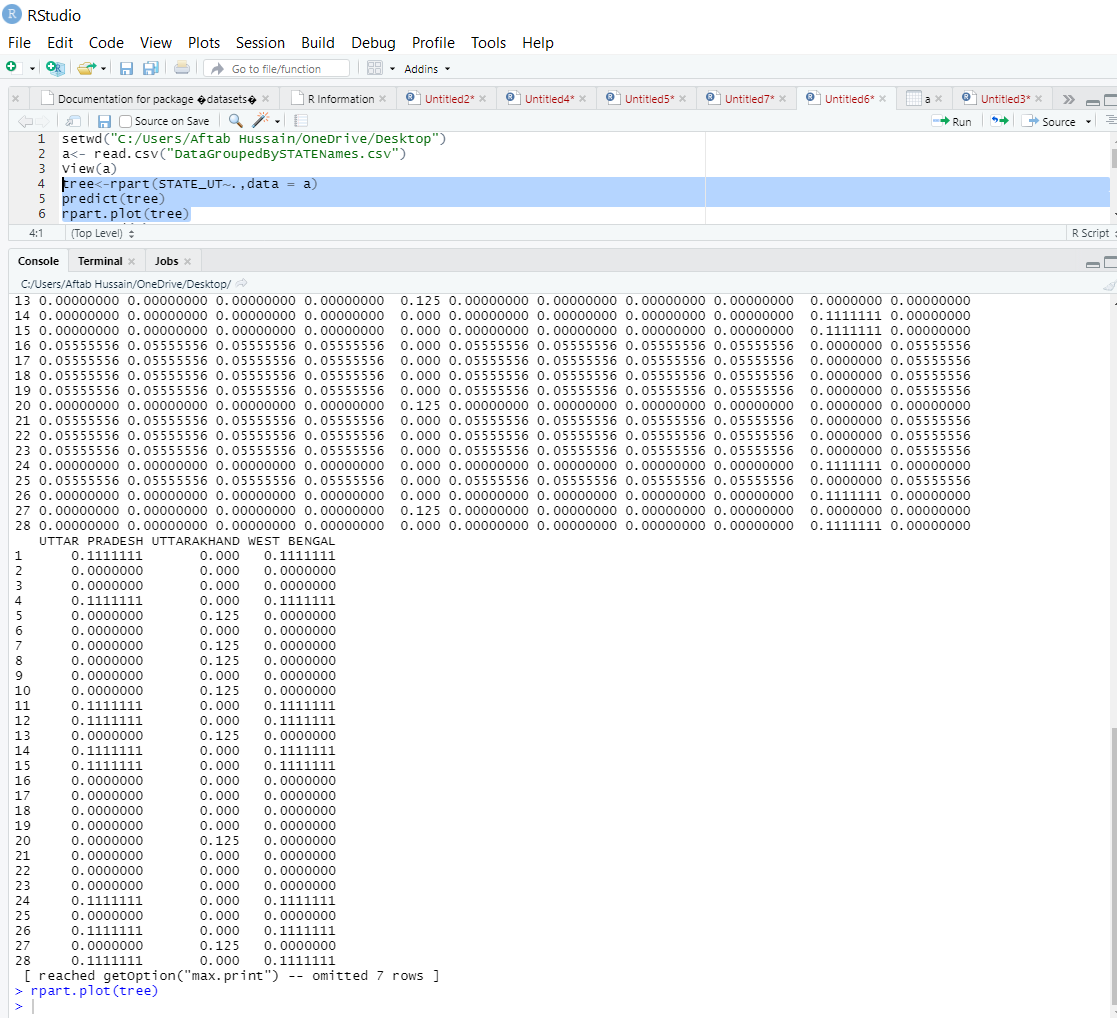
India crime data from 2001 to 2012 is taken from Kaggel.

Data Preprocessing 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 analyzes and takes data from 2001 to Crime States can be predicted in different categories . So, first step of preprocessing is to extract data from 2001 to 2012.

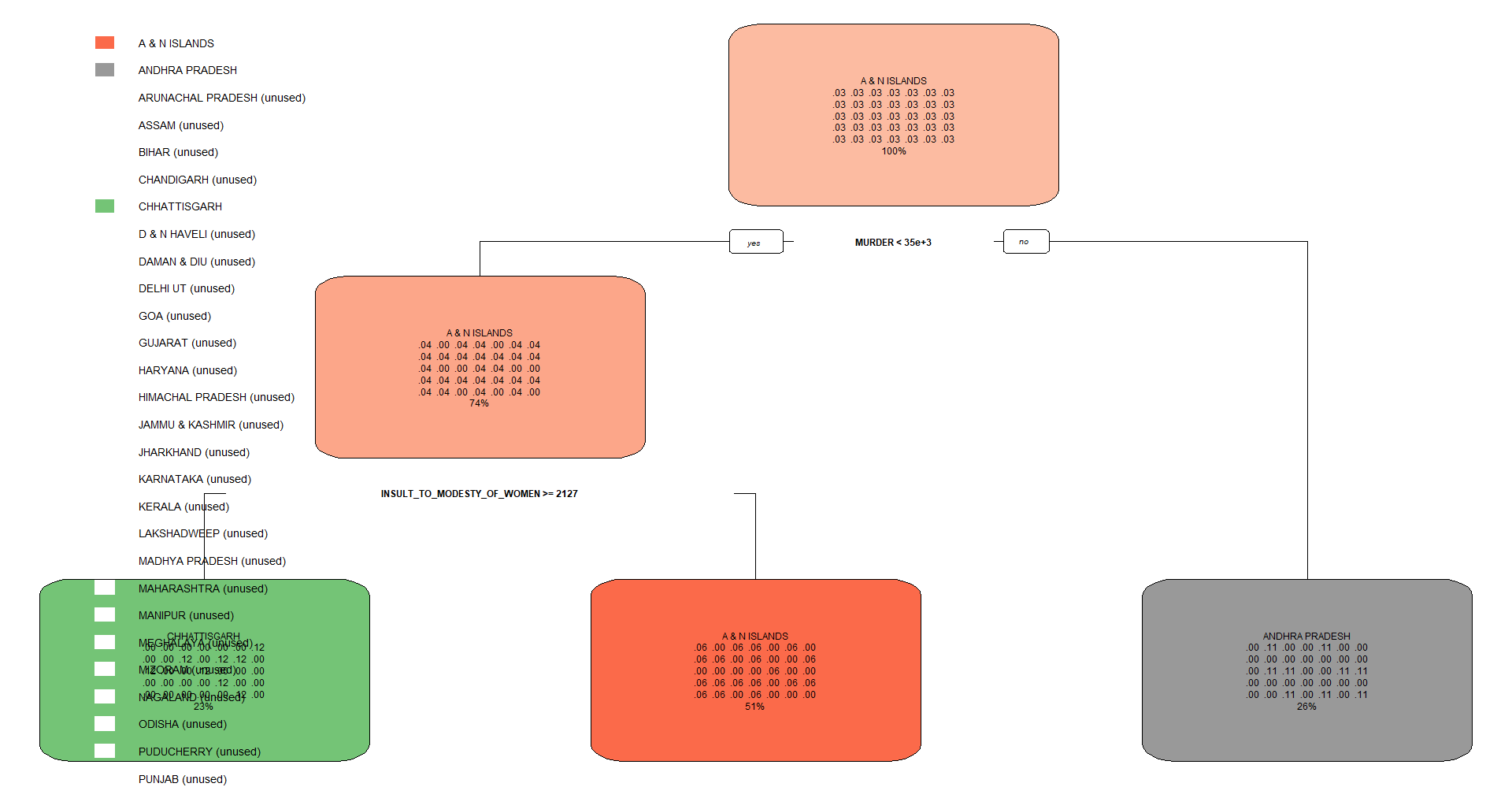


How has Crime Evolved over States in the India ?

**Well we have too many states by diagram it is difficult to explore all states and there different crime . Some of them are selected to show how they represented different states with different crimes.**



States having 28 attributes as they are calculate itself in different values.



As we can see that 3 states has been selected A & N Island , Chhattisgarh ,

Andhra Pradesh. In which First they are divided by Murder <= , In yes A & N island and In No Andhra State are divide , Second INSULT\_IN\_MOTESTY\_ OF\_WOMEN >= 2127 , Chhattisgarh and A & N Island are divide and this processes is going on. Different states with different crime will we distinguish.

# 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.

# Significance Of The Project

We expect that from our project we are able to draw useful inferences that will have potential to become useful for other researchers or crime investigation department to use it and to learn from the past trend and nature of crime rates, similarities among areas and set of crimes occur frequently. Can be used to improve crime management.

The models can be used to apply to the similar data but after 2012 whenever it is available in bulk from the data.gov.in.

# References

**Data set sources:**

<https://www.kaggle.com/rajanand/crime-in-india>

<https://data.gov.in/resources/district-wise-crime-under-various-sections-indian-penal-code-ipc-crimes-during-2001-2012>

**Machine Learning tutorials:**

**<https://www.udemy.com/course/machinelearning/>**

**Other guidance and knowledge:**

From respected Dr. Nilamadhab Mishrasir’s Advance Data Analytics I & II materials.