SUMMER TRAINING REPORT

ON APPLICATION OF R ON SOLVING STATISTICAL PROBLEMS

Project By

UTKARSH SINGH

UID:1605163

COMPUTER SCIENCE ENGINEERING DEPARTMENT, 5th SEMESTER,
KIIT UNIVERSITY,
BHUBANESWAR, ODISHA

Certificate

This is to certify that the Summer Project entitled on
“ ***APPLICATIONS OF R IN SOLVING STATISTICAL PROBLEMS***”
has been successfully completed by

UTKARSH SINGH
UID: 1605163, B.E. COMPUTER SCIENCE ENGINEERING - 5TH SEMESTER
DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
KIIT UNIVERSITY, BHUBANESWAR, ODISHA
Under the guidance of
DR. KOUSHIK MONDAL.

TRAINING IN CHARGE
OR
PROJECT IN CHARGE

DR. KOUSHIK MONDAL
SYSTEM MANAGER
OF COMPUTER CENTRE
IIT(ISM), DHANBAD

Acknowledgement

With profound gratitude, I would like to thanks to **DR. KOUSHIK MONDAL**, **System Manager of Computer Centre, IIT (ISM) , Dhanbad**, for his most valuable & inspiring guidance rendered throughout the summer training program.

He is also Training In-charge for this summer training. Special thanks to him for his helpful & special instruction to make this training a successful one. It is impossible to acknowledge sufficiently his important contribution of talent and time given unselfishly in proceeding with this work. I wish to convey my regard to him for helping me at each and every step in bringing out this report.

My sincere gratitude to the **HOD, Department Computer Science Engineering, KIIT University,** for giving me the opportunity to carry out summer training in IIT (ISM), Dhanbad.

I would also like to thank all the members associated with this training program, for giving me excellent laboratory facilities and their kind help whenever I nee

UTKARSH SINGH
UID: 1605163
DEPARTMENT OF COMPUTER SCIENCE,
KIIT UNIVERSITY,BHUBANESWAR
ODISHA

SPECIAL MENTION

This special mention is being awarded to
UTKARSH SINGH
For the successful development in summer project
"APPLICATIONS OF R IN SOLVING STATISTICAL PROBLEMS".

He is enthusiastic, creative & dedicated to work with helpful in nature.

We wish him good luck in all his future endeavors.

SYSTEM MANAGER
COMPUTER CENTRE
IIT (ISM) DHANBAD
DR. KOUSHIK MONDAL

CONTENTS

Predictive Analysis on Crimes Using Regression and Gradient Descent

B.TECH Undergraduate U. Singh

Department of Computer Science,KIIT University,Bhubaneswar,Odisha,India

Abstract

Abstracting useful information from a big data has always been a challenging task. Data mining is a powerful technology with great potential to extract knowledge based information from such data. Prediction can be done with past and related records in different fields. Risk and safety have always been an important consideration in field of crimes. Prediction of the crime rate in a particular city will help us to explore what factors of a city contribute to the crimes prevailing in a city and what measures can be ensured to take action against these crimes .So, that we can save a lot of life and cost. This paper proposes a crime prediction system with huge collection of past records by applying effective regression machine learning tools like Linear Regression, Non-Linear Regression, PCA(Principle Component Analysis), Support Vector Machine(SVM), Decision Tree(DT), Random Forest etc. because these algorithms have huge accuracy and precision to handle huge and noisy data. The methods used ,prove to handle noisy,unrelated and missing data.The prediction results are tabulated and ranges between 80% to 85%.

KEYWORDS: Regression; SVM; DT; Crime Prediction

1. INTRODUCTION

This work focus on using machine learning techniques in the process of crime prediction analysis with data about crimes in various cities in United States of America. Data from http://www.disastercenter.com/crime.htm, which records all the different kind of crimes that prevail in a particular city in a city its corresponding population and other relavent data which can be efficient in prediction of the crimes prevailing in the particular city in a particular year. Here by prediction we imply that trend which can be deduced from the dataset. Various attributes that caused the crime are analyzed using various statistical tools. Collectively a 13 attribute dataset with crimes for 51 cities over a period of 1960 to 2012 was recorded, different crimes for a particular year and particular city was entered in the dataset. This data had some errors and missing values.

On the dataset Regression was applied then after using the p-value and Adjusted Sum of Squares we remove attributes that contributed least in the prediction of the model and the applied different models such as linear regression, SVM, Decision Tree etc to the most contributing attributes.

2. PROBLEM DEFINATION

Understanding and processing an unstructured and dynamic data is a tedious work. Crimes data are dynamic and seems to be unpredictable.Crimes can be caused due to many reasons such as poverty, unemployment, etc.Thus we need to find the attributes that contribute to the crimes in a city.Thus to do this to the unpredictable data we need tools and techniques.

3. DATASET SELECTION

Data from the website(http://www.disastercenter.com/crime.htm) is used. Thousands of records are collected over a period of 1960 and 2012 for 51 cities across USA. Population, violence, property along with different types of crimes were recorded for analysis purposes.

The following attributes values are used as training data set, year, population, property, violent, Murder, Forcible. Rape, Robbery, Aggravated assault, Burglary, Larceny Theft, Vehicle Theft, abbr, state. Here abbr is the abbreviation of the names of each state, so we can remove state and keep abbr instead of it.

4. TOOLS USED

We can use many machine learning tools like R, Python, octave etc. But, here we will use R due to its statistically powerful tools and libraries which make it efficient tool for machine learning algorithm implementation.

5. WHY R ?

We are using R in this project due to its following advantages :-

(i) R is the most comprehensive statistical analysis package available.

(ii) It incorporates all the statistical tools and tests for model analysis and prediction.

(iii) R is also quiet efficient in data manipulation and dealing with unprocessed data

(iv) R has no license restriction, it is open source for anyone to use and modify.

(v) Also R gives a nice set of graphical tools.

6. BASIC COMANDS IN R

For assigning a variable:-

```
>>a=1

>>b=2

>>c<-5
```

For assigning a vector

```
>>d<-c(1,2,3,4,5)

>>d= c(1,2,3,4,5)
```

Both will work

Now working with some of the arithmetic operations

```
>>e=d+d

>>e
```

Output: 2 4 6 8 10

```
>>length( e)
```

Output: 5

Generating regular sequences

```
>> seq(-5, 5, by=.2)
```

Creates a vector of numbers from -5 to 5 which have a difference of .2

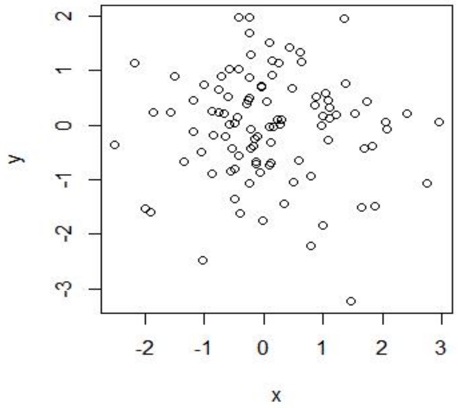
Loading a dataset:

```
>> dataset= read.csv(‘filename.csv’, sep=’ , ’)
```

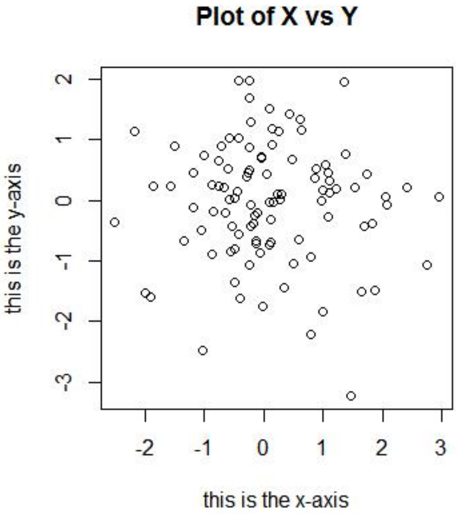
The dataset is being loaded in the dataset variable

Graphical Instructions:

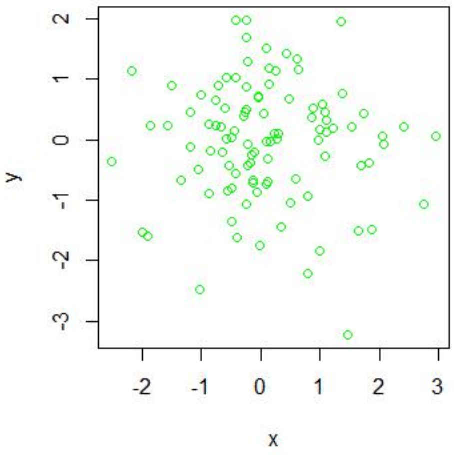
```
>>x=rnorm(100)

>>y=rnorm(100)

>>plot(x,y)
```



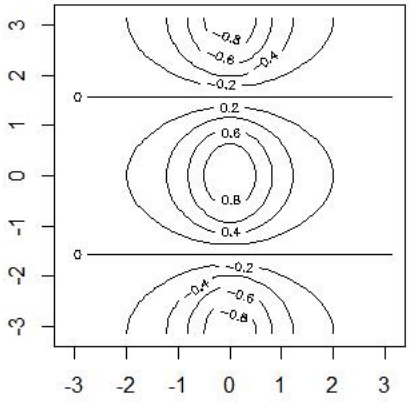
```
>>plot(x,y,xlab="this is the x-axis",ylab="this is the y-axis",main="Plot of X vs Y")
```



```
>>plot(x,y,col="green")
```



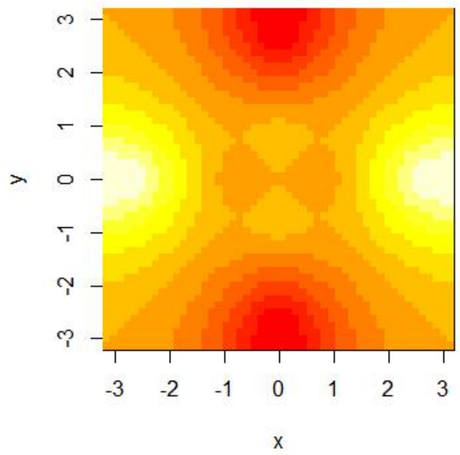
```
>>x=seq(-pi,pi,length=50)

>>y=x

>>f=outer(x,y,function(x,y)cos(y)/(1+x^2))

>>contour(x,y,f)
```



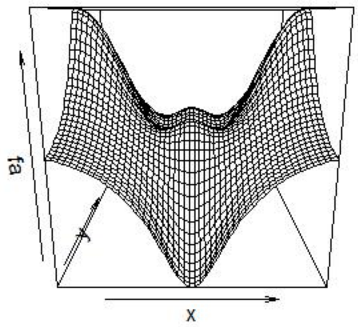
Here the contour and plot functions are the plotting functions

Some more plotting functions:

```
>> fa=(f-t(f))/2

>>image(x,y,fa)
```

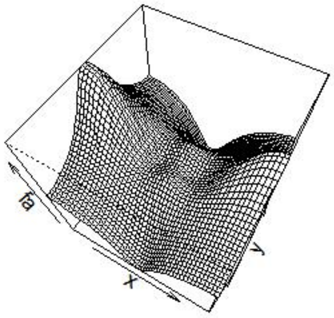


```
>>persp(x,y,fa)
```



We can adjust its view by using attributes like theta and phi

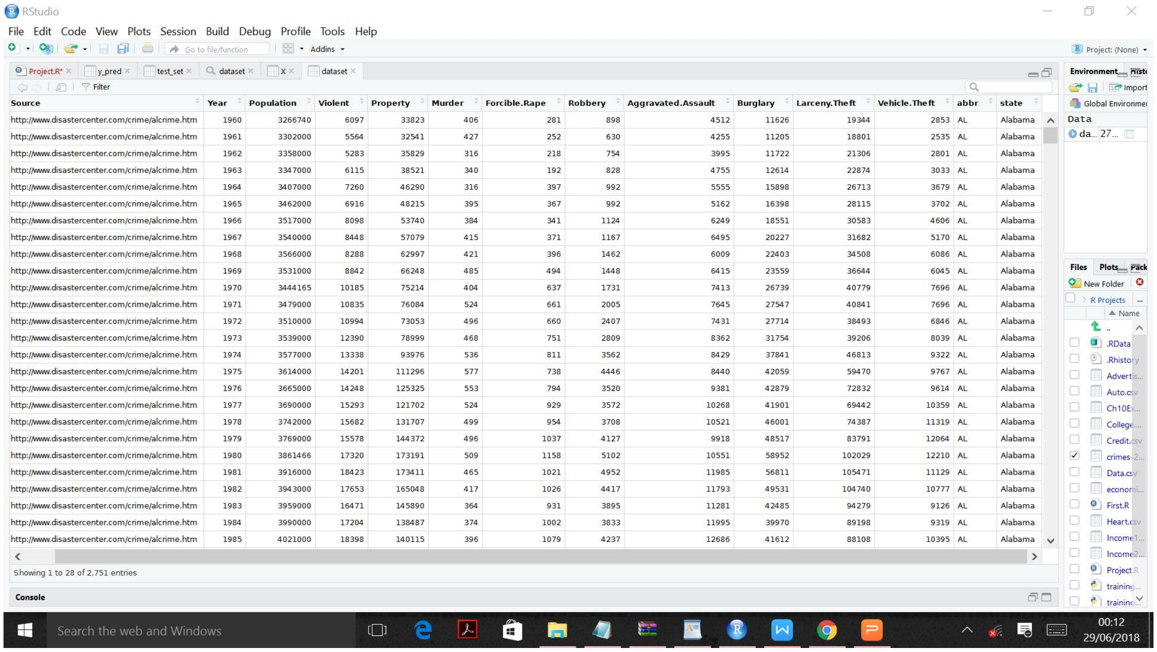
```
>>persp(x,y,fa,theta=30,phi=60)
```



Now after knowing little bit about the different syntax and functions of R we can begin with our analytic. A bit about the different statistical tools and analytic functions of R will be covered in the later sections as we do our project on the dataset.

7 BEGINNING WITH OUR DATASET

7.1 Loading the dataset



Source	Year	Population	Violent	Property	Murder	Forcible.Rape	Robbery	Aggravated.Assault	Burglary	Larceny.Theft	Vehicle.Theft	abbr	state
http://www.disastercenter.com/crime/alcrime.htm	1960	3266740	6097	33823	406	281	898	4512	11626	19344	2853	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1961	3302000	5564	32541	427	252	630	4255	11205	18801	2535	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1962	3358000	5283	35829	316	218	754	3995	11722	21306	2801	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1963	3347000	6115	38521	340	192	828	4755	12614	22874	3033	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1964	3407000	7260	46290	316	397	992	5555	15898	26713	3679	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1965	3462000	6916	48215	395	367	992	5162	16398	28115	3702	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1966	3517000	8098	53740	384	341	1124	6249	18551	30583	4606	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1967	3540000	8448	57079	415	371	1167	6495	20227	31682	5170	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1968	3566000	8288	62997	421	396	1462	6009	22403	34508	6086	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1969	3531000	8842	66248	485	494	1448	6415	23559	36644	6045	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1970	3444165	10185	75214	404	637	1731	7413	26739	40779	7696	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1971	3479000	10835	76084	524	661	2005	7645	27547	40841	7696	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1972	3510000	10994	73053	496	660	2407	7431	27714	38493	6846	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1973	3539000	12390	78999	468	751	2809	8362	31754	39206	8039	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1974	3577000	13338	93976	536	811	3562	8429	37841	46813	9322	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1975	3614000	14201	111296	577	738	4446	8440	42059	59470	9767	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1976	3665000	14248	125325	553	794	3520	9381	42879	72832	9614	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1977	3690000	15293	121702	524	929	3572	10268	41901	69442	10359	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1978	3742000	15682	131707	499	954	3708	10521	46001	74387	11319	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1979	3769000	15578	144372	496	1037	4127	9918	48517	83791	12064	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1980	3861466	17320	173191	509	1158	5102	10551	58952	102029	12210	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1981	3916000	18423	173411	465	1021	4952	11985	56811	105471	11129	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1982	3943000	17653	165048	417	1026	4417	11793	49531	104740	10777	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1983	3959000	16471	145890	364	931	3895	11281	42485	94279	9126	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1984	3990000	17204	138487	374	1002	3833	11995	39970	89198	9319	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1985	4021000	18398	140115	396	1079	4237	12686	41612	88108	10395	AL	Alabama

Showing 1 to 28 of 2,751 entries

RAW DATA

First step of any data analytic is to load our dataset which can be done in R in very easy manner

```
>>dataset=read.csv('crimes-2012.csv')
```

The dataset got loaded we can have a good look at it using the fix() function

```
>>fix(dataset)
```

A window opens and where u have the data in a tabular form ,here you can edit the data by mere clicking and typing.

7.2 Data Preprocessing

7.2.1 Dealing with missing values

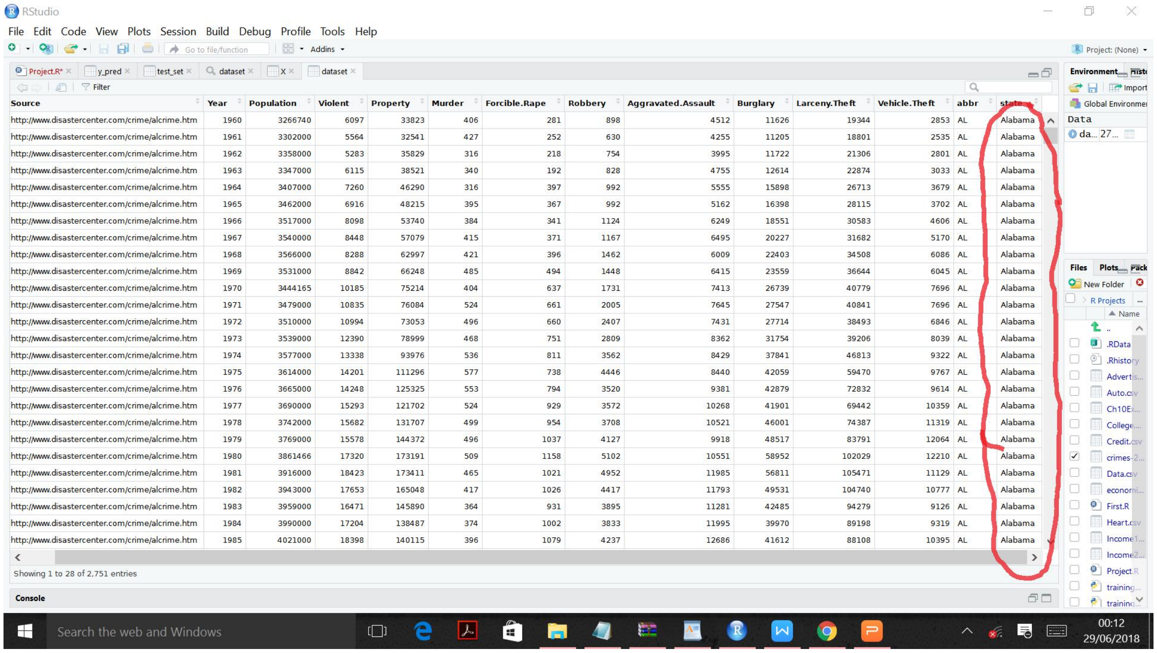
The first step of any data preprocessing step is to remove all the missing values present in the dataset. In R missing values are denoted by nan. Here is the way to deal with missing values present in the dataset.

```
>>dataset=na.omit(dataset)
```

After doing this the row containing the nan values disappears.

7.2.2 Label Encoding the data

The second step is to encode the data which are in the character or letter form and convert them to numbers. In this dataset the abbr has to be label encoded.



Source	Year	Population	Violent	Property	Murder	Forcible.Rape	Robbery	Aggravated.Assault	Burglary	Larceny.Theft	Vehicle.Theft	abbr	state
http://www.disastercenter.com/crime/alcrime.htm	1960	3266740	6097	33823	406	281	898	4512	11626	19344	2853	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1961	3302000	5564	32541	427	252	630	4255	11205	18801	2535	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1962	3358000	5283	35829	316	218	754	3995	11722	21306	2801	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1963	3347000	6115	38521	340	192	828	4755	12614	22874	3033	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1964	3407000	7260	46290	316	397	992	5555	15898	26713	3679	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1965	3462000	6916	48215	395	367	992	5162	16398	28115	3702	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1966	3517000	8098	53740	384	341	1124	6249	18551	30583	4606	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1967	3540000	8448	57079	415	371	1167	6495	20227	31682	5170	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1968	3566000	8288	62997	421	396	1462	6009	22403	34508	6086	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1969	3531000	8842	66248	485	494	1448	6415	23559	36644	6045	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1970	3444165	10185	75214	404	637	1731	7413	26739	40779	7696	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1971	3479000	10835	76084	524	661	2005	7645	27547	40841	7696	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1972	3510000	10994	73053	496	660	2407	7431	27714	38493	6846	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1973	3539000	12390	78999	468	751	2809	8362	31754	39206	8039	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1974	3577000	13338	93976	536	811	3562	8429	37841	46813	9322	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1975	3614000	14201	111296	577	738	4446	8440	42059	59470	9767	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1976	3665000	14248	125325	553	794	3520	9381	42879	72832	9614	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1977	3690000	15293	121702	524	929	3572	10268	41901	69442	10359	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1978	3742000	15682	131707	499	954	3708	10521	46001	74387	11319	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1979	3769000	15578	144372	496	1037	4127	9918	48517	83791	12064	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1980	3861466	17320	173191	509	1158	5102	10551	58952	102029	12210	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1981	3916000	18423	173411	465	1021	4952	11985	56811	105471	11129	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1982	3943000	17653	165048	417	1026	4417	11793	49531	104740	10777	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1983	3959000	16471	145890	364	931	3895	11281	42485	94279	9126	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1984	3990000	17204	138487	374	1002	3833	11995	39970	89198	9319	AL	Alabama
http://www.disastercenter.com/crime/alcrime.htm	1985	4021000	18398	140115	396	1079	4237	12686	41612	88108	10395	AL	Alabama

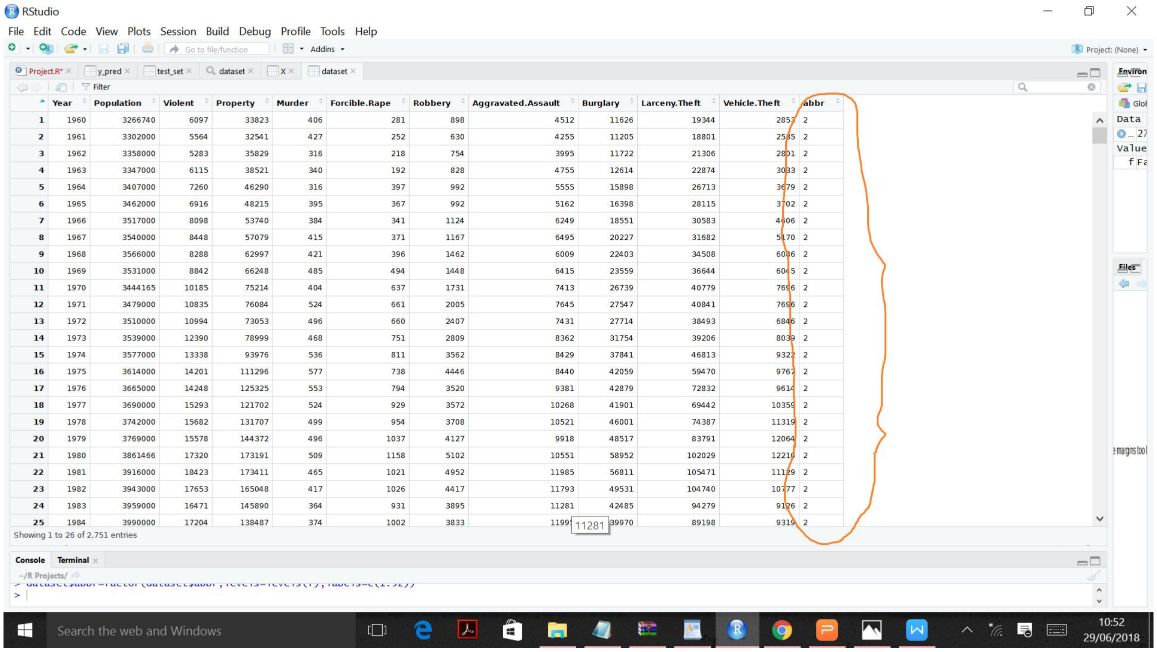
Showing 1 to 28 of 2,751 entries

Before encoding

```
>>f=factor(dataset$abbr)
```

```
>>levels(f)
```

```
>>dataset$abbr=factor(dataset$abbr,levels=levels(f),labels=c(1:52))
```



	Year	Population	Violent	Property	Murder	Forcible.Rape	Robbery	Aggravated.Assault	Burglary	Larceny.Theft	Vehicle.Theft	abbr
1	1960	3266740	6097	33823	406	281	898	4512	11626	19344	2853	2
2	1961	3302000	5564	32541	427	252	630	4255	11205	18801	2535	2
3	1962	3358000	5283	35829	316	218	754	3995	11722	21306	2801	2
4	1963	3347000	6115	38521	340	192	828	4755	12614	22874	3033	2
5	1964	3407000	7260	46290	316	397	992	5555	15898	26713	3679	2
6	1965	3462000	6916	48215	395	367	992	5162	16398	28115	3702	2
7	1966	3517000	8098	53740	384	341	1124	6249	18551	30583	4606	2
8	1967	3540000	8448	57079	415	371	1167	6495	20227	31682	5170	2
9	1968	3566000	8288	62997	421	396	1462	6009	22403	34508	6036	2
10	1969	3531000	8842	66248	485	494	1448	6415	23559	36644	6045	2
11	1970	3444165	10185	75214	404	637	1731	7413	26739	40779	7696	2
12	1971	3479000	10835	76084	524	661	2005	7645	27547	40841	7696	2
13	1972	3510000	10994	73053	496	660	2407	7431	27714	38493	6846	2
14	1973	3539000	12390	78999	468	751	2809	8362	31754	39206	8039	2
15	1974	3577000	13338	93976	536	811	3562	8429	37841	46813	9322	2
16	1975	3614000	14201	111296	577	738	4446	8440	42059	59470	9767	2
17	1976	3665000	14248	125325	553	794	3520	9381	42879	72832	9614	2
18	1977	3690000	15293	121702	524	929	3572	10268	41901	69442	10359	2
19	1978	3742000	15682	131707	499	954	3708	10521	46001	74387	11319	2
20	1979	3769000	15578	144372	496	1037	4127	9918	48517	83791	12064	2
21	1980	3861466	17320	173191	509	1158	5102	10551	58952	102029	12210	2
22	1981	3916000	18423	173411	465	1021	4952	11985	56811	105471	11129	2
23	1982	3943000	17653	165048	417	1026	4417	11793	49531	104740	10777	2
24	1983	3959000	16471	145890	364	931	3895	11281	42485	94279	9126	2
25	1984	3990000	17204	138487	374	1002	3833	1199	39970	89198	9319	2

Showing 1 to 26 of 2,751 entries

```
~/R Projects/
> dataset$abbr=factor(dataset$abbr,levels=levels(f),labels=c(1:52))
> 
```

After encoding

So following these steps our dataset will be cleaned and now it is ready to be used for our analysis.

7.3 Fitting our first linear model and analyzing the model

So here we will start with our linear regression model. Now before fitting the data what we need to do is to summarize all the crimes prevailing in a city in a particular year and create a column named Total_crimes

```
>>dataset$total_crime=
dataset$Murder+dataset$Forcible.Rape+dataset$Robbery+dataset$Aggravated.
Assault+dataset$Burglary+dataset$Larceny.Theft+dataset$Vehicle.Theft
```

This will create a new column named total_crime and it will sum up all the crimes and fill the column

Now after doing this second step is to split the dataset into Dependent and Independent Variables.In our case the year, population, property, violent and abbr are the independent variables and the total_crime column is the dependent variable.

```
>>X=dataset[,1:4]

>>X$abbr= dataset$abbr

>>Y=dataset$total_crime
```

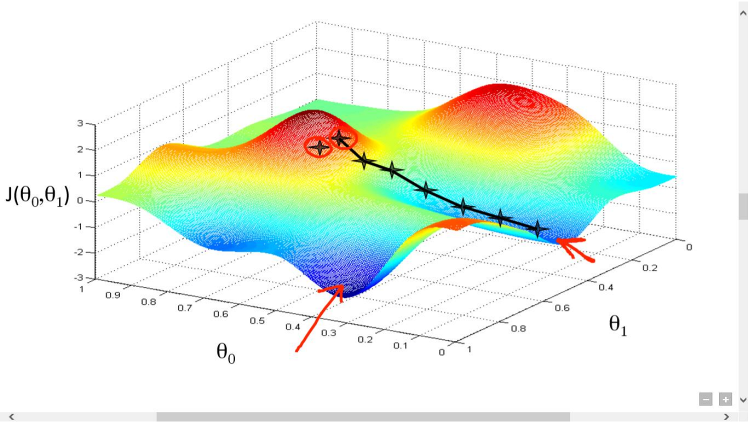
So this will create a matrix of independent variables as ‘X’ and vector of independent variables ‘Y’.

Now at last we are ready to fit our first real life model on the dataset available for fitting the model we have a function[lm( )] which will help us fit a model in an effective manner. The fitting of the model works on basis that for any set of attributes the value of coefficients should be such that the Residual Sum of Squares become very less.

Residual Sum of Squares= $\sum (y\_predicted-y\_true)^2$

GRADIENT DESCENT

For this purpose we can use a lot of algorithms such as gradient descent and others. Gradient descent algorithm is simple that to choose a set of coefficients let say L(θ) so that it reduces RSS, we will choose a starting initial values of the coefficient and then from that point we will descent using the steepest slope and get to the optimal set of coefficient which reduces the values of the RSS(Residual Sum of Squares).



Here in this diagram we can clearly see J(θ_0,θ_1) as a function of RSS parametrized by θ_0 and θ_1 and the black path shows the path by which it will descent to the optimal value.

Now lets move on to the programming part of the dataset

Now we split the dataset into two parts the training and the test set and then fit our model according to the data. This can be done using ”caTools” library and here is the code

```
>>library(caTools)
```

This function imports the caTools library

```
>>set.seed(123)
```

This will set the random values for the seed constant

```
>>split=sample.split(dataset$total_crime, SplitRatio=0.8)
```

This will split the test set into 20% test set and 80% training set

```
>>training_set=subset(dataset, split=TRUE)
```

This will split the whole dataset on the basis of the split set made before.

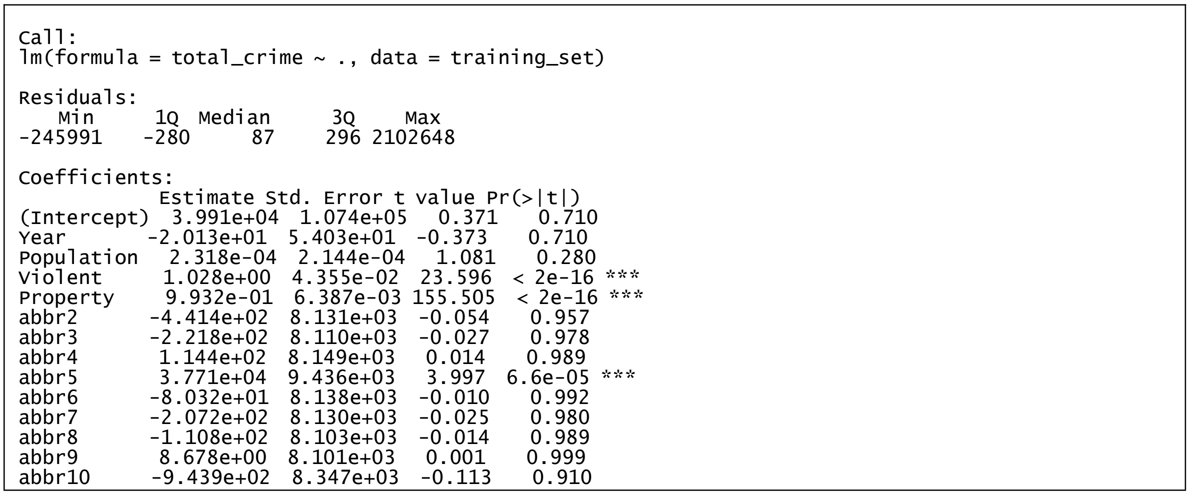
```
>>test_set=subset(dataset, split=FALSE)
```

So, now we will create a regressor to fit the data and fit the dataset using multiple linear regression. Multiple because the we have more than one attribute and linear as we assume that the dependent and independent variables have a linear relationship.

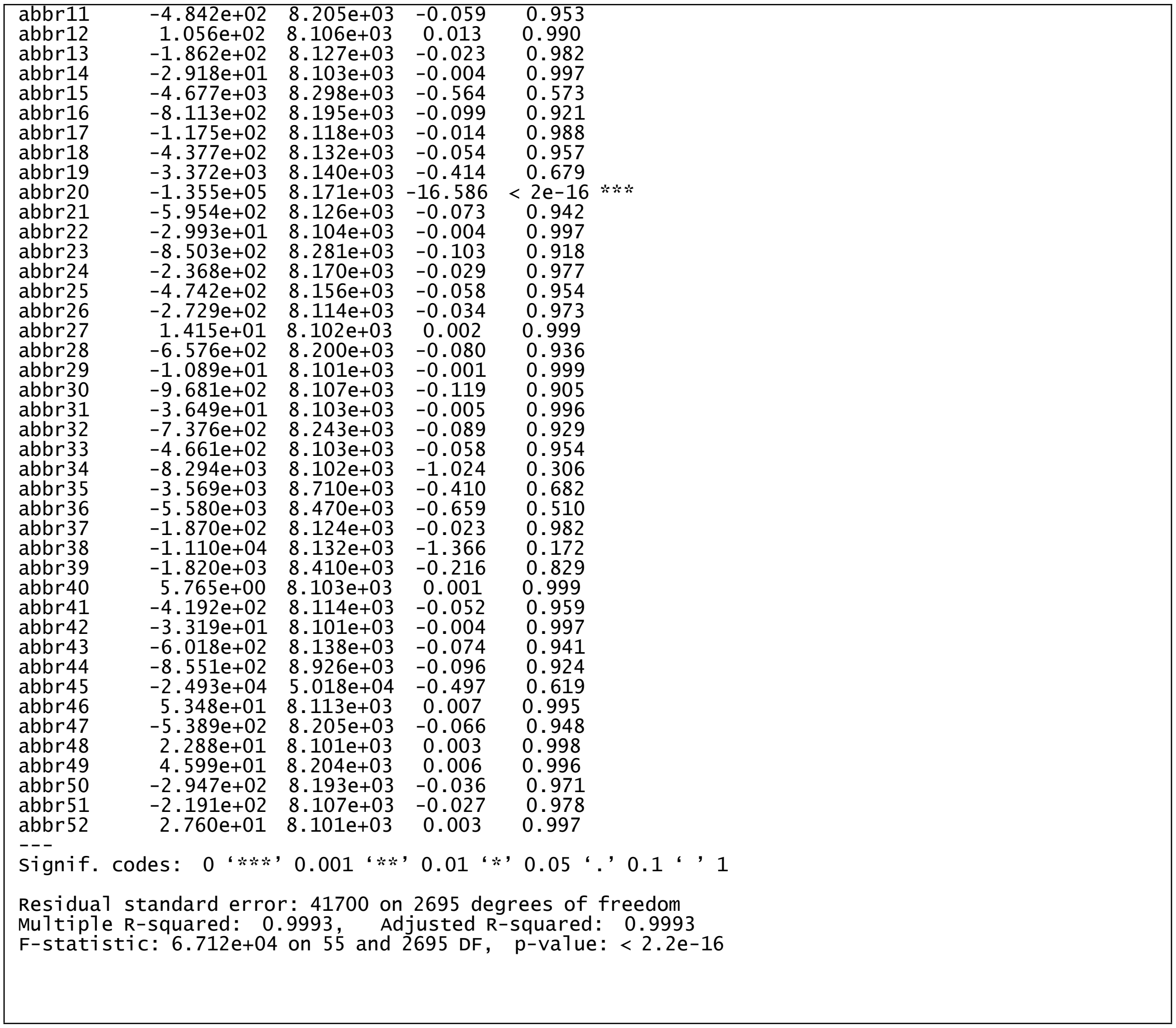
```
>>regressor=lm(dataset$total_crime~.,data=dataset)
```

The ‘ . ’ sign shows that all the attributes are to be taken apart from total_crime.

Now we will analyze the regressor using the summary() function.



```
Call:
lm(formula = total_crime ~ ., data = training_set)

Residuals:
   Min     1Q  Median     3Q     Max 
-245991   -280     87    296 2102648 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.991e+04  1.074e+05   0.371    0.710    
Year        -2.013e+01  5.403e+01  -0.373    0.710    
Population   2.318e-04  2.144e-04   1.081    0.280    
Violent      1.028e+00  4.355e-02  23.596  < 2e-16 ***
Property     9.932e-01  6.387e-03 155.505  < 2e-16 ***
abbr2       -4.414e+02  8.131e+03  -0.054    0.957    
abbr3       -2.218e+02  8.110e+03  -0.027    0.978    
abbr4        1.144e+02  8.149e+03   0.014    0.989    
abbr5        3.771e+04  9.436e+03   3.997  6.6e-05 ***
abbr6       -8.032e+01  8.138e+03  -0.010    0.992    
abbr7       -2.072e+02  8.130e+03  -0.025    0.980    
abbr8       -1.108e+02  8.103e+03  -0.014    0.989    
abbr9        8.678e+00  8.101e+03   0.001    0.999    
abbr10      -9.439e+02  8.347e+03  -0.113    0.910    
```



```
abbr11      -4.842e+02  8.205e+03  -0.059    0.953
abbr12       1.056e+02  8.106e+03   0.013    0.990
abbr13      -1.862e+02  8.127e+03  -0.023    0.982
abbr14      -2.918e+01  8.103e+03  -0.004    0.997
abbr15      -4.677e+03  8.298e+03  -0.564    0.573
abbr16      -8.113e+02  8.195e+03  -0.099    0.921
abbr17      -1.175e+02  8.118e+03  -0.014    0.988
abbr18      -4.377e+02  8.132e+03  -0.054    0.957
abbr19      -3.372e+03  8.140e+03  -0.414    0.679
abbr20      -1.355e+05  8.171e+03 -16.586  < 2e-16 ***
abbr21      -5.954e+02  8.126e+03  -0.073    0.942
abbr22      -2.993e+01  8.104e+03  -0.004    0.997
abbr23      -8.503e+02  8.281e+03  -0.103    0.918
abbr24      -2.368e+02  8.170e+03  -0.029    0.977
abbr25      -4.742e+02  8.156e+03  -0.058    0.954
abbr26      -2.729e+02  8.114e+03  -0.034    0.973
abbr27       1.415e+01  8.102e+03   0.002    0.999
abbr28      -6.576e+02  8.200e+03  -0.080    0.936
abbr29      -1.089e+01  8.101e+03  -0.001    0.999
abbr30      -9.681e+02  8.107e+03  -0.119    0.905
abbr31      -3.649e+01  8.103e+03  -0.005    0.996
abbr32      -7.376e+02  8.243e+03  -0.089    0.929
abbr33      -4.661e+02  8.103e+03  -0.058    0.954
abbr34      -8.294e+03  8.102e+03  -1.024    0.306
abbr35      -3.569e+03  8.710e+03  -0.410    0.682
abbr36      -5.580e+03  8.470e+03  -0.659    0.510
abbr37      -1.870e+02  8.124e+03  -0.023    0.982
abbr38      -1.110e+04  8.132e+03  -1.366    0.172
abbr39      -1.820e+03  8.410e+03  -0.216    0.829
abbr40       5.765e+00  8.103e+03   0.001    0.999
abbr41      -4.192e+02  8.114e+03  -0.052    0.959
abbr42      -3.319e+01  8.101e+03  -0.004    0.997
abbr43      -6.018e+02  8.138e+03  -0.074    0.941
abbr44      -8.551e+02  8.926e+03  -0.096    0.924
abbr45      -2.493e+04  5.018e+04  -0.497    0.619
abbr46       5.348e+01  8.113e+03   0.007    0.995
abbr47      -5.389e+02  8.205e+03  -0.066    0.948
abbr48       2.288e+01  8.101e+03   0.003    0.998
abbr49       4.599e+01  8.204e+03   0.006    0.996
abbr50      -2.947e+02  8.193e+03  -0.036    0.971
abbr51      -2.191e+02  8.107e+03  -0.027    0.978
abbr52       2.760e+01  8.101e+03   0.003    0.997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41700 on 2695 degrees of freedom
Multiple R-squared:  0.9993,    Adjusted R-squared:  0.9993
F-statistic: 6.712e+04 on 55 and 2695 DF,  p-value: < 2.2e-16
```

The above table gives the summary of the regressor.

Statistical Analysis of the summary and FEATURE ELIMINATION

Before analyzing the summary we need to know some of the statistical terminologies that we are going to discuss in the following part. Some of which are listed below:-

(i)p-value

(ii)Adjusted sum of squares

(iii) Confidence interval

(i) P-Value

The **P value**, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested

(ii) Adjusted sum of squares

The **adjusted sum of squares** for a term is the increase in the regression **sum of squares** compared to a model with only the other terms. It quantifies the amount of variation in the response data that is **explained** by each term in the model. Adj SS Error. The error **sum of squares** is the **sum** of the **squared** residuals.

(iii)Confidence interval

In **statistics**, a **confidence interval** (CI) is a type of **interval** estimate, computed from the **statistics** of the observed data, that might contain the true value of an unknown population parameter. ... Most commonly, the 95% **confidence** level is used. However, other **confidence** levels can be used, for example, 90% and 99%.

NOW ANALYZING

Now lets start analyzing the results we got from regressor and then deduce the contribution of each of the attributes and eliminate the factors which are least significant. Hence we need to first select a significance level which will decide the contributing attributes in the model. Now lets assume we select the significance level at 5% and now attributes having p-value less than this significant level gets eliminated. For doing this we apply subset selection method, to be more precise stepwise subset selection method as it is quiet efficient.Here we can see that all the attributes except population and violence have a p-value above significance level.So population and violent column has a huge impact on the model whereas other attributes contribute less.

So we use the backward elimination technique to remove the unwanted attributes that do not contribute much in the model

BACKWARD ELIMINATION ALGO:-

STEP 1:-Select a significance level to stay in the model.(for e.g 0.05)

STEP 2:-Fit the full model with all the possible attributes.

STEP 3:-Consider the attributes with highest p-value .If p-value of the attribute greater than significance level go to step 4 or otherwise finish.

STEP 4:- Remove the selected attribute.

STEP 5:- Refit the model with remaining attributes and repeat step 3.

FINISH

Now applying the algorithm on the model

So firstly we remove the abbr attribute as it has a huge p-value

```
>>regressor=lm(total_crime~.-abbr,data=training_set)

>>summary(regressor)
```

```
7
Call:
lm(formula = total_crime ~ . - abbr, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max 
-247680    2098    2546    2941 2144533 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.969e+04  1.152e+05   0.345    0.730    
Year        -2.131e+01  5.800e+01  -0.367    0.713    
Population   8.276e-05  1.061e-04   0.780    0.436    
Violent      1.053e+00  3.238e-02  32.528   <2e-16 ***
Property     9.917e-01  5.124e-03 193.531   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45680 on 2746 degrees of freedom
Multiple R-squared:  0.9991,	Adjusted R-squared:  0.9991 
F-statistic: 7.688e+05 on 4 and 2746 DF,  p-value: < 2.2e-16
```

Now the Year has a huge p-value we remove Year and refit the model

```
>>regressor=lm(total_crime~.-Year-abbr,data=training_data)

>>summary(regressor)
```

```
Call:
lm(formula = total_crime ~ . - Year - abbr, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max 
-248110    2336    2635    2751 2144326 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.631e+03  9.046e+02  -2.909  0.00366 ** 
```

```
Population   8.242e-05  1.061e-04   0.777  0.43735    
Violent     1.051e+00  3.187e-02  32.989  < 2e-16 ***
Property    9.920e-01  5.071e-03 195.612  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45680 on 2747 degrees of freedom
Multiple R-squared:  0.9991,   Adjusted R-squared:  0.9991 
F-statistic: 1.025e+06 on 3 and 2747 DF,  p-value: < 2.2e-
```

Now we can see that Population has the highest p-value

So remove Population and refit the model.

>>regressor=lm(total_crime~.-abbr-Year-Population,data=training_set)

>>summary(regressor)

```
Call:
lm(formula = total_crime ~ . - abbr - Year - Population, data = training_
set)

Residuals:
    Min      1Q  Median      3Q     Max 
-248083    2320    2617    2727 2144936 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.581e+03  9.021e+02   -2.86  0.00426 ** 
Violent      1.049e+00  3.175e-02   33.05  < 2e-16 ***
Property     9.943e-01  4.152e-03  239.50  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45670 on 2748 degrees of freedom
Multiple R-squared:  0.9991,   Adjusted R-squared:  0.9991 
F-statistic: 1.538e+06 on 2 and 2748 DF,  p-value: < 2.2e-16
```

Now our model is ready and has a p-value less than the significance level so our model is well fit with attributes property and violent.

Predicting the values of the test set

Now, we will predict the values using our model

>>y_pred=predict (regressor, newdata=test_set)

Output will be the response of new dataset based on the model that has been created from the training set.

And when we check the results it nearly matches the real dependent response values that were provided to us in the dataset.

```
Call:
lm(formula = total_crime ~ . - abbr - Year - Population, data = training_
set)

Residuals:
    Min      1Q  Median      3Q     Max 
-248083    2320    2617    2727 2144936 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.581e+03  9.021e+02   -2.86  0.00426 ** 
Violent      1.049e+00  3.175e-02   33.05  < 2e-16 ***
Property     9.943e-01  4.152e-03  239.50  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45670 on 2748 degrees of freedom
Multiple R-squared:  0.9991,	Adjusted R-squared:  0.9991 
F-statistic: 1.538e+06 on 2 and 2748 DF,  p-value: < 2.2e-16
```

FINAL SUMMARY

CONCLUSION

1. Here we used various programming techniques and algorithms in the powerful statistical programming language R to achieve our goal.

2. Also from the following dataset we concluded that the total crime prevailing in a city in a particular year is a linear function of its property and violence attribute.

3. Hence, we can ensure less crime by stopping violent behaviour in a particular city ,also we conclude that area having a lot of property in it gives an impact on the society.

4.Also both these attributes property and violent have positive coefficients .For property(9.943e-01) and for violent (1.049e+00) also the intercept of the equation is (-2.581e+03) and hence the multiple linear regression becomes as follows

TOTAL_CRIME=9.943e-01(PROPERTY)+1.049e+00(VIOLENT)-2.581e+03

5.So looking at the regression formula we conclude that according to my hypothesis,if we stop crimes at an early level in a locality it will decrease the total crimes caused in the locality

REFERENCES

- An Introduction to Statistical Learning with application in R.(BY:- Gareth James , Daniela Witten ,Trevor Hastie, Robert Tibshirani)
- www.CARN.org
- www.google.com
- www.wikipedia.org
- Introduction to R (from CARN.org)

THANK YOU