A project report on

COMPARITIVE AND PREDICTIVE ANALYSIS OF CRIME IN FIRST WORLD AND THIRD WORLD COUNTRIES

(INDIA vs. USA)

Submitted in partial fulfilment for the award of the degree of

BSc Computer Science

By

PREETI RACHEL JASPER (17BCS0003)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SITE

APRIL, 2020

A project report on

COMPARITIVE AND PREDICTIVE ANALYSIS OF CRIME IN FIRST WORLD AND THIRD WORLD COUNTRIES

(INDIA vs. USA)

Submitted in partial fulfilment for the award of the degree of

BSc Computer Science

By

PREETI RACHEL JASPER (17BCS0003)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SITE

APRIL, 2020

DECLARATION

I hereby declare that the thesis entitled "COMPARITIVE AND PREDICTIVE ANALYSIS OF CRIME IN FIRST WORLD AND THIRD WORLD COUNTRIES" submitted by me, for the award of the degree of BSc Computer Science, VIT is a record of the bona fide work carried out by me under the supervision of Prof Chemmalar Selvi G.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Preeti

Date: 22/5/2020

Signature of the Candidate

CERTIFICATE

This is to certify that the thesis entitled "COMPARITIVE AND PREDICTIVE ANALYSIS OF CRIME IN FIRST WORLD AND THIRD WORLD COUNTRIES" submitted by PREETI RACHEL JASPER (17BCS0003), SITE, VIT, for the award of the degree of BSc Computer Science is a record of bona fide work carried out by him/her under my supervision.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The Project report fulfils the requirements and regulations of VIT and in my opinion meets the necessary standards for submission.

Signature of the Guide **Signature of the HoD**

Internal Examiner **External Examiner**

ABSTRACT

According to the National Crime Records Bureau of India, a crime against women is committed every 3 minutes, and the rate of crime showed an increase of approximately 50% from the years 2011- 2015 (in four years). However, data studies show that crime against women per capita in India is low compared to other countries. This new found information helps us understand that maybe the increased rate of crime in India might be directly related to population growth India has experienced in the past decade. This poses the question, what will be the state of our country in future years? We strive for a safer country for our mothers and sisters and this project aims to provide insight regarding the question. Exploratory data analytics techniques will be used to assess the datasets and supervised learning techniques will be applied to draw a predictive conclusion.

Predictive analysis is concerned with the branch of data mining used to predict future patterns and trends. This modelling technique can be used to aid society. This research aims to foresee the crime patterns against women in India. In recent years, crime against women has skyrocketed and understanding past data can help us come up with insightful patterns that describe the current state of crime and assault in India. The datasets of the past years will be studied using extensive EDA (Exploratory Data Analysis) techniques that will help us understand the problems women and girls of India face. The data will then be cleaned and supervised learning techniques will be executed to predict future trends in crime rates with time. The project aims to help the women of India by using technological advancements in the area of data science to predict the future state of the country.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Mrs Chemmalar Selvi G, Assistant Professor, School of Information Technology and Engineering, Vellore Institute of Technology, for her constant guidance, continual encouragement, and understanding; more than all, she taught me patience in my endeavour. My association with her is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Data Analytics.

I would like to express my gratitude to G. Viswanathan, Sekar Viswanathan, Dr. Anand A. Samuel, Dr. S. Narayanan, and Dr. Balakrushna Tripathy, SITE, for providing with an environment to work in and for his inspiration during the tenure of the course.

In jubilant mood I express ingeniously my whole-hearted thanks to Dr. Malaserene. Head of Department, all teaching staff and members working as limbs of our university for their not-self-centered enthusiasm coupled with timely encouragements showered on me with zeal, which prompted the acquirement of the requisite knowledge to finalize my course study successfully. I would like to thank my parents for their support.

It is indeed a pleasure to thank my friends who persuaded and encouraged me to take up and complete this task. At last but not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly toward the successful completion of this project.

Preeti

Place: Vellore

Date: 22/5/2020 **Preeti Rachel Jasper**

CONTENTS

Page no

CONTENTS (contd) page no

LIST OF FIGURES

Page no

LIST OF FIGURES (contd)

LIST OF TABLES

Page no

LIST OF ABBREVATIONS

NCRB- National Crime Records Bureau of India

EDA- Exploratory Data Analysis

LR- Linear Regression

DBSCAN- Density-Based Spatial Clustering of Applications with Noise

PCA- Principal Component Analysis

USA- United States of America

OGD- Open Government Data

DF- Data Frame

CSV- Comma-Separated Values

JSON- JavaScript Object Notation

TR- Table Row

TD- Table Definition

TH-Table Head

DISP- Distribution Plot

HTML- Hypertext Markup Language

CSS- Cascading Style Sheets

Chapter 1

Introduction

1.1 MOTIVATION AND OBJECTIVE:

According to the National Crime Records Bureau of India, a crime against women is committed every 3 minutes, and the rate of crime showed an increase of approximately 50% from the years 2011- 2015 (in four years). However, data studies show that crime against women per capita in India is low compared to other countries. This new found information helps us understand that maybe the increased rate of crime in India might be directly related to population growth India has experienced in the past decade. This poses the question, what will be the state of our country in future years? We strive for a safer country for our mothers and sisters and this project aims to provide insight regarding the question. Exploratory data analytics techniques will be used to assess the datasets and supervised learning techniques will be applied to draw a predictive conclusion.

1.2 OVERVIEW

Predictive analysis is concerned with the branch of data mining used to predict future patterns and trends. This modelling technique can be used to aid society. This research aims to foresee the crime patterns against women in India. In recent years, crime against women has skyrocketed and understanding past data can help us come up with insightful patterns that describe the current state of crime and assault in India. The datasets of the past years will be studied using extensive EDA (Exploratory Data Analysis) techniques that will help us understand the problems women and girls of India face. The data will then be cleaned and supervised learning techniques will be executed to predict future trends in crime rates with time. The project aims to help the women of India by using technological advancements in the area of data science to predict the future state of the country

1.3 ADVANTAGES OF DATA ANALYTICS

Data analysis is the process of evaluating data using analytical and statistical tools to discover useful information and aid in decision making. There are a several data analysis methods including data mining, text analytics, business intelligence and data visualization.

- It detects and corrects the errors from data sets with the help of data cleansing. This helps in improving the quality of data and consecutively benefits both customers and institutions such as banks, insurance, and finance companies.

- It removes duplicate information from data sets and hence saves a large amount of memory space. This decreases the cost to the company.
- It helps in showing applicable notices on internet shopping sites dependent on notable information and buy conduct of the clients.
- It decreases banking dangers by recognizing likely false clients dependent on notable information examination. This helps establishments in choosing whether to give advance or charge cards to the candidates or not.
- It is utilized by security organizations for surveillance and checking reason dependent on information gathered by the colossal number of sensors.

Chapter 2

Literature Review

2.1 REVIEW OF PAST WORK:

Latest developments in predictive analytics include ‘Prediction of Crime Rate Using Data Clustering Technique’ where A. Anitha discusses the comparative study of various clustering algorithms that can be used to predict the rate of crime. The paper focused solely on the district of West Bengal and K means, Fuzzy C, DBSCAN and agglomerative methods were used to predict crime rates per district in West Bengal. The usage of unsupervised learning techniques in this case provides the reader with possible groups in which future crime may fall into. This work has fuelled our motivation to consider the accuracy provided by supervised learning techniques such as Linear Regression in predicting future crime rates in India compared to other countries, without limiting the study to just one state.’ Approach of Predictive Modelling on Crime Against Women Problem’, provides insights into the crime against women problem by using least squares simple Linear Regression problem. The author states that taking the population of the country into consideration while performing predictive analysis will provide more accurate results. This paper considers the various categories that come under the umbrella term of ‘Crime against Women’ and narrows down the increased rate of crime per year in areas such as ‘Domestic Violence and Cruelty by Husband or his relatives’. The study concluded that more women were coming forward with complaints in the present years which then led to a possible increase in the number of filed complaints in the past few years.

Both studies show an increase in reported cases from the past decade. In 2018, ‘Crime rate prediction using data clustering algorithms’ conducted a comparative study using K means and Fuzzy C clustering techniques on unstructured data including audio and video tapes to develop possible patterns in the nature of crime which can aid the law enforcements to predict future crime patterns and take necessary precautions. The research proposes in-depth studies focussing on smaller and static data sets which can help the user narrow down more accurate predictions. These studies help us understand the demand for predictive analysis to analyse the current state of India’s crime statistics. ‘Crime against Indian Women –Women Crime Susceptibility Indexes: A Principal Component Analysis’ uses PCA to narrow down the four major areas of Crime against women that affect the statistics the most. Cruelty by husband and relatives (45%), Assault with intent to outrage modesty (44%), Dowry Deaths (42%), Kidnapping and Abduction (38%) are the variables contributing to Principal component 1 and maximum variance. By providing these features more importance crime prediction and by using supervised learning techniques, we will aim to yield more accurate answers in our study.

2.2 OUR WORK:

We have the data for years 2001 to 2012 for crimes committed against women under different categories. In our work we use extensive EDA (Exploratory Data Analysis) to study the past data to understand the crime statistics against women in India. From the EDA we have understood the states with the highest and lowest rates of rape assaults in India. We provide line charts that help in visualizing the increase in rape from the year 2001 to 2012. We also made use of the Folium library to showcase the rapes per capita in different countries across the world. The data is cleared of all null values to prevent inaccurate results. Predictive analysis will then be done on the data to predict the future trends in crime against women with time. Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. Since the data that we have are labelled and structured, we have decided to use linear regression prediction methods to calculate the future crime rates against women with time.

RAPE STATISTICS BY COUNTRY 2020

Rape is unlawful sexual activity typically involving sexual intercourse done forcibly or under threat of injury against a person's will. Rape is a worldwide problem.

It is estimated that approximately 35% of women worldwide have experienced some form of sexual harassment in their lifetime. In the majority of countries that have data available on rape report that less than 40% of women who experience sexual violence seek help. Less than 10% seek help from law enforcement.

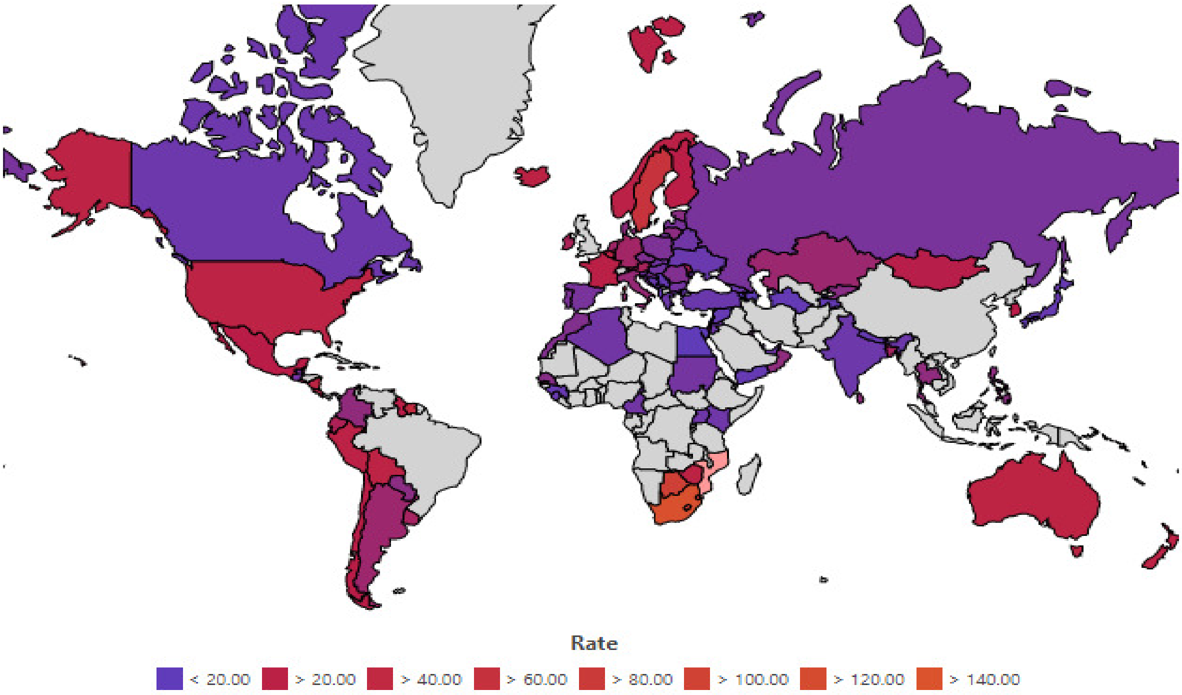
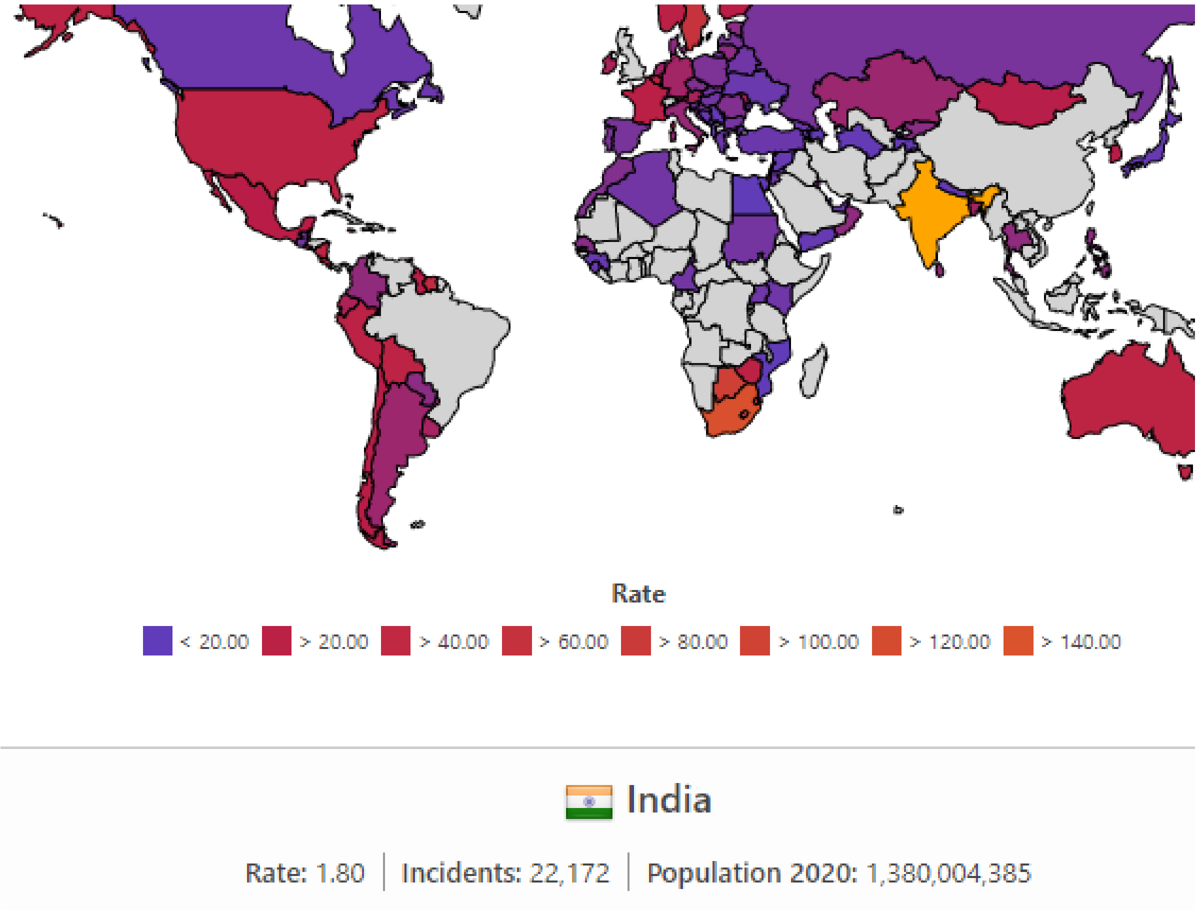


Figure 1

Because many women who experience sexual violence rarely report or come forward about their incidences, exact rape numbers are challenging to report. While many countries have laws against the act of sexual assault and violence, many of them are insufficient, inconsistent, and not systematically enforced. While people mostly hear about rape and sexual assault against women, men around the world also experience sexual harassment, sexual assault, and rape every day. Women ages 16-19 are four times more likely to be victims of rape or sexual assault and female college students ages 18-24 are three times more likely to experience sexual assault. Transgender people and those with disabilities are twice as likely to be victims of sexual assault or rape. In the United States, 70% of rape is committed by someone the victim knows.



Rate: 1.80 | Incidents: 22,172 | Population 2020: 1,380,004,385

Figure 2

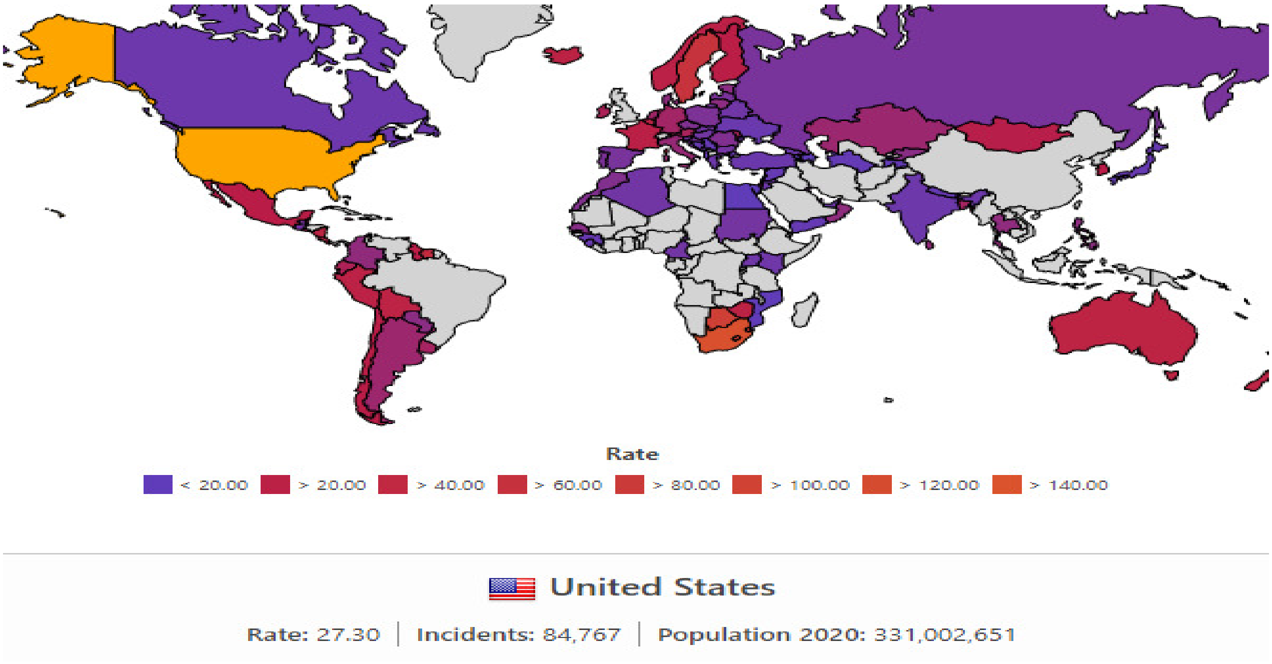
The United States has a rape rate(number of incidents per 100,000 citizens) of 27.3. As in many other countries, rape is grossly underreported in the United States due to victim shaming, fear of reprisal, fear of family knowing, cases not being taken seriously by law enforcement, and possible lack of prosecution for the perpetrator. Only 9% of rapists in the US get prosecuted and only 3% of rapists will spend a day in prison. 97% of rapists in the United States will walk free.

The ten countries with the highest rates of rape are:

- South Africa (132.4)
- Botswana (92.9)
- Lesotho (82.7)
- Swaziland (77.5)
- Bermuda (67.3)
- Sweden (63.5)
- Suriname (45.2)
- Costa Rica (36.7)
- Nicaragua (31.6)
- Grenada (30.6)

India is nowhere to be seen among these countries. However, South Africa which is another third world country continues to have high rates of rape. This makes us question whether economy has any role to play in the high crime rates against women in different countries

South Africa has the highest rate of rape in the world of 132.4 incidents per 100,000 people. According to a survey conducted by the South African Medical Research Council, approximately one in four men surveyed admitted to committing rape. Although the Parliament of South Africa enacted the Criminal Law (Sexual Offences and Related Matters) Amendment Act in 2007 attempting to amend and strengthen all laws dealing with sexual violence, the rates of reported rape, sexual abuse of children and domestic violence have continued to rise.



United States

Rate: 27.30 | Incidents: 84,767 | Population 2020: 331,002,651

Figure 3

INTERESTING OBSERVATIONS ABOUT CRIME AND RAPE RATES

- South Africa has had the highest rape rate since 2004.
- Sweden ranked first for rape rate amongst European Union in 2010.
- All of the top 3 countries by rape rate are Sub-Saharan African.
- Belgium ranked first for rape rate amongst NATO countries in 2010.
- Australia ranked second for rape rate amongst High income OECD countries in 2010.
- All of the top 6 countries by rape rate are Christian.
- United States ranked third for rape rate amongst Cold countries in 2010.
- France ranked second for rape rate amongst Eurozone in 2009.
- Iceland ranked third for rape rate amongst Europe in 2009.
- Moldova ranked first for rape rate amongst Eastern Europe in 2010.

Chapter 3

Design and Methodology

3.1 METHODOLOGY:

Data analytics is the science of analyzing raw data in order to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system.

The process involved in data analysis involves several different steps:

1. The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.
2. The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.
3. Once the data is collected, it must be organized so it can be analyzed. Organization may take place on a spreadsheet or other form of software that can take statistical data.
4. The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analyzed.
5. Predictive analysis must be done on cleaned data to get the most accurate results.

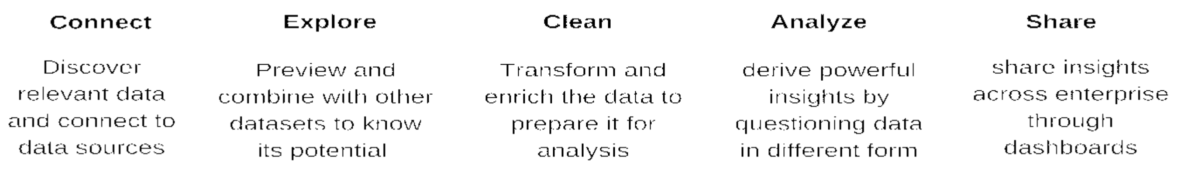

Connect
Discover relevant data and connect to data sources
Explore
Preview and combine with other datasets to know its potential
Clean
Transform and enrich the data to prepare it for analysis
Analyze
derive powerful insights by questioning data in different form
Share
share insights across enterprise through dashboards

Figure 4

3.2 DATASET SELECTION

For this research we have used six datasets. All are from Open Government Data (OGD) Platform India (data.gov.in).

1. Crime against Women during 2001-2012
2. World crime against women statistics

3. District-wise crimes committed against Women during 2001-2012
4. Age and sex wise persons arrested under crime against women during 2012
5. Persons arrested under crime against Women during 2001-2012.
6. USA rape statistics 2000-2015
7. Json files for world maps

Datasets 1 is contributed by Ministry of Home Affairs, Department of States, National Crime Records Bureau (NCRB). Datasets 3 and 5 are contributed by Ministry of Home Affairs, Department of States, National Crime Records Bureau (NCRB).

3.3 EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

- Maximize insight into a data set
- Uncover underlying structure
- Extract important variables
- Detect outliers and anomalies
- Test underlying assumptions
- Develop parsimonious models
- Determine optimal factor settings.

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data.

3.4 DATA CLEANING

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results. The nulls of the dataset will be removed so that we may get accurate predictions. Predictive analysis must be done on cleaned data to get the most accurate results.

3.5 PREDICTIVE ANALYSIS

Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. Since the data that we have are labelled and structured, we have decided to use linear regression prediction methods to calculate the future crime rates against women with time.

3.6 LINEAR REGRESSION USED FOR PREDICTIVE ANALYSIS

Data analysts often use a linear relationship to predict the (average) numerical value of Y for a given value of X using a straight line (called the regression line). If you know the slope and the y-intercept of that regression line, then you can plug in a value for X and predict the average value for Y. In other words, you predict (the average) Y from X.
Simple linear regression makes use of the formula y= mx +c. By calculating the values of the slope and Y-intercept, we can predict the value of the dependent variable using the known values of the independent variable.

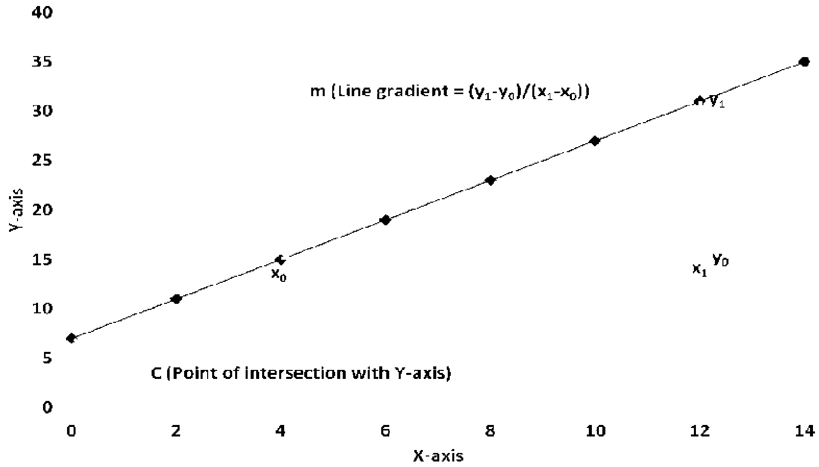


Figure 5

3.7 K MEANS

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k means algorithm works is as follows:

- Specify number of clusters K.
- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).

- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

3.8 ARCHITECTURE AND DESIGN

3.8.1 FLOWCHART:

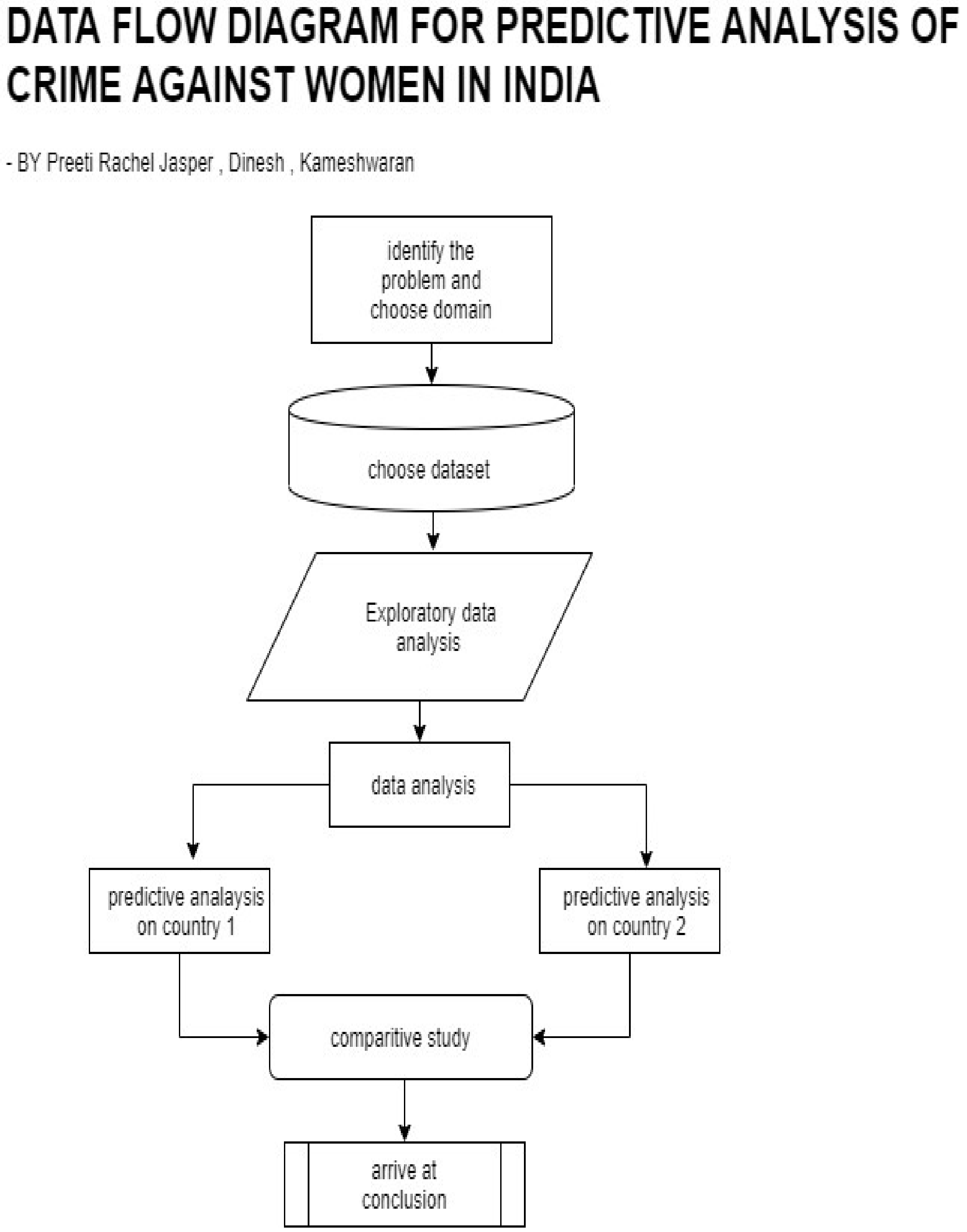

DATA FLOW DIAGRAM FOR PREDICTIVE ANALYSIS OF CRIME AGAINST WOMEN IN INDIA
- BY Preeti Rachel Jasper , Dinesh , Kameshwaran
identify the problem and choose domain
choose dataset
Exploratory data analysis
data analysis
predictive analaysis on country 1
predictive analysis on country 2
comparitive study
arrive at conclusion

Figure 6

3.8.2 USE CASE DIAGRAM

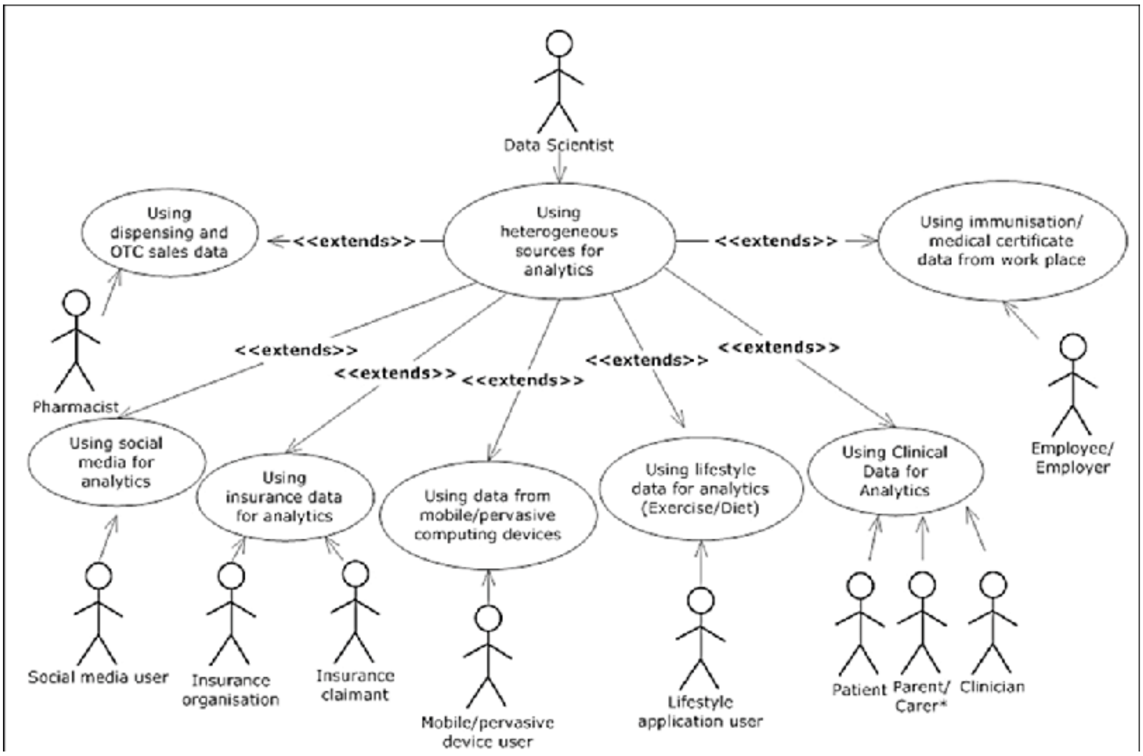

Data Scientist
Using heterogeneous sources for analytics
Using dispensing and OTC sales data
<<extends>>
Using immunisation/ medical certificate data from work place
<<extends>>
Pharmacist
<<extends>>
<<extends>>
<<extends>>
<<extends>>
<<extends>>
Employee/ Employer
Using social media for analytics
Using insurance data for analytics
Using data from mobile/pervasive computing devices
Using lifestyle data for analytics (Exercise/Diet)
Using Clinical Data for Analytics
Social media user
Insurance organisation
Insurance claimant
Mobile/pervasive device user
Lifestyle application user
Patient
Parent/ Carer*
Clinician

Figure 7

3.8.3 CLASS DIAGRAM

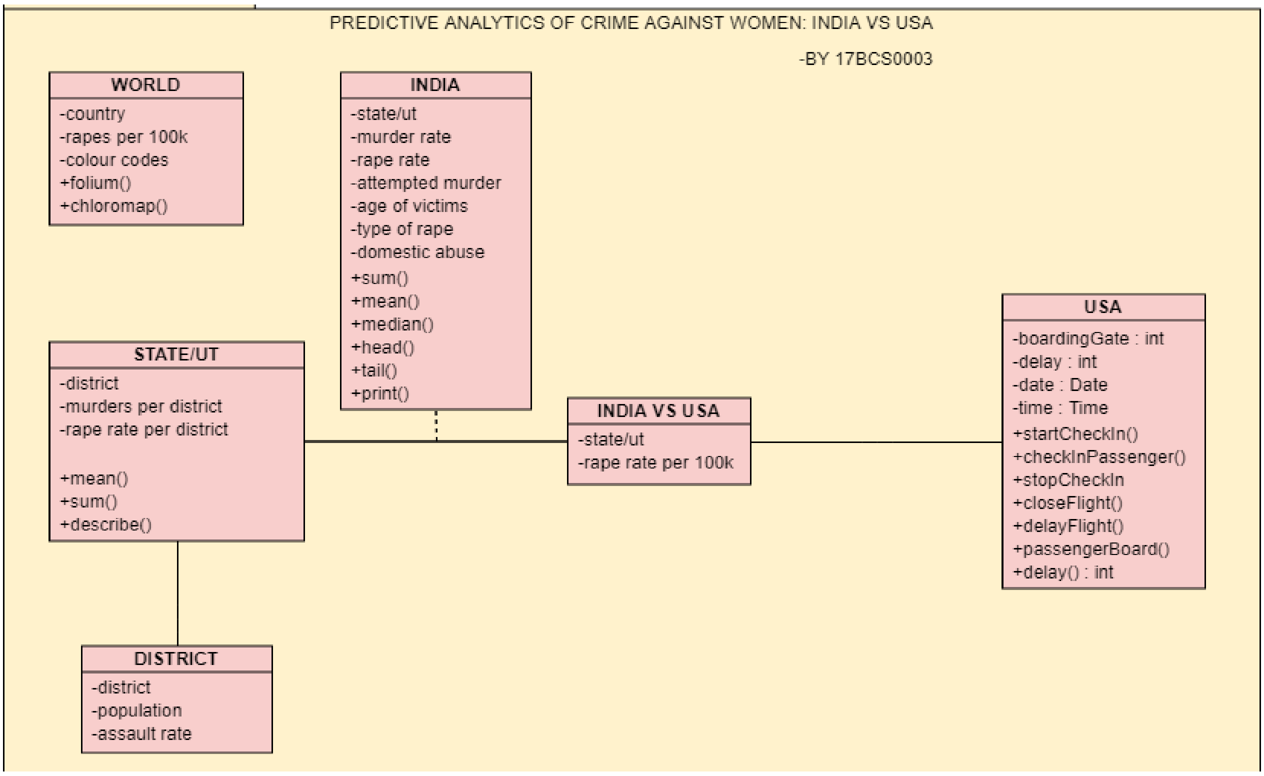


Figure 8

3.8.4 DATA FLOW DIAGRAM

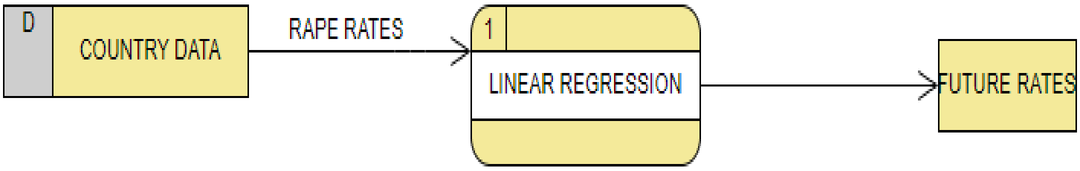


Figure 9

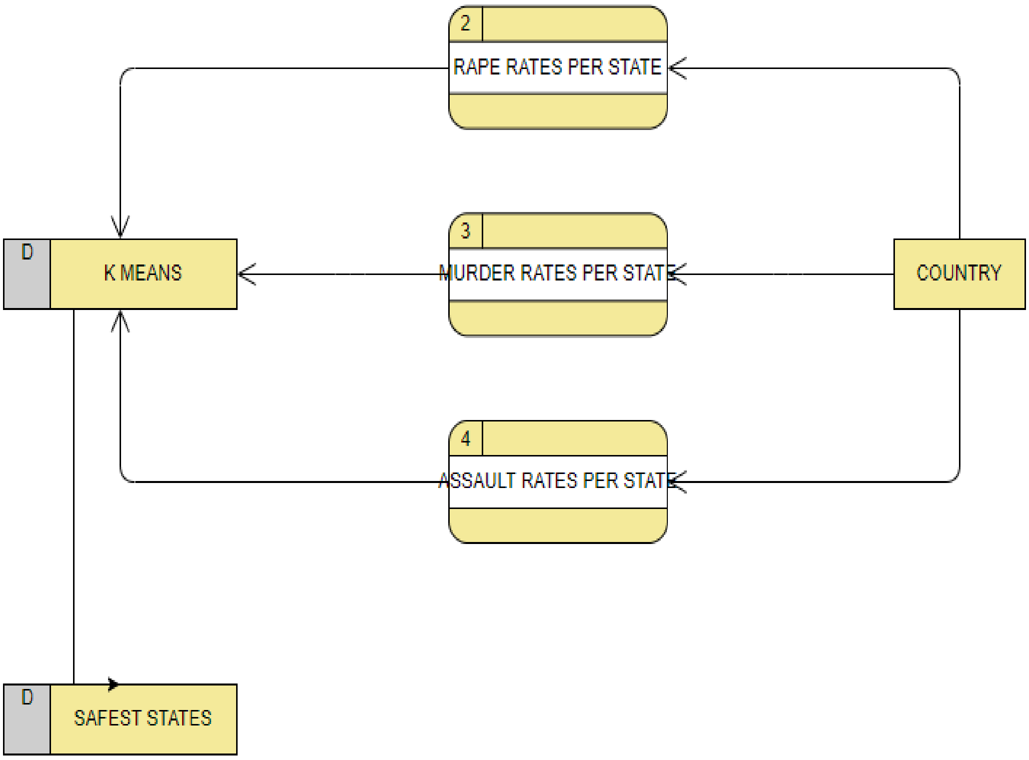


Figure 10

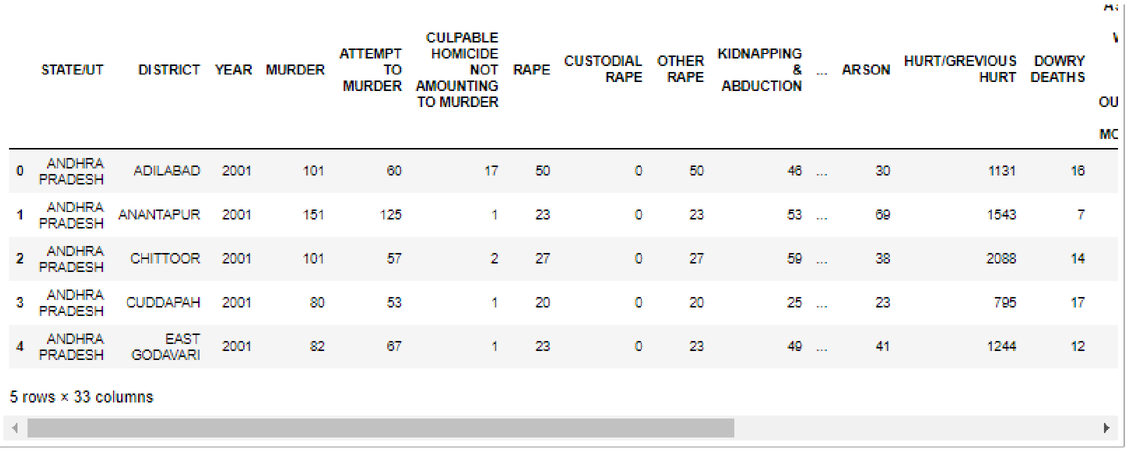
Chapter 4

Implementation

4.1 INDIA

India is considered to be a Third World country and is also a developing country today. India has a high poverty rate, corruption, a very prevalent caste system, and other significant issues that people say are causes for violence and inequality in India. According to the National Crime Records Bureau of India, a crime against women is committed every 3 minutes, and the rate of crime showed an increase of approximately 50% from the years 2011- 2015 (in four years). However, we must understand that India has a population of approximately 136 crores. This has not been taken into account by most studies done. Here, we aim to find the underlying issues behind rape and sexual abuse in India as well as deep dive into the condition of each state. We will also compare India, a third world country with a first world country to understand how development and economy sets us apart. Meanwhile, we hope to bust some of the myths perpetuated about India in the world by faulty statistical studies.

Our first step is to study the dataset provided to better understand the study. The data collected is first imported to the notebook. Using the pandas library the csv file is imported and ready for use. Now, by using df.head() function, by default the first five rows can be seen.

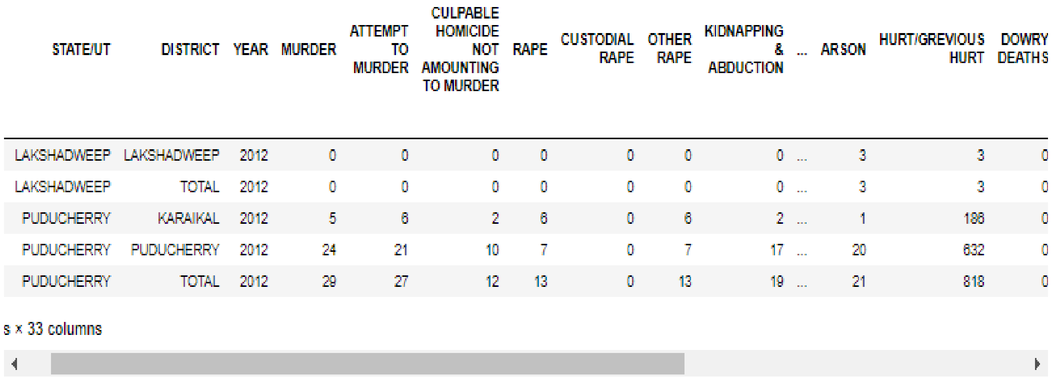


	STATE/UT	DISTRICT	YEAR	MURDER	ATTEMPT TO MURDER	CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	RAPE	CUSTODIAL RAPE	OTHER RAPE	KIDNAPPING & ABDUCTION	...	ARSON	HURT/GREVIOUS HURT	DOWRY DEATHS	A: V OU MC
0	ANDHRA PRADESH	ADILABAD	2001	101	60	17	50	0	50	46	...	30	1131	16	
1	ANDHRA PRADESH	ANANTAPUR	2001	151	125	1	23	0	23	53	...	69	1543	7	
2	ANDHRA PRADESH	CHITTOOR	2001	101	57	2	27	0	27	59	...	38	2088	14	
3	ANDHRA PRADESH	CUDDAPAH	2001	80	53	1	20	0	20	25	...	23	795	17	
4	ANDHRA PRADESH	EAST GODAVARI	2001	82	67	1	23	0	23	49	...	41	1244	12	

5 rows × 33 columns

Table 1

Now, by using the df.tail() function the last five rows are seen. These functions are used to make sure the csv file has been imported correctly.

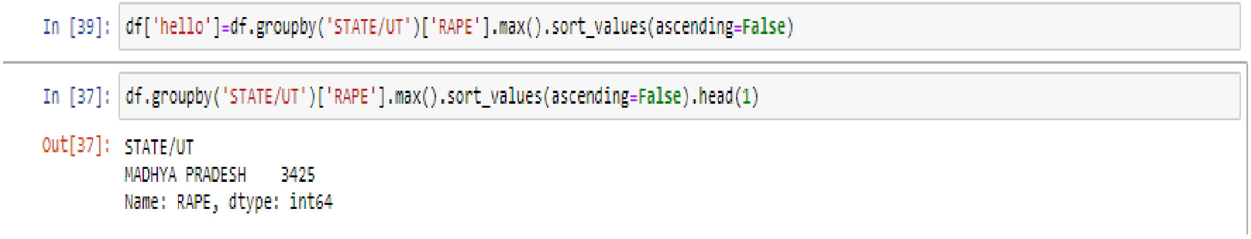


	STATE/UT	DISTRICT	YEAR	MURDER	ATTEMPT TO MURDER	CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	RAPE	CUSTODIAL RAPE	OTHER RAPE	KIDNAPPING & ABDUCTION	...	ARSON	HURT/GREVIOUS HURT	DOWRY DEATHS
	LAKSHADWEEP	LAKSHADWEEP	2012	0	0	0	0	0	0	0	...	3	3	0
	LAKSHADWEEP	TOTAL	2012	0	0	0	0	0	0	0	...	3	3	0
	PUDUCHERRY	KARAIKAL	2012	5	6	2	6	0	6	2	...	1	186	0
	PUDUCHERRY	PUDUCHERRY	2012	24	21	10	7	0	7	17	...	20	632	0
	PUDUCHERRY	TOTAL	2012	29	27	12	13	0	13	19	...	21	818	0

s × 33 columns

Table 2

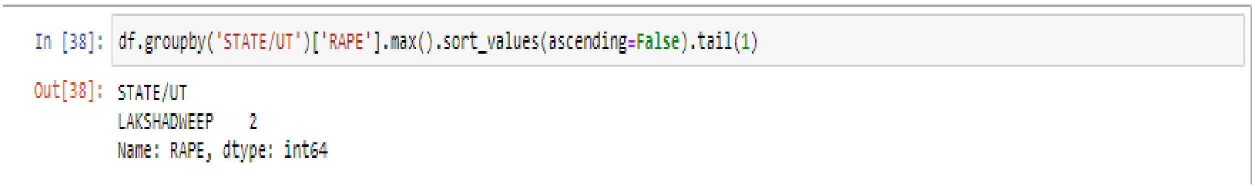
In India, Madhya Pradesh is infamous for the large number of sexual assaults caused. This should reflect in our datasets to make sure the dataset provided is correct. Madhya Pradesh leads over all other states with a colossal amount of 3425 rape assaults. Now, while performing EDA, we find the state which has the most number of reported sexual assaults against women below.



```
In [39]: df['hello']=df.groupby('STATE/UT')['RAPE'].max().sort_values(ascending=False)

In [37]: df.groupby('STATE/UT')['RAPE'].max().sort_values(ascending=False).head(1)

Out[37]: STATE/UT
         MADHYA PRADESH    3425
         Name: RAPE, dtype: int64
```

Table 3

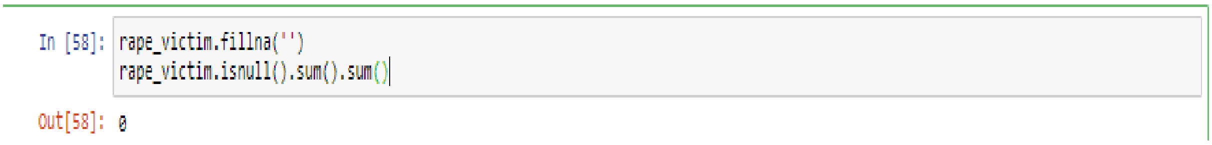
However, the Union Territoty Lakshadweep seems to have the lowest reporting of rape consecutively. Now,we find the state which has the lowest number of sexual assaults.



```
In [38]: df.groupby('STATE/UT')['RAPE'].max().sort_values(ascending=False).tail(1)

Out[38]: STATE/UT
         LAKSHADWEEP    2
         Name: RAPE, dtype: int64
```

Table 4

The null values are replaced with ‘’ and then the sum of null values is calculated to find the sum of null values.



```
In [58]: rape_victim.fillna('')
         rape_victim.isnull().sum().sum()

Out[58]: 0
```

Table 5

Correlogram of the input variables is found to represent graphically, the correlation each variable has on the other. The lighter shades from the graph show higher correlation. It is very interesting to note that RAPE and ASSAULT AGAINST A WOMAN WITH THE INTENT TO OUTRAGE HER MODESTY has one of the highest levels of correlation. We will further prove this through mathematical means while conducting clustering.

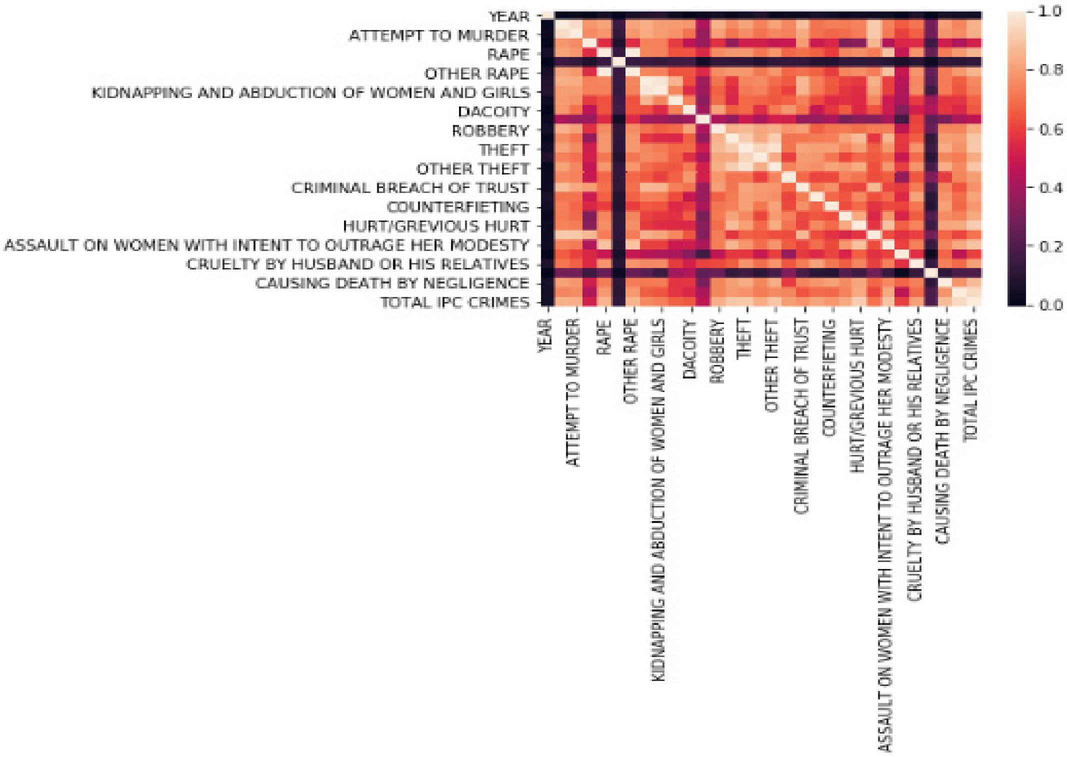


Figure 11

In this box plot we see that over the decade, the intensity of rape cases increases almost steadily. This is very interesting and alarming to note because as the country continues to develop, the conditions of women are alarmingly decreasing.

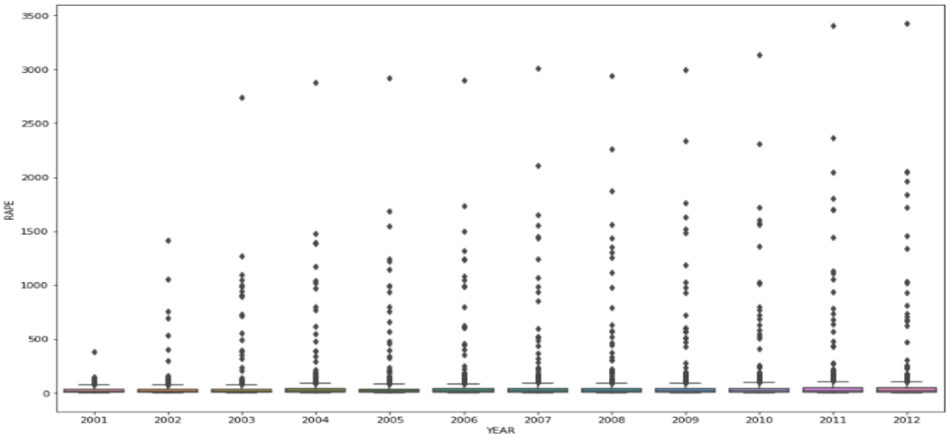


Figure 12

4.2 EXPLORATORY DATA ANALYSIS (EDA)

Over the decade, the condition of women has slowly declined. With a steep and almost steady climb, the number of rapes in India every year is at an all time low .Now, we use python to find and visualize the line graph showing the condition of the country over the past decade. One can see a considerable raise in the number of crimes reported over the decade. From the graph, we can see that over the decade the numbers have gone from approximately 15000 to 22000.

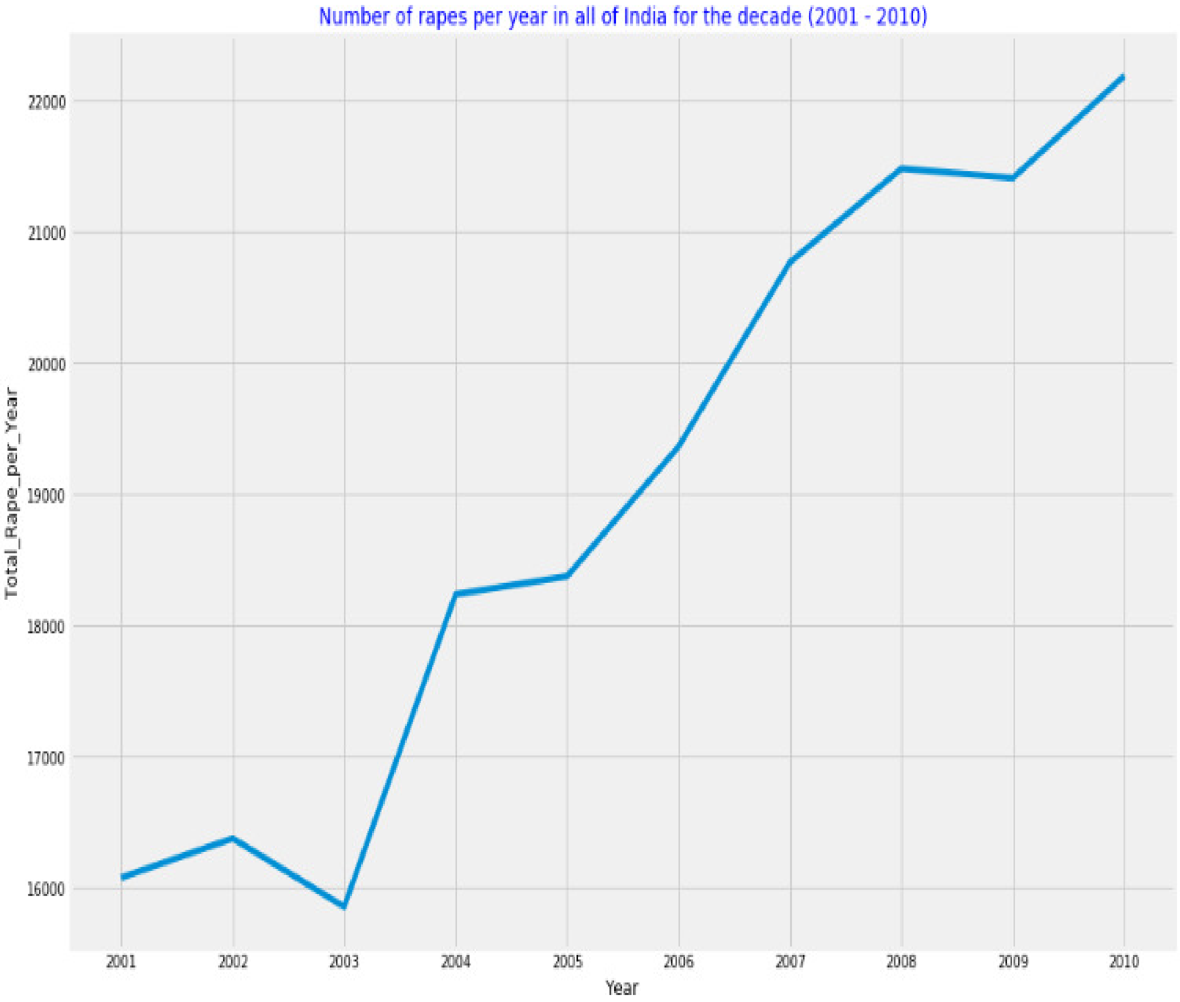


Figure 13

The folium library is used to visualize the crime data against women per capita. This shows that the per capita rates in India are low. This supports the hypothesis that the population spurt in India over the decade could be the reason. The code is written the visualise the .json file for better understanding.

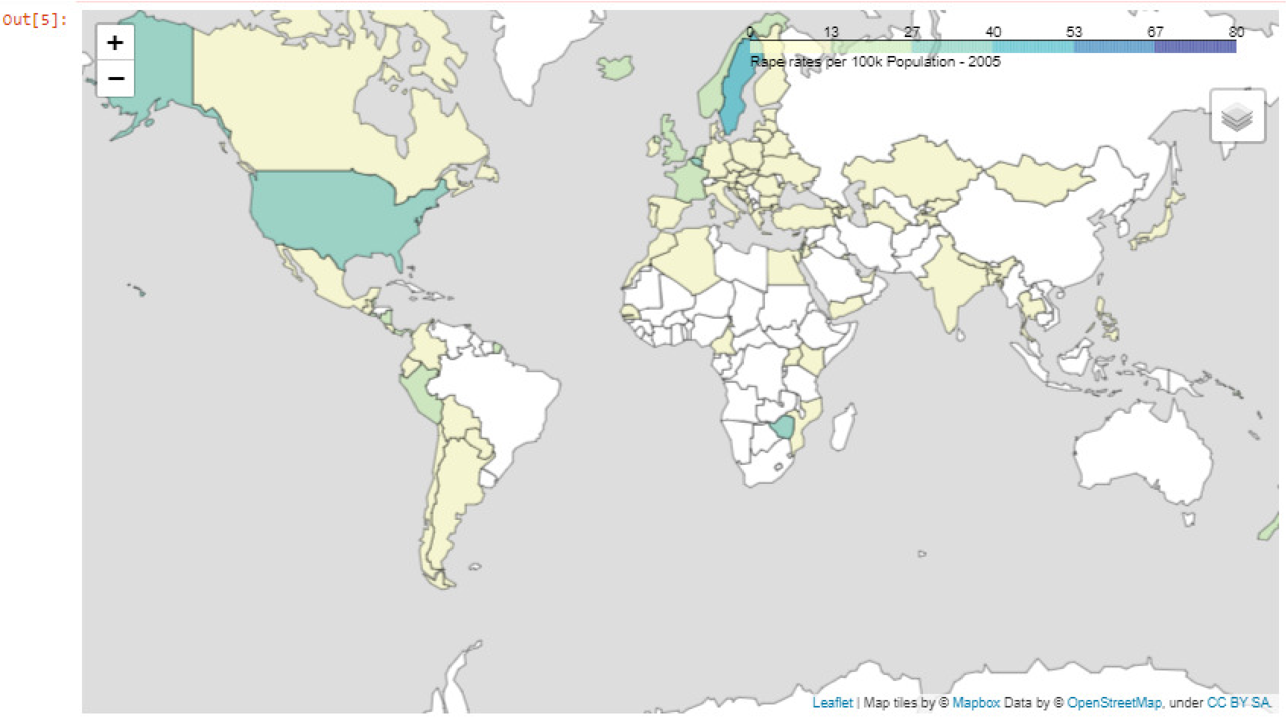

Out[5]:
+
−
0
13
27
40
53
67
80
Rape rates per 100k Population - 2005
Leaflet | Map tiles by © Mapbox Data by © OpenStreetMap, under CC BY SA.

Figure 14

However, even though the number of assaults against women has increased over the decade. Many rumours spread about the state of our country seem to be false. When the number of rapes have been calculated per capita , we can see that India has very low rates of sexual assault in the world

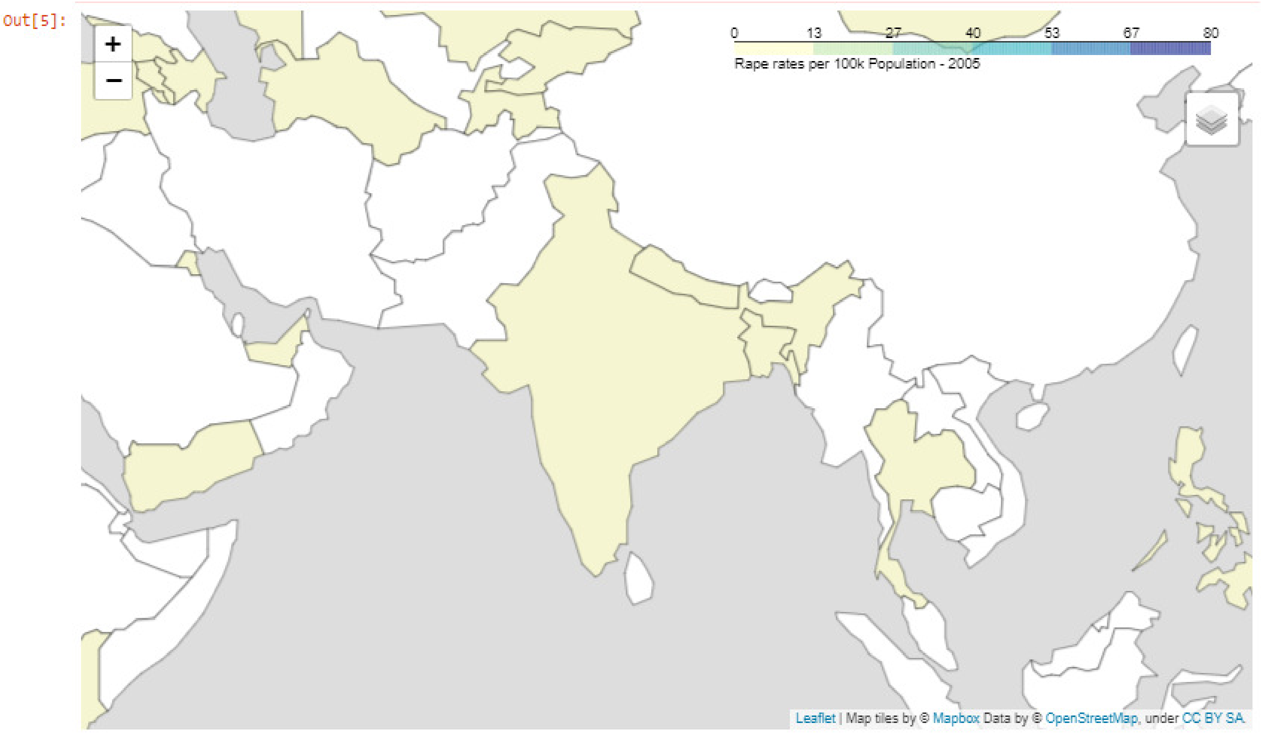
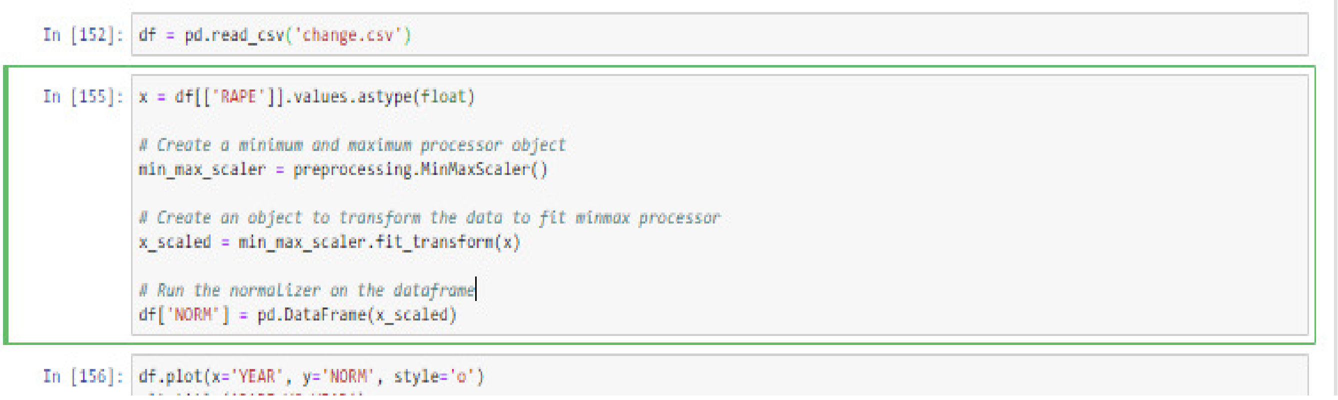

Out[5]:
+
−
0
13
27
40
53
67
80
Rape rates per 100k Population - 2005
Leaflet | Map tiles by © Mapbox Data by © OpenStreetMap, under CC BY SA.

Figure 15

Now the data is cleaned and normalized. From the csv file, the total data is deleted and then the data is normalized. The code for normalization is done using the MinMaxScaler. The df_normalized attributes are seen below.



```
In [152]: df = pd.read_csv('change.csv')

In [155]: x = df[['RAPE']].values.astype(float)

          # Create a minimum and maximum processor object
          min_max_scaler = preprocessing.MinMaxScaler()

          # Create an object to transform the data to fit minmax processor
          x_scaled = min_max_scaler.fit_transform(x)

          # Run the normalizer on the dataframe
          df['NORM'] = pd.DataFrame(x_scaled)

In [156]: df.plot(x='YEAR', y='NORM', style='o')
```

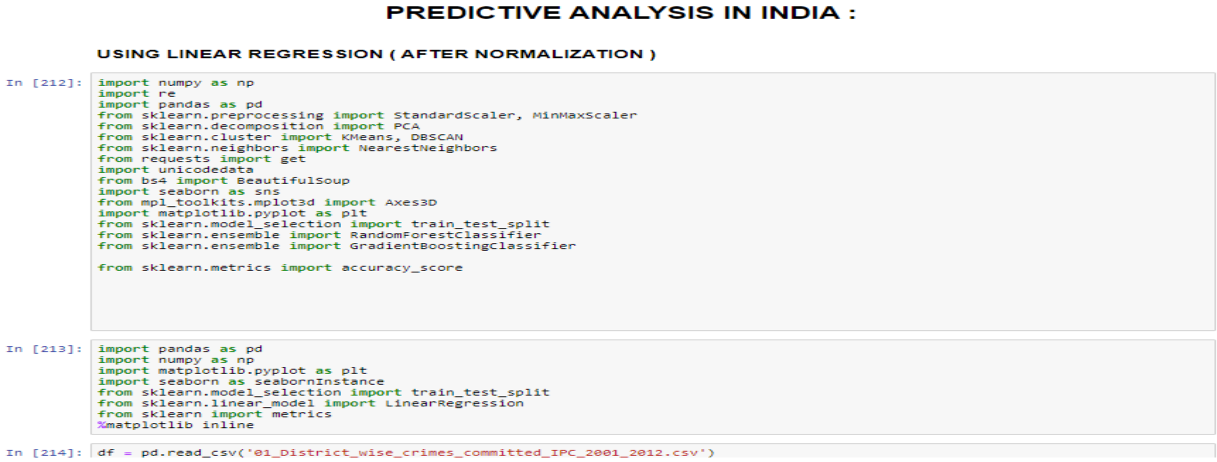
Figure 16

From conducting Exploratory data analysis we have better understood the past data. To summarise, let's look at the various interesting findings we have made.

- Firstly, we find that the state with the highest number of rapes is Madhya Pradesh with an approximate number of 3000
- The one with the lowest number of rapes is Lakshadweep with the approximate value of two per year.
- It is important to note that when you calculate the mean number of rape cases in India, you can see that over the decade it has been approximately ~50
- We then wanted to know the age group for women mostly targeted by rapists and we have found that the women who are at the prime of their age and are usually in colleges or work targeted. The average of age group targeted is from the age of 18-30.
- From the correlogram , you can see that the highest correlation exists between rapes and assault against a woman with the intent to outrage her modesty. This can lead one to believe that rape is conducted with the intent to shame the women or to cause question of her modesty.
- From the world map, you can notice that the number of rapes per capita in India is quite low. This is very contradictory to popular belief that women are not at all safe in India. However, this can also be due to a number of reasons, the main one being that the many number of rapes do not get justice or in many cases, do not even get reported in India due to the various cultural and religious constraints and judgements cast on a woman subject to such abuse.

4.3 LINEAR REGRESSION (INDIA)

If you establish at least a moderate correlation between *X* and *Y* through both a correlation coefficient and a scatter plot, then you know they have some type of linear relationship. In our case, we have a time series. We want to predict the number of rapes that might take place in the future years. Given that we understand the data set, we can arrive at a reasonable number close to the average of rapes over the decades in various states all over India. First lets import all the libraries required for carrying out linear regression effectively.



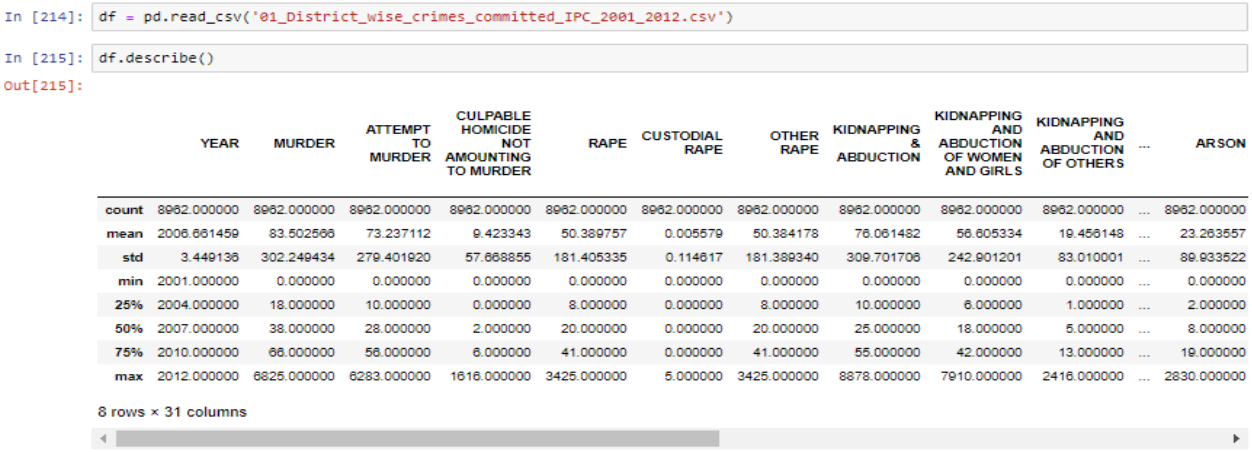
PREDICTIVE ANALYSIS IN INDIA :

USING LINEAR REGRESSION (AFTER NORMALIZATION)

```
In [212]: import numpy as np
          import re
          import pandas as pd
          from sklearn.preprocessing import StandardScaler, MinMaxScaler
          from sklearn.decomposition import PCA
          from sklearn.cluster import KMeans, DBSCAN
          from sklearn.neighbors import NearestNeighbors
          from requests import get
          import unicodedata
          from bs4 import BeautifulSoup
          import seaborn as sns
          from mpl_toolkits.mplot3d import Axes3D
          import matplotlib.pyplot as plt
          from sklearn.model_selection import train_test_split
          from sklearn.ensemble import RandomForestClassifier
          from sklearn.ensemble import GradientBoostingClassifier

          from sklearn.metrics import accuracy_score

In [213]: import pandas as pd
          import numpy as np
          import matplotlib.pyplot as plt
          import seaborn as seabornInstance
          from sklearn.model_selection import train_test_split
          from sklearn.linear_model import LinearRegression
          from sklearn import metrics
          %matplotlib inline

In [214]: df = pd.read_csv('01_District_wise_crimes_committed_IPC_2001_2012.csv')
```

Figure 17

The describe() function helps us understand the dataset we are going to be working on a little better. We understand that the average number of rapes in India is close to 40-50 in districts. So, we must keep this in mind while calculating the result and then testing the values to see if they are accurate.



```
In [214]: df = pd.read_csv('01_District_wise_crimes_committed_IPC_2001_2012.csv')

In [215]: df.describe()

Out[215]:
```

	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	...	ARSON
count	8962.000000	8962.000000	8962.000000	8962.000000	8962.000000	8962.000000	8962.000000	8962.000000	8962.000000	8962.000000	...	8962.000000
mean	2006.661459	83.502566	73.237112	9.423343	50.389757	0.005579	50.384178	76.061482	56.605334	19.456148	...	23.263557
std	3.449136	302.249434	279.401920	57.668855	181.405335	0.114617	181.389340	309.701706	242.901201	83.010001	...	89.933522
min	2001.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000
25%	2004.000000	18.000000	10.000000	0.000000	8.000000	0.000000	8.000000	10.000000	6.000000	1.000000	...	2.000000
50%	2007.000000	38.000000	28.000000	2.000000	20.000000	0.000000	20.000000	25.000000	18.000000	5.000000	...	8.000000
75%	2010.000000	66.000000	56.000000	6.000000	41.000000	0.000000	41.000000	55.000000	42.000000	13.000000	...	19.000000
max	2012.000000	6825.000000	6283.000000	1616.000000	3425.000000	5.000000	3425.000000	8878.000000	7910.000000	2416.000000	...	2830.000000

8 rows × 31 columns

Table 6

The scatter plot must first be executed for the two variables we are going to use for the linear regression . From the linear regression, we see that the intensity of rapes averages around

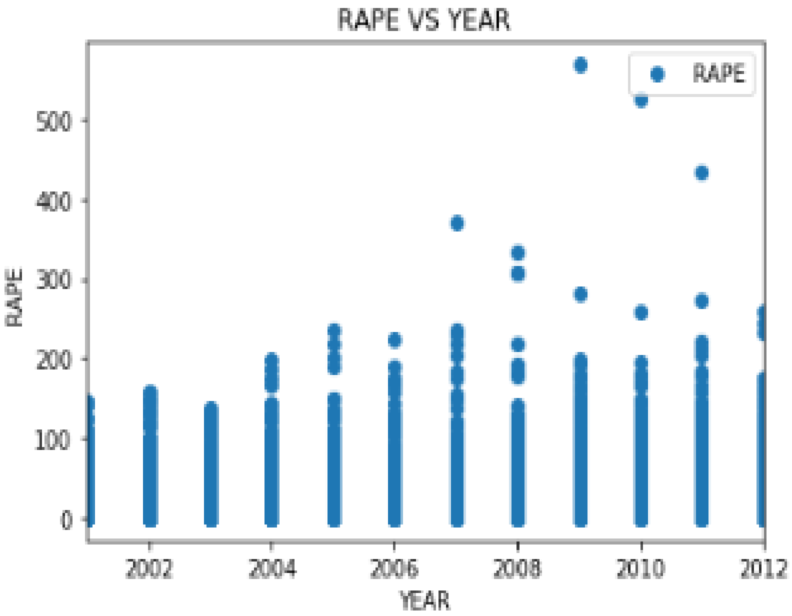


Figure 18

Now let us perform normalization so that we can perform result analysis. After normalization, let us look at the scatter plot.

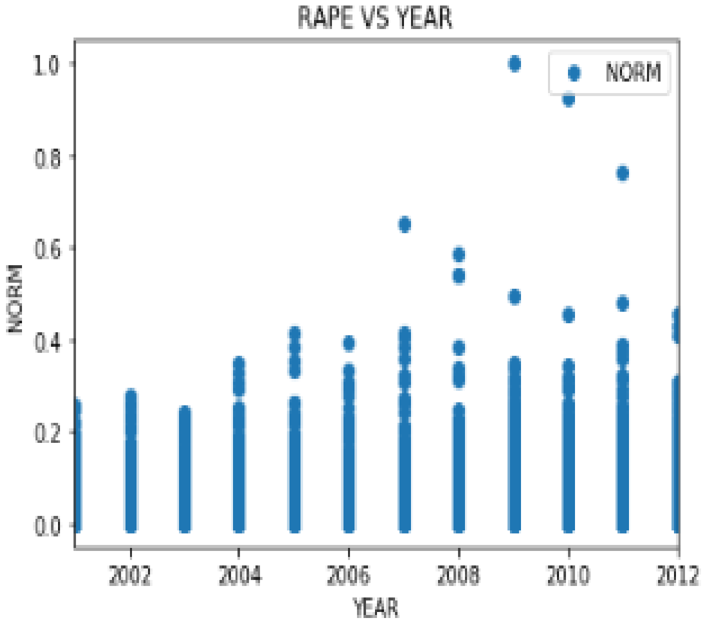
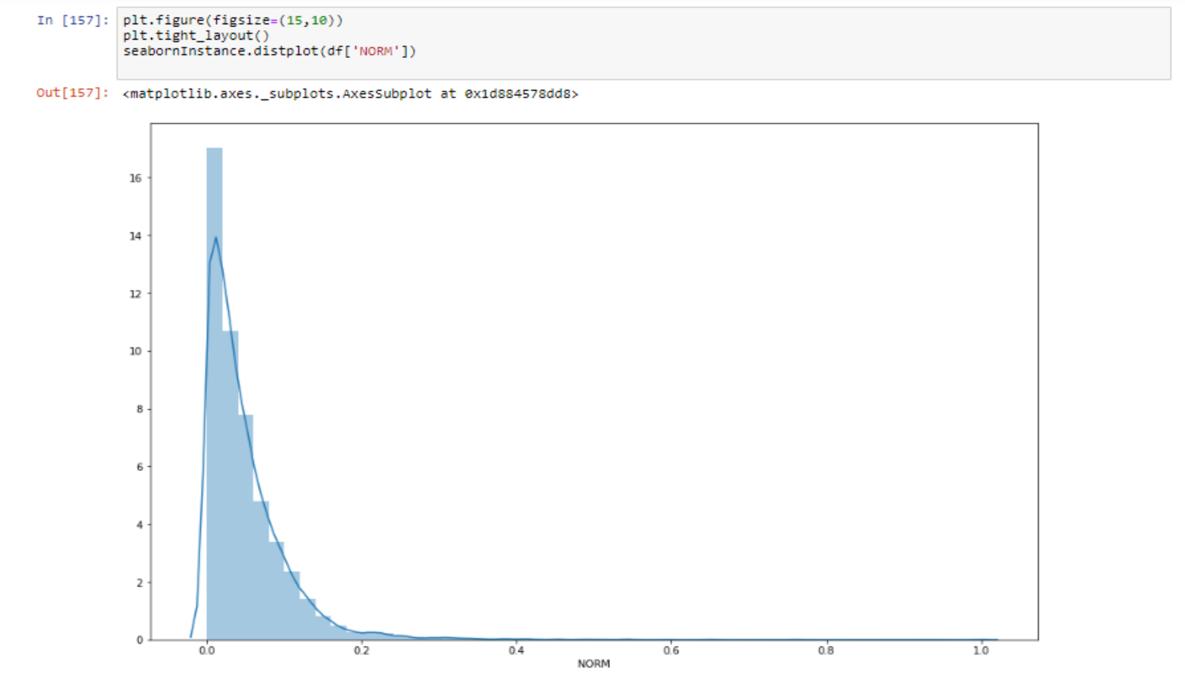


Figure 19

A distribution plot displays a distribution and range of a set of numeric values plotted against a dimension. You can display this chart in three different ways, you can just have the value points displayed showing the distribution, or you can display the bounding box which shows the range or use a combination of both. In the distribution plot shown below, you can see there a range and distribution of the rape values displayed for 0.0- 1.0. Each range and distribution box show how data values for a product group is distributed over the average rape rates per district.



```
In [157]: plt.figure(figsize=(15,10))
          plt.tight_layout()
          seabornInstance.distplot(df['NORM'])

Out[157]: <matplotlib.axes._subplots.AxesSubplot at 0x1d884578dd8>
```

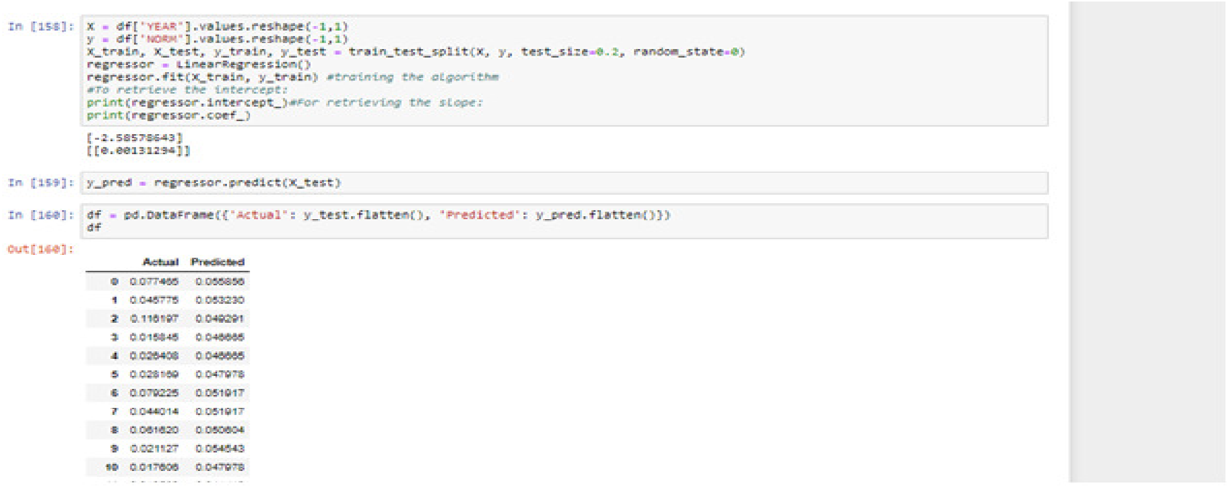
Figure 20

In statistics, the actual value is the value that is obtained by observation or by measuring the available data. It is also called the observed value. The predicted value is the value of the variable predicted based on the regression analysis. The difference between the actual value or observed value and the predicted value is called the residual in regression analysis.

The difference between the actual and the predicted value is the residual which is defined as:

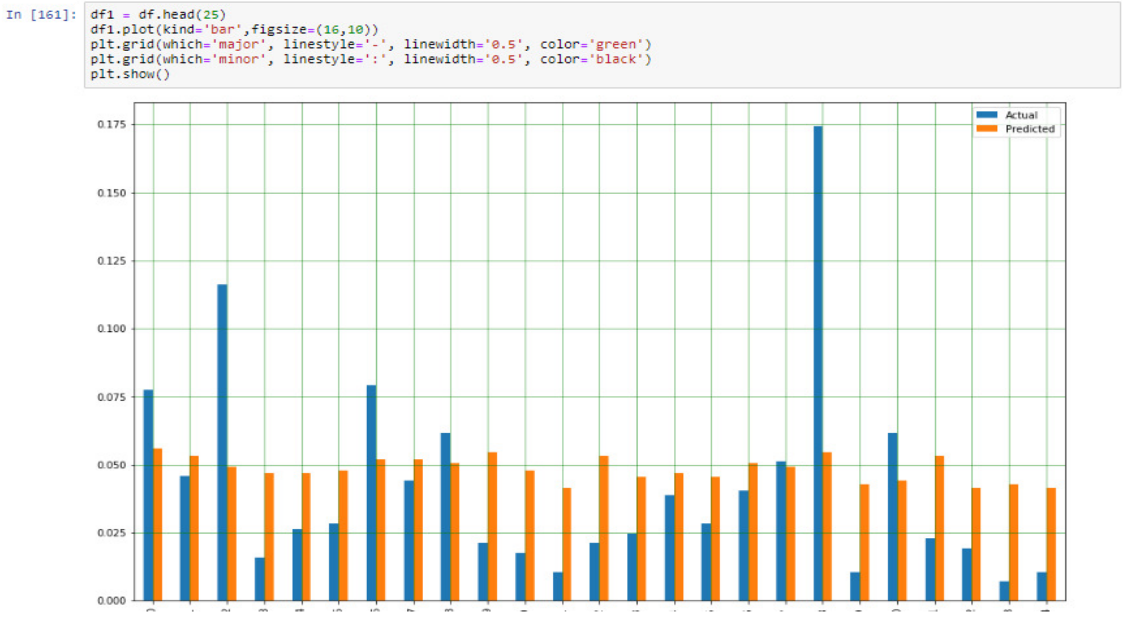
$e = y - \hat{y}$

Here, e is the residual, y is the observed or actual value and $\hat{y}$ is the predicted value. Each actual value has a predicted value and hence each data point has one residual. If the difference between the actual value and the predicted value is positive, then the data points are above the regression line. If the difference between the actual value and the predicted value is negative, then the data points are below the regression line. If the difference is zero, then that data points lie on the regression line. If the line of best fit is the best fit then the sum of the difference between the actual value and the predicted values is always zero. The residuals play a vital role to validate the obtained regression model. Residuals are represented graphically by means of a residual plot. If the data points on the residual plot are spread around the horizontal axis, it indicates the appropriateness of a linear regression model. If it is not spread around the horizontal axis, it indicates the appropriateness of non-linear regression model.



```
In [158]: X = df['YEAR'].values.reshape(-1,1)
          y = df['NORM'].values.reshape(-1,1)
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
          regressor = LinearRegression()
          regressor.fit(X_train, y_train) #training the algorithm
          #To retrieve the intercept:
          print(regressor.intercept_)#For retrieving the slope:
          print(regressor.coef_)

          [-2.58578643]
          [[0.00131294]]

In [159]: y_pred = regressor.predict(X_test)

In [160]: df = pd.DataFrame({'Actual': y_test.flatten(), 'Predicted': y_pred.flatten()})
          df

Out[160]:
             Actual  Predicted
          0  0.077465  0.055856
          1  0.045775  0.053230
          2  0.116197  0.049291
          3  0.015845  0.046665
          4  0.026408  0.046665
          5  0.028169  0.047978
          6  0.079225  0.051917
          7  0.044014  0.051917
          8  0.061620  0.050604
          9  0.021127  0.054543
         10  0.017606  0.047978
```

Figure 21

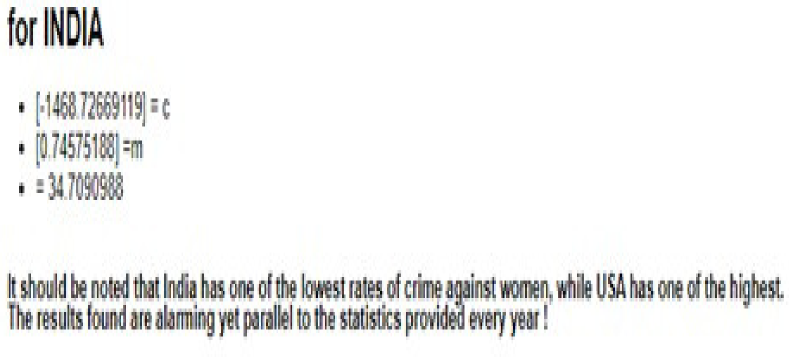


```
In [161]: df1 = df.head(25)
          df1.plot(kind='bar',figsize=(16,10))
          plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
          plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
          plt.show()
```

Figure 22

From the above calculations we can see that the y intercept has a value of -2.585 and the slope has a value of 0.0001. When we apply this to the equation, y=mx+c, to find the value of the dependable variable y. This will help us predict the rape rates per state for the future years. This can help us understand and better prepare for the future law enforcement.

Let us calculate the predicted value of the average value of rapes per state in India. Once, this is done the value of the predicted variable is ~34.7 per state for a single year.



for INDIA

- [-1468.72669119] = c
- [0.74575188] =m
- = 34.7090988

It should be noted that India has one of the lowest rates of crime against women, while USA has one of the highest.
The results found are alarming yet parallel to the statistics provided every year !

This proves the point that the per capita rates of rape in India is one of the lowest rates in the world.

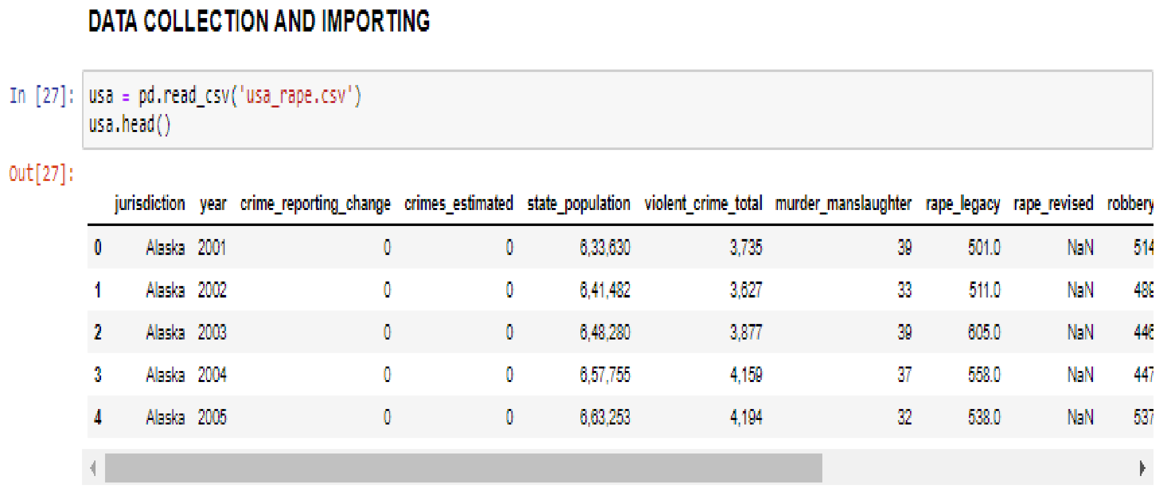
4.4 UNDERSTANDING USA

Modern connotation of the word ‘first world’ ,which USA is both politically and in terms of privilege does refer to the privileged people, who live in a stable economy by virtue of a stable political system, where people have food, shelter, clothing, transportation, healthcare and an infrastructure in place, and yet they are complaining and whining about little small things not aligning their ways and are oblivious to the real issues that people in the world face, of walking long distances for water, of having children that have stunted growth because they can’t feed them baby food and they have osteoporosis and anorexia because of lack of nutrition, gathering in a large open space to watch a movie or never be entertained, or eat meat only on a festive day once a year, something that is hard to imagine for some of the western people.

However, on average, there are 433,648 victims (age 12 or older) of rape and sexual assault each year in the United States. Ages 12-34 are the highest risk years for rape and sexual assault. While those age 65 and older are 92% less likely than 12-24 year olds to be a victim of rape or sexual assault, and 83% less likely than 25-49 year olds. Millions of women in the United States have experienced rape. As of 1998, an estimated 17.7 million American women had been victims of attempted or completed rape. Young women are especially at risk. 82% of all juvenile victims are female. 90% of adult rape victims are female. Females ages 16-19 are 4 times more likely than the general population to be victims of rape, attempted rape, or sexual assault. Women ages 18-24 who are college students are 3 times more likely than women in general to experience sexual violence. Females of the same age who are not enrolled in college are 4 times more likely.

4.5 EXPLORATORY DATA ANALYSIS (USA)

Our first step is to study the dataset provided to better understand the study. The data collected is first imported to the notebook. Using the pandas library the csv file is imported and ready for use. Now, by using df.head() function, by default the first five rows can be seen.



DATA COLLECTION AND IMPORTING

In [27]:
```
usa = pd.read_csv('usa_rape.csv')
usa.head()
```

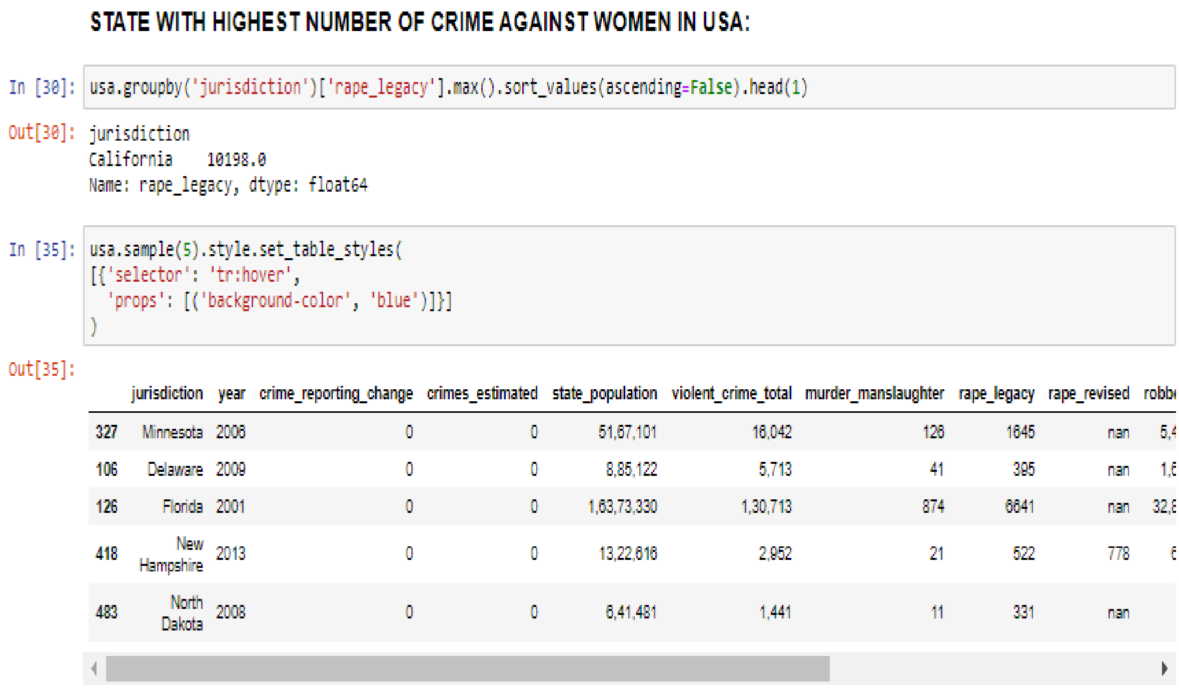
Out[27]:

	jurisdiction	year	crime_reporting_change	crimes_estimated	state_population	violent_crime_total	murder_manslaughter	rape_legacy	rape_revised	robbery
0	Alaska	2001	0	0	6,33,630	3,735	39	501.0	NaN	514
1	Alaska	2002	0	0	6,41,482	3,627	33	511.0	NaN	48
2	Alaska	2003	0	0	6,48,280	3,877	39	605.0	NaN	44
3	Alaska	2004	0	0	6,57,755	4,159	37	558.0	NaN	447
4	Alaska	2005	0	0	6,63,253	4,194	32	538.0	NaN	537

Table 7

In 2018, California had the highest number of forcible rape cases in the United States, with 15,505 reported rapes. Vermont had the lowest number of reported forcible rape cases at 243. It is perhaps unsurprising that California had the highest number of reported rapes in the United States in 2018, as California is the state with the highest population. When looking at the rape rate, or the number of rapes per 100,000 of the population, a very different picture is painted: Alaska was the state with the highest rape rate in the country in 2017, with California ranking as 35th in the nation.

.

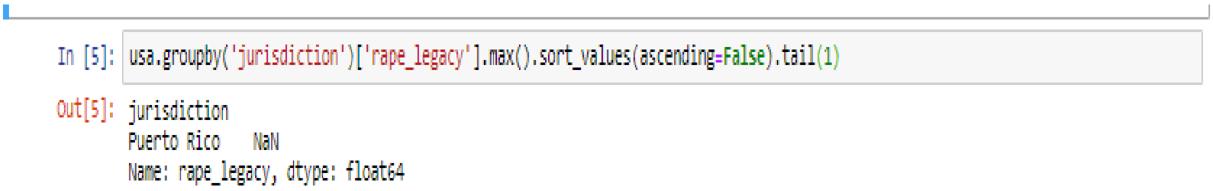


STATE WITH HIGHEST NUMBER OF CRIME AGAINST WOMEN IN USA:

```
In [30]: usa.groupby('jurisdiction')['rape_legacy'].max().sort_values(ascending=False).head(1)

Out[30]: jurisdiction
         California    10198.0
         Name: rape_legacy, dtype: float64
```

```
In [35]: usa.sample(5).style.set_table_styles(
         [{'selector': 'tr:hover',
           'props': [('background-color', 'blue')]}]
         )

Out[35]:
```

	jurisdiction	year	crime_reporting_change	crimes_estimated	state_population	violent_crime_total	murder_manslaughter	rape_legacy	rape_revised	robbe
327	Minnesota	2006	0	0	51,67,101	16,042	126	1645	nan	5,4
106	Delaware	2009	0	0	8,85,122	5,713	41	395	nan	1,6
126	Florida	2001	0	0	1,63,73,330	1,30,713	874	6641	nan	32,8
418	New Hampshire	2013	0	0	13,22,616	2,952	21	522	778	6
483	North Dakota	2008	0	0	6,41,481	1,441	11	331	nan	



```
In [5]: usa.groupby('jurisdiction')['rape_legacy'].max().sort_values(ascending=False).tail(1)

Out[5]: jurisdiction
        Puerto Rico    NaN
        Name: rape_legacy, dtype: float64
```

Table 8

Rape prevalence among women in the U.S. (the percentage of women who experienced rape at least once in their lifetime so far) is in the range of 15–20% according to different studies (National Violence against Women survey, 1995, found 17.6% prevalence rate.A 2007 national study for the Department of Justice on rape found 18% prevalence rate. According to a March 2013 report from the U.S. Department of Justice's Bureau of Justice Statistics, from 1995 to 2010, the estimated annual rate of female rape or sexual assault declined 58%, from 5.0 victimizations per 1,000 females age 12 or older to 2.1 per 1,000.

The 2018 Uniform Crime Report (UCR), which measures rapes that are known to police, estimated that there were 127,258 rapes reported to law enforcement in 2018. However, in 2016 National Crime Victimization Survey (NCVS), which measures sexual assaults and rapes that may not have been reported to the police, estimated that there were 431,840 incidents of rape or sexual assault in 2015.

This is the graph showing you the number of crimes against women in USA over the decade:

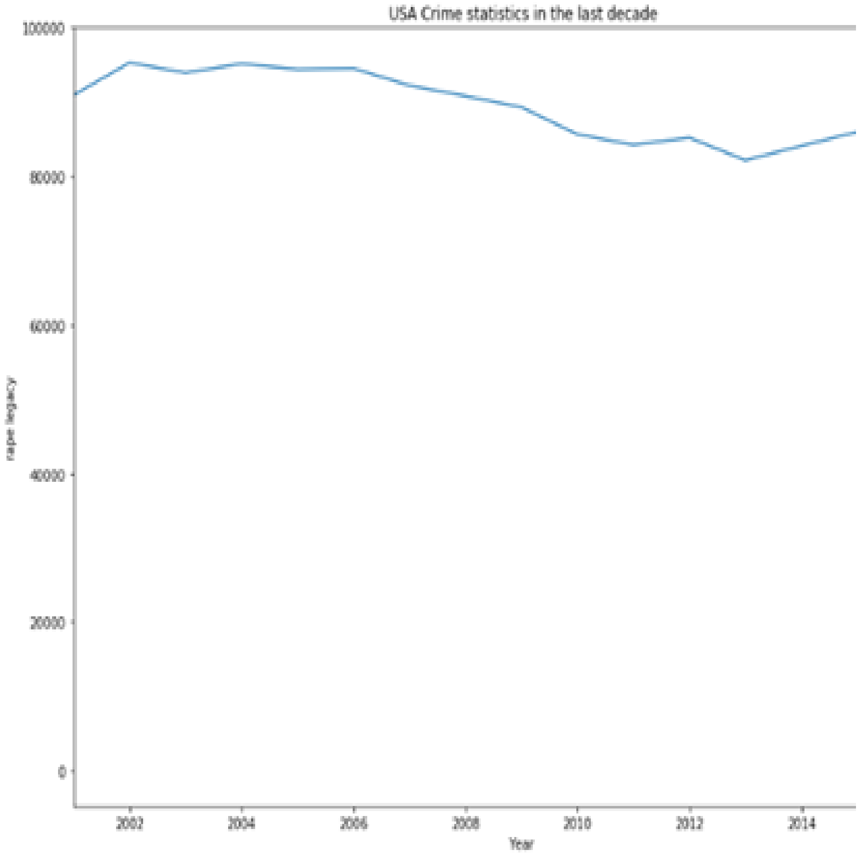


Figure 23

In 2018, the rate of forcible rapes in the United States stood at 30.9 per 100,000 inhabitants. While this figure is about the same as it was in 2007, when the rate was 30.6, it has decreased from 1990, when there were 41.2 forcible rapes per 100,000 inhabitants.

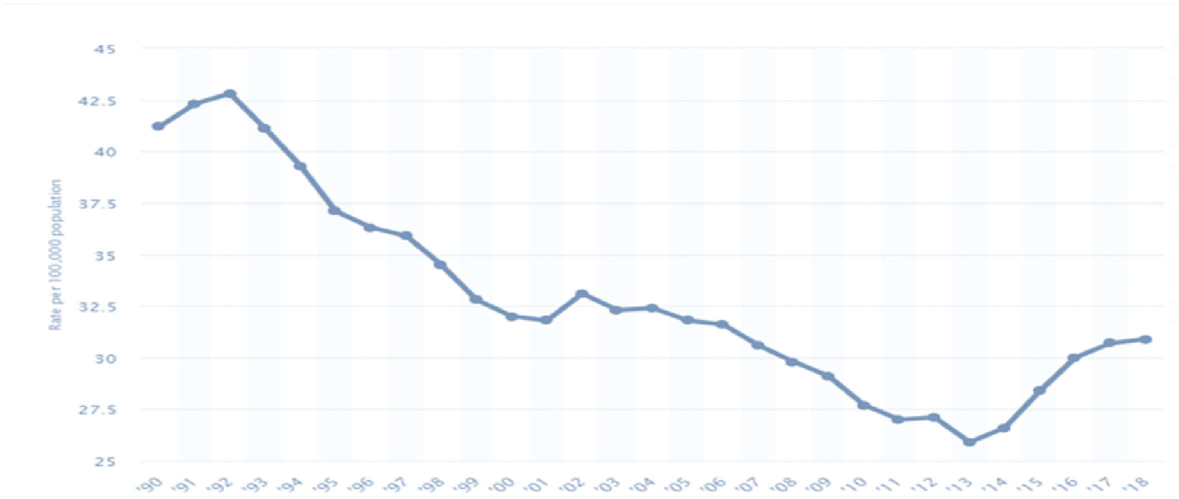
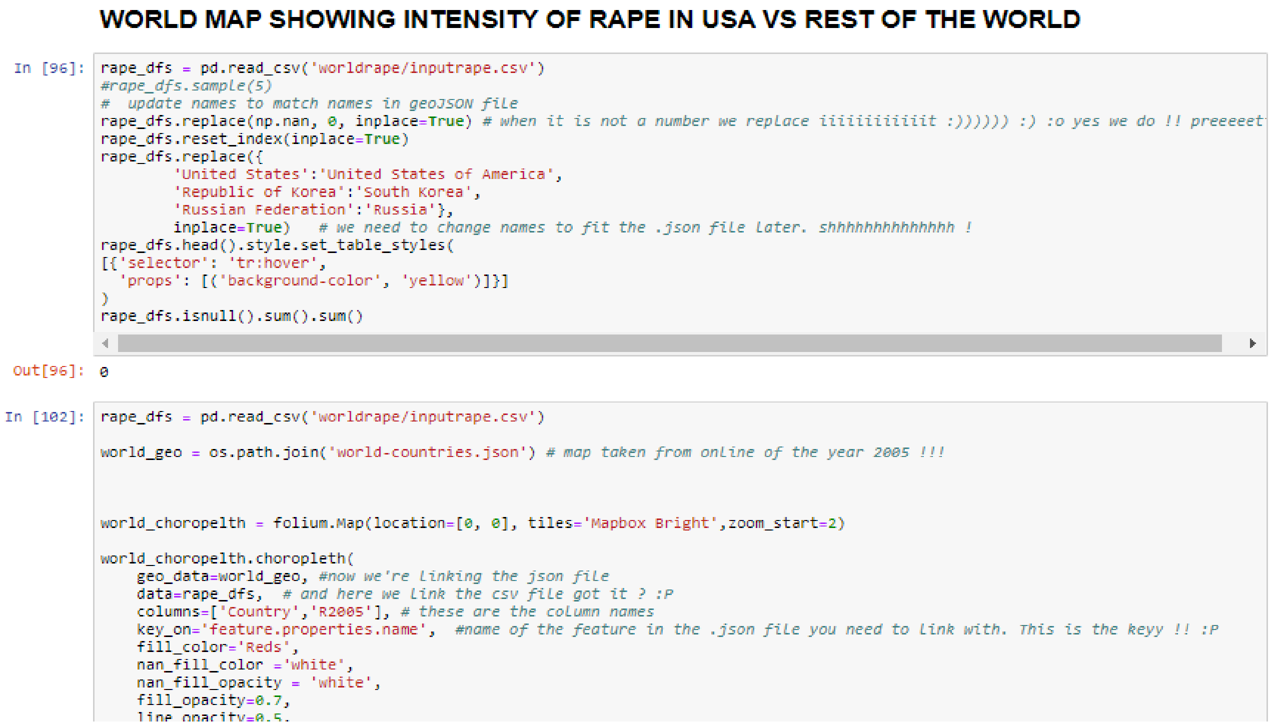


Figure 24



WORLD MAP SHOWING INTENSITY OF RAPE IN USA VS REST OF THE WORLD

```
In [96]: rape_dfs = pd.read_csv('worldrape/inputrape.csv')
         #rape_dfs.sample(5)
         #  update names to match names in geoJSON file
         rape_dfs.replace(np.nan, 0, inplace=True) # when it is not a number we replace iiiiiiiiiiit :)))))) :) :o yes we do !! preeeeet
         rape_dfs.reset_index(inplace=True)
         rape_dfs.replace({
                 'United States':'United States of America',
                 'Republic of Korea':'South Korea',
                 'Russian Federation':'Russia'},
                 inplace=True)   # we need to change names to fit the .json file later. shhhhhhhhhhhhhh !
         rape_dfs.head().style.set_table_styles(
         [{'selector': 'tr:hover',
           'props': [('background-color', 'yellow')]}]
         )
         rape_dfs.isnull().sum().sum()

Out[96]: 0
```

```
In [102]: rape_dfs = pd.read_csv('worldrape/inputrape.csv')

          world_geo = os.path.join('world-countries.json') # map taken from online of the year 2005 !!!


          world_choropelth = folium.Map(location=[0, 0], tiles='Mapbox Bright',zoom_start=2)

          world_choropelth.choropleth(
              geo_data=world_geo, #now we're linking the json file
              data=rape_dfs,  # and here we link the csv file got it ? :P
              columns=['Country','R2005'], # these are the column names
              key_on='feature.properties.name',  #name of the feature in the .json file you need to link with. This is the keyy !! :P
              fill_color='Reds',
              nan_fill_color ='white',
              nan_fill_opacity = 'white',
              fill_opacity=0.7,
              line_opacity=0.5,
```

Figure 25

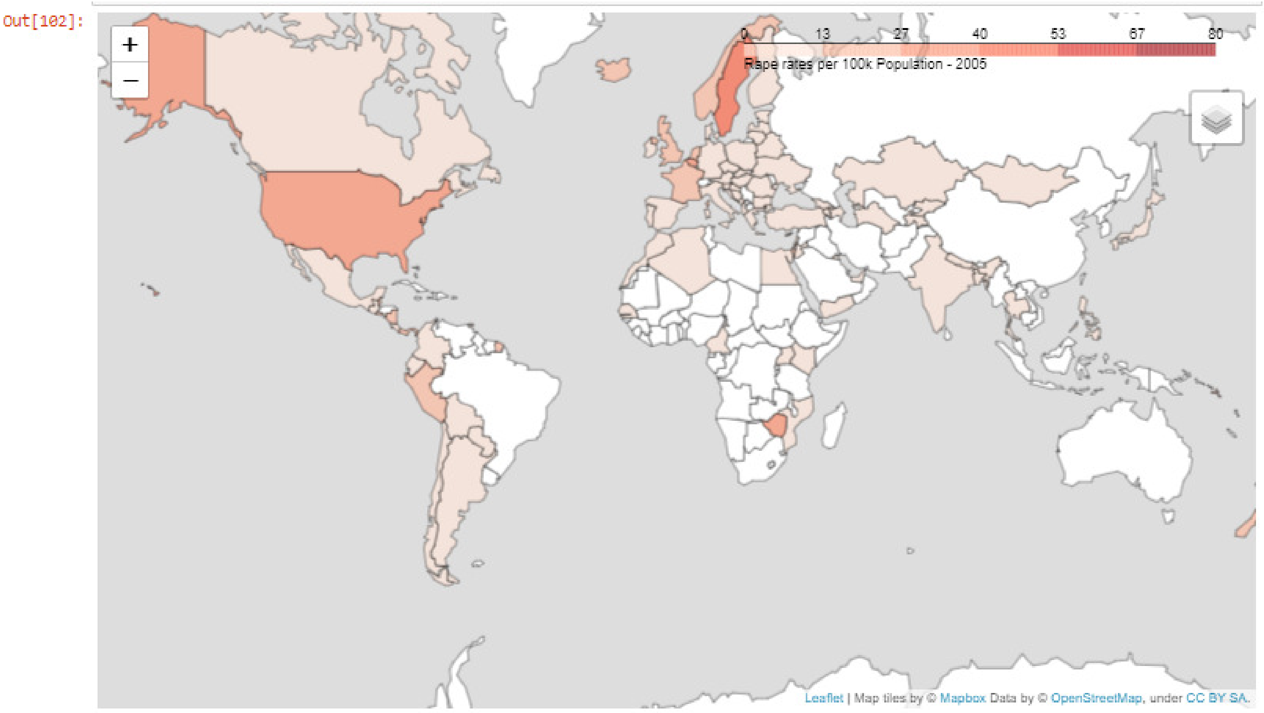
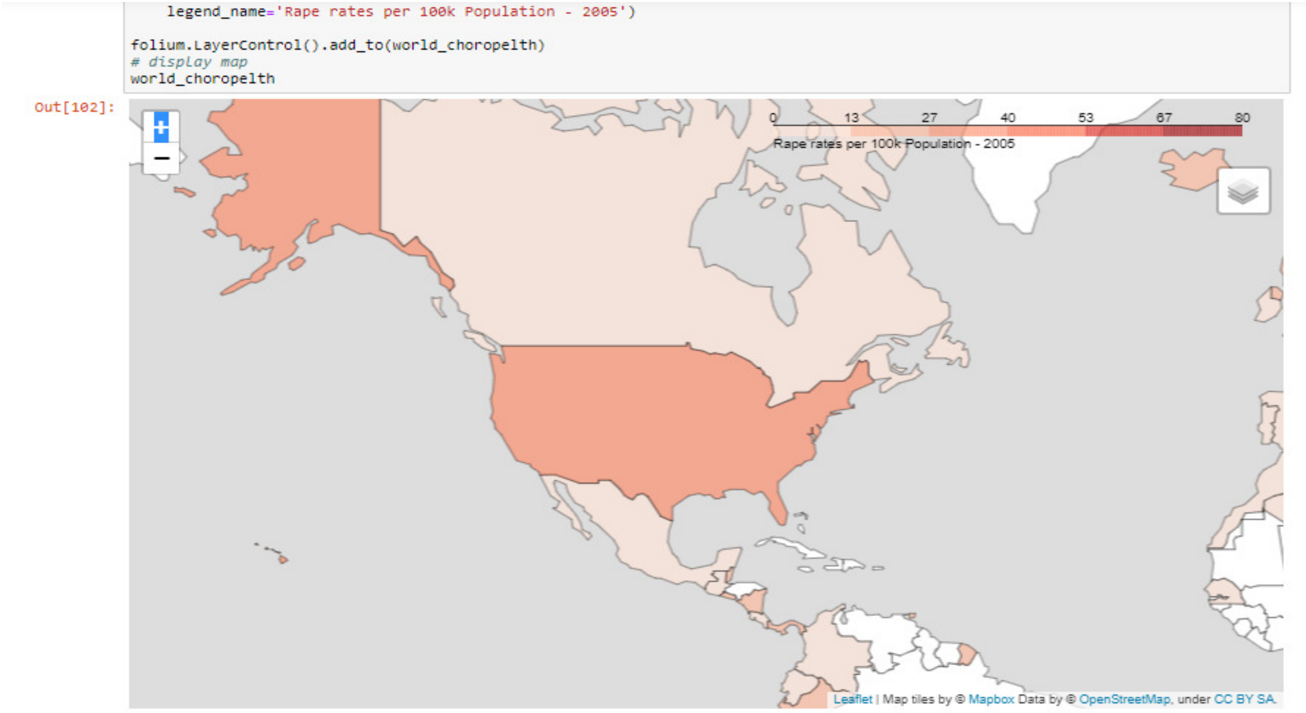

Out[102]:
+
–
0
13
27
40
53
67
80
Rape rates per 100k Population - 2005
Leaflet | Map tiles by © Mapbox Data by © OpenStreetMap, under CC BY SA.

Figure 26

We can see that USA has one of the highest numbers of rapes in the world:



```
    legend_name='Rape rates per 100k Population - 2005')

folium.LayerControl().add_to(world_choropelth)
# display map
world_choropelth
```

Out[102]:
+
–
0
13
27
40
53
67
80
Rape rates per 100k Population - 2005
Leaflet | Map tiles by © Mapbox Data by © OpenStreetMap, under CC BY SA.

Figure 27

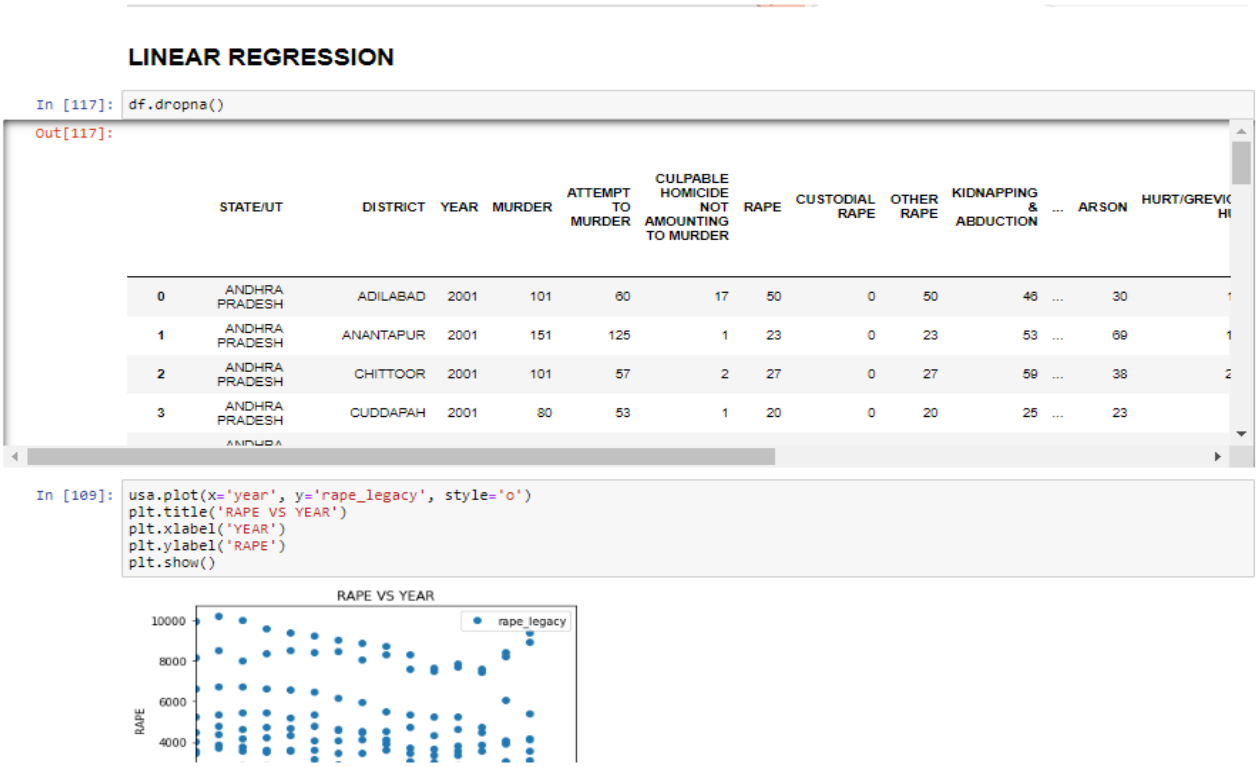
According to a March 2013 report from the U.S. Department of Justice's Bureau of Justice Statistics, from 1995 to 2010, the estimated annual rate of female rape or sexual assault declined 58%, from 5.0 victimizations per 1,000 females age 12 or older to 2.1 per 1,000.

The 2018 Uniform Crime Report (UCR), which measures rapes that are known to police, estimated that there were 127,258 rapes reported to law enforcement in 2018. However, in 2016 National Crime Victimization Survey (NCVS), which measures sexual assaults and rapes that may not have been reported to the police, estimated that there were 431,840 incidents of rape or sexual assault in 2015.

4.6 LINEAR REGRESSION (USA)

Linear regression is one of the most fundamental regression technique used in machine learning. It is based on the premise that one can draw a linear surface such as a line, plane or a hyper-plane that best fits the datapoints in the dataset. Once such linear surface is found using using Linear regression model, then we can easily predict future values using the same linear equation.

The accuracy of model depends on whether the dataset is suited for linear model or is it more suited for non-linear surface such as curve etc.



LINEAR REGRESSION

```
In [117]: df.dropna()
Out[117]:
```

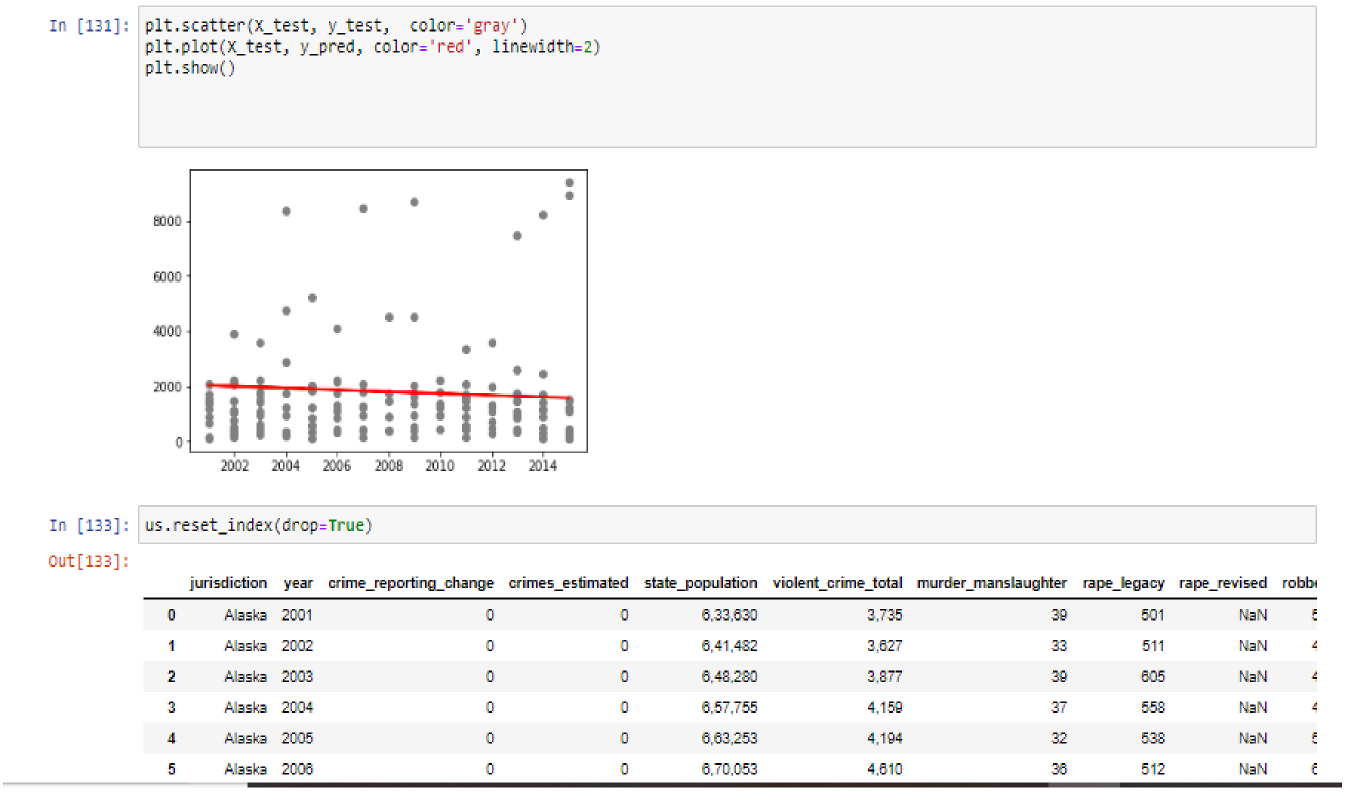
	STATE/UT	DISTRICT	YEAR	MURDER	ATTEMPT TO MURDER	CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	RAPE	CUSTODIAL RAPE	OTHER RAPE	KIDNAPPING & ABDUCTION	...	ARSON	HURT/GREVIC HI
0	ANDHRA PRADESH	ADILABAD	2001	101	60	17	50	0	50	46	...	30	1
1	ANDHRA PRADESH	ANANTAPUR	2001	151	125	1	23	0	23	53	...	69	1
2	ANDHRA PRADESH	CHITTOOR	2001	101	57	2	27	0	27	59	...	38	2
3	ANDHRA PRADESH	CUDDAPAH	2001	80	53	1	20	0	20	25	...	23	
	ANDHRA												

```
In [109]: usa.plot(x='year', y='rape_legacy', style='o')
          plt.title('RAPE VS YEAR')
          plt.xlabel('YEAR')
          plt.ylabel('RAPE')
          plt.show()
```

Table 9

The regression line we get clearly shows that there has been a decrease in the number of rapes over the decade. While in India, we have seen a steady raise. This poses many questions. One such question is what the values of rape will be in the future years.

You can see that the regression line is now plotted based on the predicted values:



```
In [131]: plt.scatter(X_test, y_test,  color='gray')
          plt.plot(X_test, y_pred, color='red', linewidth=2)
          plt.show()
```

```
In [133]: us.reset_index(drop=True)
```

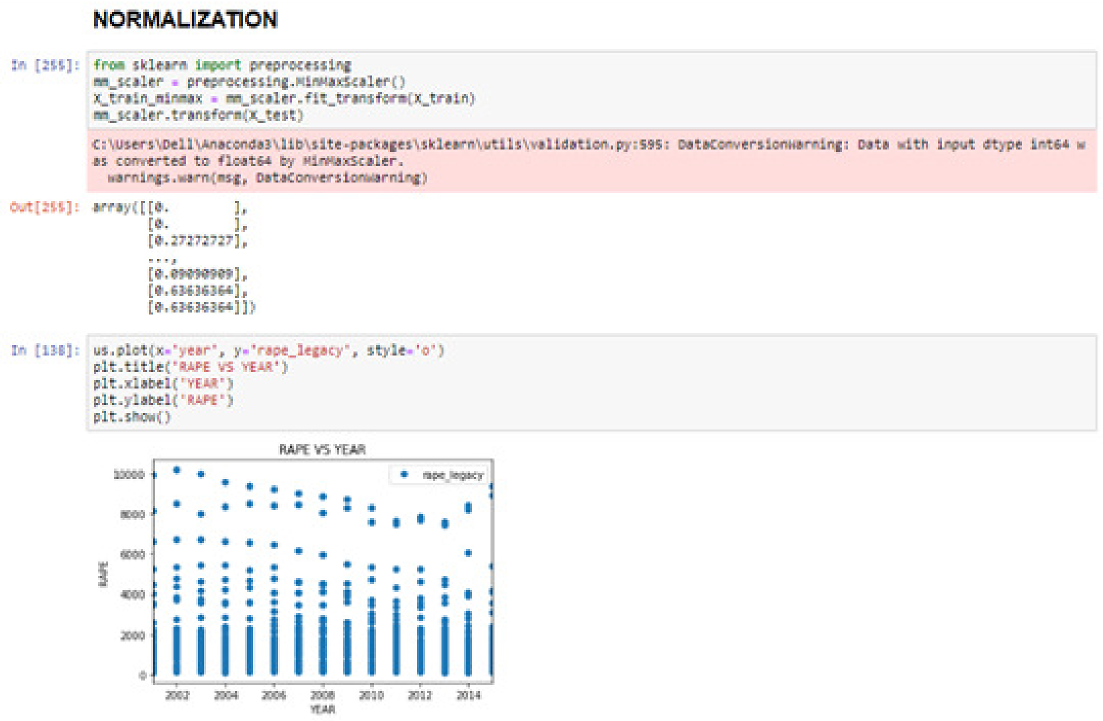
Out[133]:

	jurisdiction	year	crime_reporting_change	crimes_estimated	state_population	violent_crime_total	murder_manslaughter	rape_legacy	rape_revised	robb
0	Alaska	2001	0	0	6,33,630	3,735	39	501	NaN	5
1	Alaska	2002	0	0	6,41,482	3,627	33	511	NaN	4
2	Alaska	2003	0	0	6,48,280	3,877	39	605	NaN	4
3	Alaska	2004	0	0	6,57,755	4,159	37	558	NaN	4
4	Alaska	2005	0	0	6,63,253	4,194	32	538	NaN	5
5	Alaska	2006	0	0	6,70,053	4,610	36	512	NaN	6

Figure 28

The word "normalization" is used informally in statistics, and so the term *normalized data* can have multiple meanings. In most cases, when you normalize data you eliminate the units of measurement for data, enabling you to more easily compare data from different places. **<u>Rescaling data</u> to have values between 0 and 1.** This is usually called *feature scaling*. One possible formula to achieve this is:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$



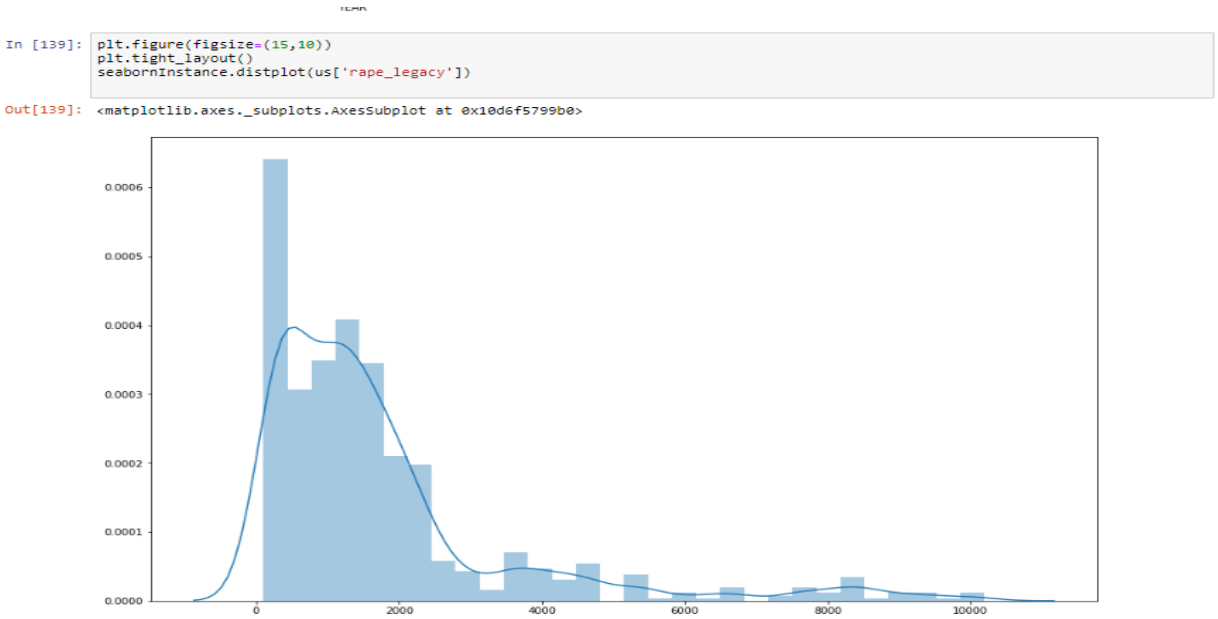
NORMALIZATION

```
In [255]: from sklearn import preprocessing
          mm_scaler = preprocessing.MinMaxScaler()
          X_train_minmax = mm_scaler.fit_transform(X_train)
          mm_scaler.transform(X_test)

          C:\Users\Dell\Anaconda3\lib\site-packages\sklearn\utils\validation.py:595: DataConversionWarning: Data with input dtype int64 w
          as converted to float64 by MinMaxScaler.
            warnings.warn(msg, DataConversionWarning)

Out[255]: array([[0.        ],
                 [0.        ],
                 [0.27272727],
                 ...,
                 [0.09090909],
                 [0.63636364],
                 [0.63636364]])
```

```
In [138]: us.plot(x='year', y='rape_legacy', style='o')
          plt.title('RAPE VS YEAR')
          plt.xlabel('YEAR')
          plt.ylabel('RAPE')
          plt.show()
```

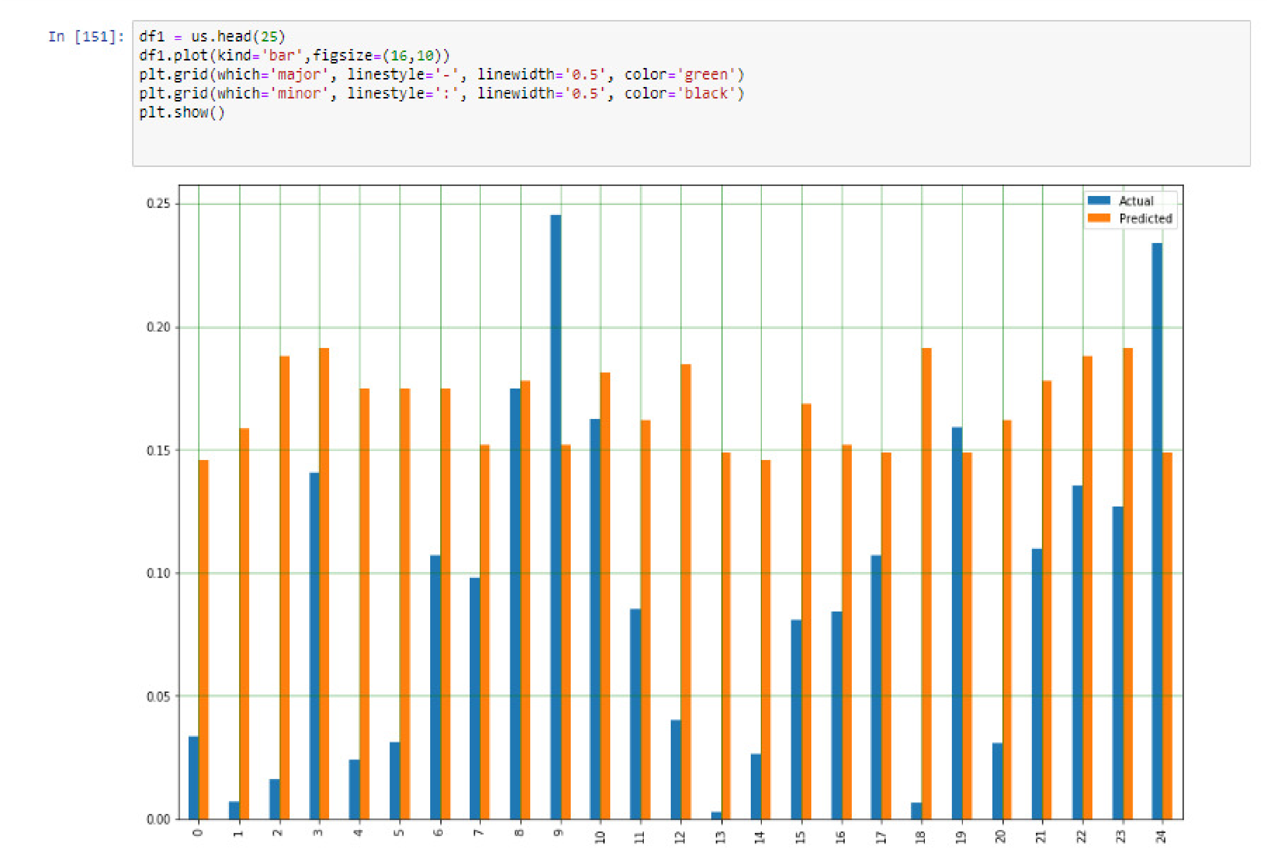
Figure 29

A distribution plot displays a distribution and range of a set of numeric values plotted against a dimension. You can display this chart in three different ways, you can just have the value points displayed showing the distribution, or you can display the bounding box which shows the range or use a combination of both. In the distribution plot shown below, you can see there a range and distribution of the rape values displayed for 0.0- 1.0. Each range and distribution box show how data values for a product group is distributed over the average rape rates per district.



```
In [139]: plt.figure(figsize=(15,10))
          plt.tight_layout()
          seabornInstance.distplot(us['rape_legacy'])

Out[139]: <matplotlib.axes._subplots.AxesSubplot at 0x10d6f5799b0>
```

Figure 30



```
In [151]: df1 = us.head(25)
          df1.plot(kind='bar',figsize=(16,10))
          plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
          plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
          plt.show()
```

Figure 31

This graph shows the relation between the actual and predicted values.

From the above calculations we can see that the y intercept has a value of 67881.2 and the slope has a value of -32.90874403. When we apply this to the equation, y=mx+c, to find the value of the dependable variable y. This will help us predict the rape rates per state for the future years. This can help us understand and better prepare for the future law enforcement.

Let us calculate the predicted value of the average value of rapes per state in USA. Once, this is done the value of the predicted variable is ~ 1537.18 per state for a single year.

.

4.7 COMPARISON OF INDIA VS USA

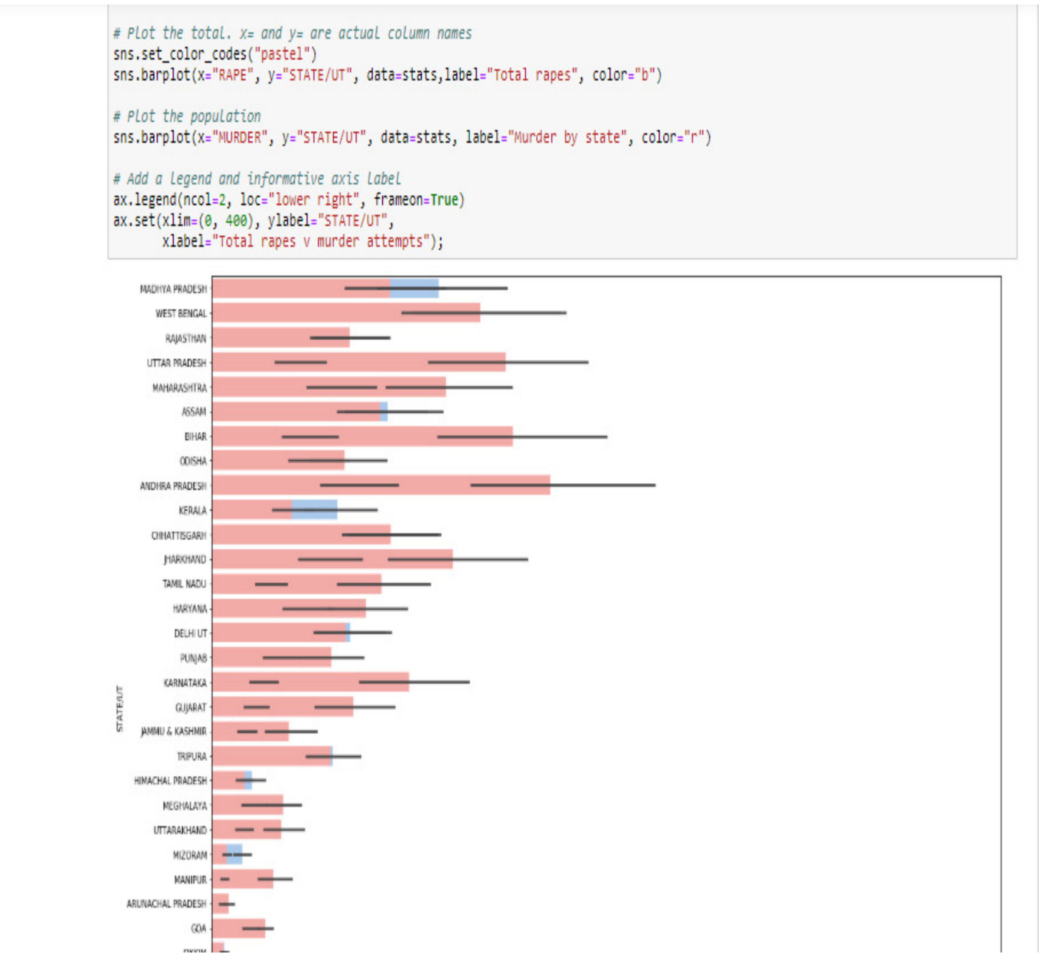
From conducting exploratory data analysis we have better understood the past data. To summarise, let’s look at the various interesting findings we have made.

- Firstly, we find that the state with the highest number of rapes is Madhya Pradesh with an approximate number of 3000
- The one with the lowest number of rapes is Lakshadweep with the approximate value of two per year.
- It is important to note that when you calculate the mean number of rape cases in India, you can see that over the decade it has been approximately ~50
- We then wanted to know the age group for women mostly targeted by rapists and we have found that the women who are at the prime of their age and are usually in colleges or work targeted. The average of age group targeted is from the age of 18-30.
- From the correlogram , you can see that the highest correlation exists between rapes and assault against a woman with the intent to outrage her modesty. This can lead one to believe that rape is conducted with the intent to shame the women or to cause question of her modesty.
- From the world map, you can notice that the number of rapes per capita in India is quite low. This is very contradictory to popular belief that women are not at all safe in India. However, this can also be due to a number of reasons, the main one being that the many number of rapes do not get justice or in many cases, do not even get reported in India due to the various cultural and religious constraints and judgements cast on a woman subject to such abuse.
- We find that the state with the highest number of rapes is California with an approximate number of 11000.
- It is important to note that when you calculate the mean number of rape cases in India, you can see that over the decade it has been approximately ~ 1800
- We then wanted to know the age group for women mostly targeted by rapists and we have found that the women who are at the prime of their age and are usually in colleges or work targeted. The average of age group targeted is from the age of 12-30..
- From the world map, you can notice that the number of rapes per capita in USA is very high. This is very contradictory to popular belief that women are all safe in USA

4.8 K MEANS CLUSTERING (INDIA)

K- means is one of the clustering methods. Clustering is an unsupervised machine learning algorithm. Clustering is nothing but grouping similar records together in a given dataset. K-Means is a non- hierarchical clustering algorithm. It is an iterative algorithm which tries to partition the dataset into a pre-defined number of non-overlapping clusters(k) by minimizing the sum of squared distances from each data point to the cluster centroid (Within sum of squared distances- WSS). Let us see how it works:

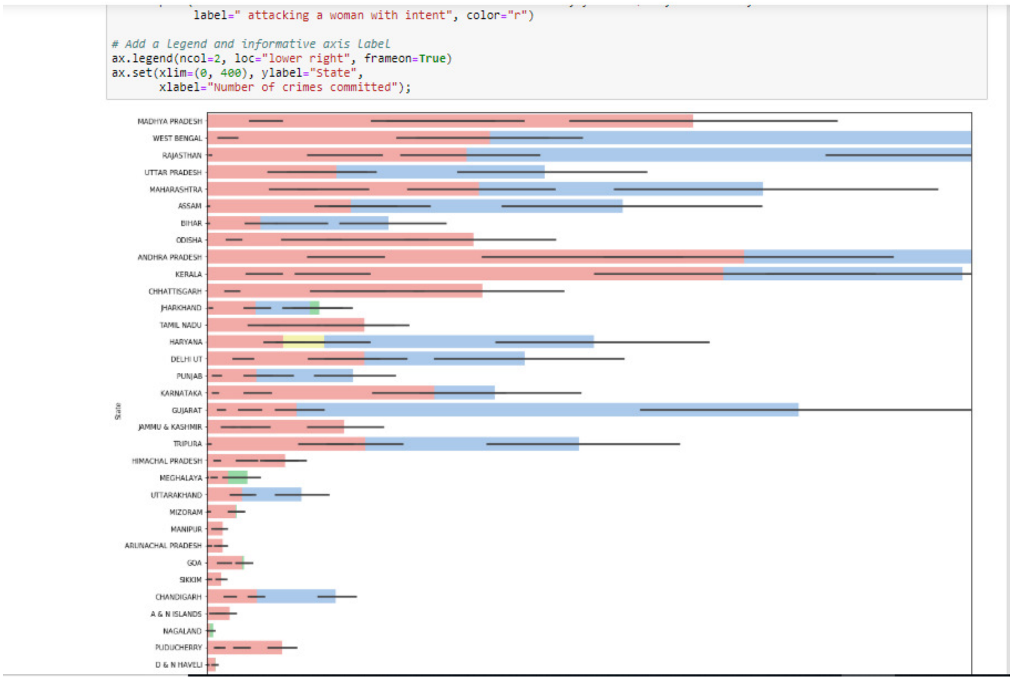
- Pre-define the no. of clusters (k)
- Initialize cluster centroids by randomly selecting K data points.
- Calculate the WSS distances between each data point to all the cluster centroids.
- Assign each data point to the closest centers
- Steps 3 and 4 will be iterated until there is no reassignment of data points is required.



```
# Plot the total. x= and y= are actual column names
sns.set_color_codes("pastel")
sns.barplot(x="RAPE", y="STATE/UT", data=stats,label="Total rapes", color="b")

# Plot the population
sns.barplot(x="MURDER", y="STATE/UT", data=stats, label="Murder by state", color="r")

# Add a legend and informative axis label
ax.legend(ncol=2, loc="lower right", frameon=True)
ax.set(xlim=(0, 400), ylabel="STATE/UT",
       xlabel="Total rapes v murder attempts");
```

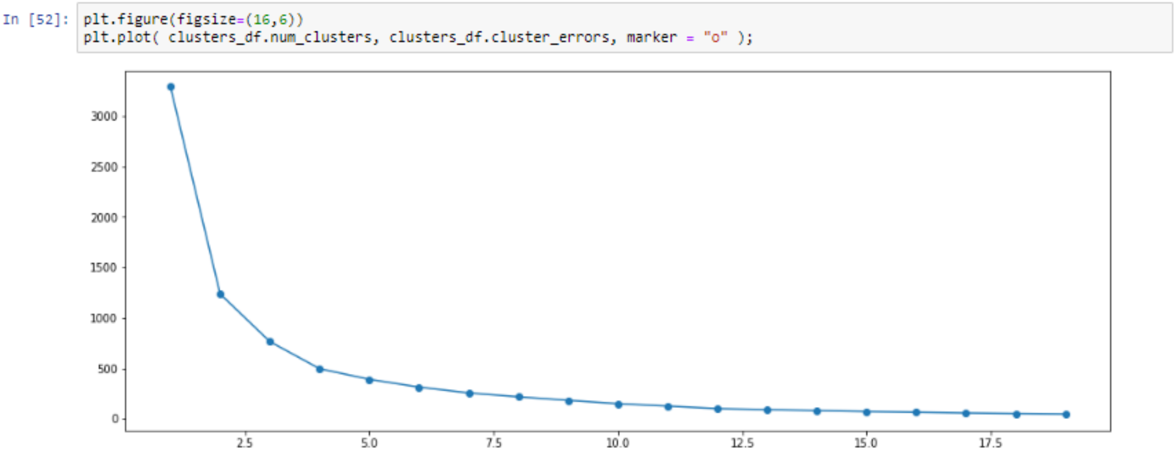
Figure 32

Now, let’s look at the same graph with multiple variables:



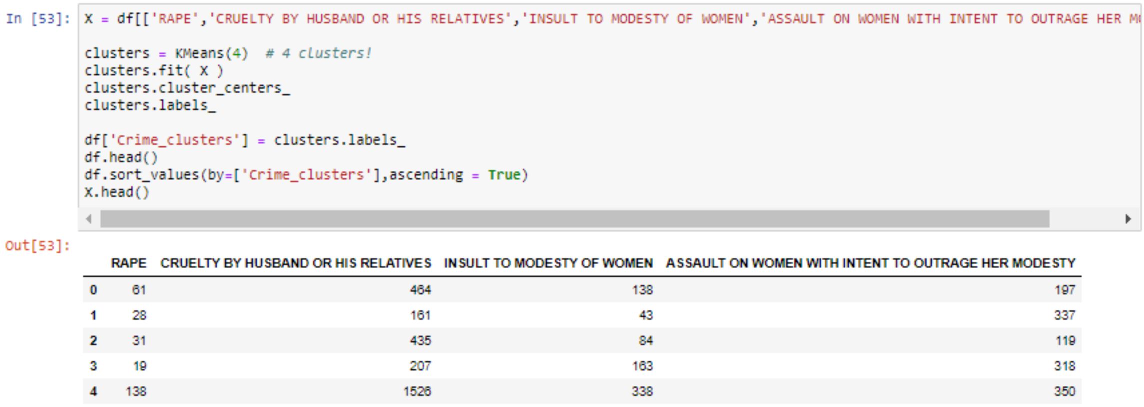
```
        label=" attacking a woman with intent", color="r")

# Add a legend and informative axis label
ax.legend(ncol=2, loc="lower right", frameon=True)
ax.set(xlim=(0, 400), ylabel="State",
       xlabel="Number of crimes committed");
```

Figure 33



```
In [52]: plt.figure(figsize=(16,6))
         plt.plot( clusters_df.num_clusters, clusters_df.cluster_errors, marker = "o" );
```

Figure 34



```
In [53]: X = df[['RAPE','CRUELTY BY HUSBAND OR HIS RELATIVES','INSULT TO MODESTY OF WOMEN','ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MO

         clusters = KMeans(4)  # 4 clusters!
         clusters.fit( X )
         clusters.cluster_centers_
         clusters.labels_

         df['Crime_clusters'] = clusters.labels_
         df.head()
         df.sort_values(by=['Crime_clusters'],ascending = True)
         X.head()
```

Out[53]:

	RAPE	CRUELTY BY HUSBAND OR HIS RELATIVES	INSULT TO MODESTY OF WOMEN	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY
0	61	464	138	197
1	28	161	43	337
2	31	435	84	119
3	19	207	163	318
4	138	1526	338	350

Figure 35

Now, we have to calculate the number of clusters that should be made.

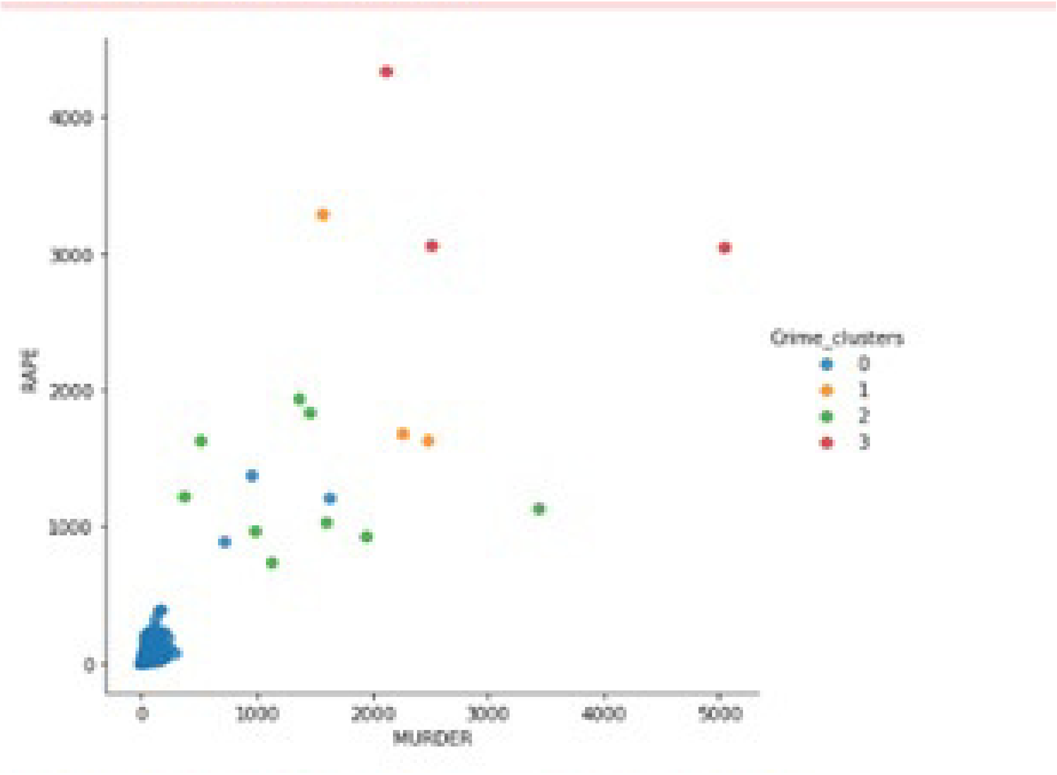


Figure 36

Here,we look at the four clusters formed on the different states in India when we bring in the two variables RAPE and MURDER to better understand the safer and most dangerous states in India.

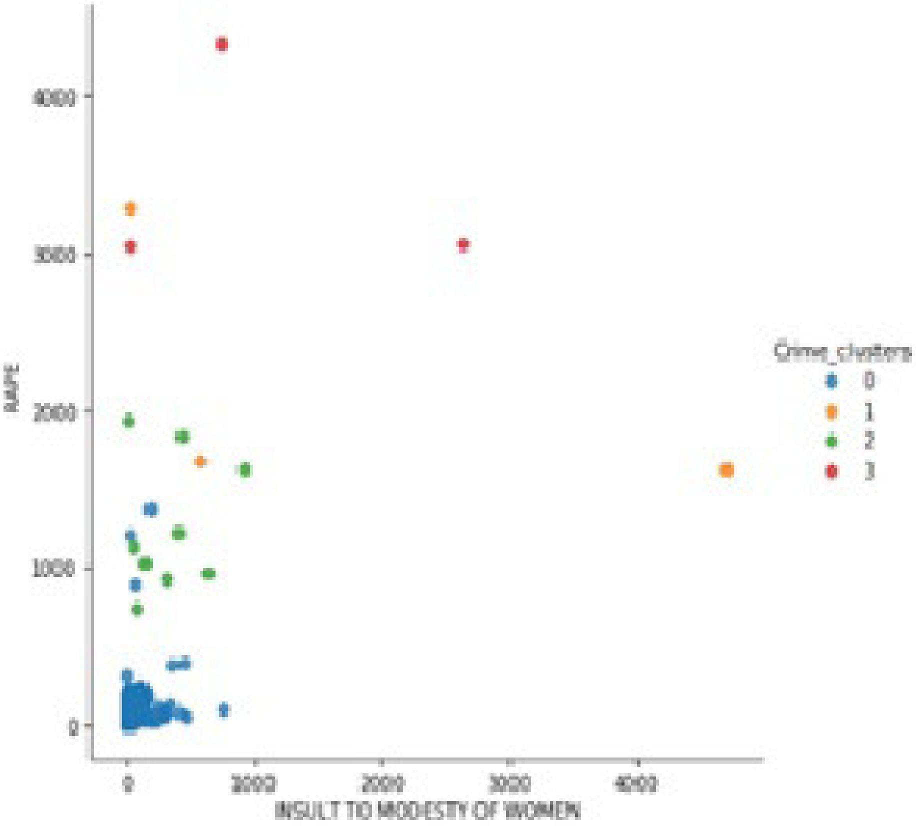


Figure 37

Here, we look at the four clusters formed on the different states in India when we bring in the two variables RAPE and INSULT TO THE MODESTY OF WOMEN to better understand the safer and most dangerous states in India.

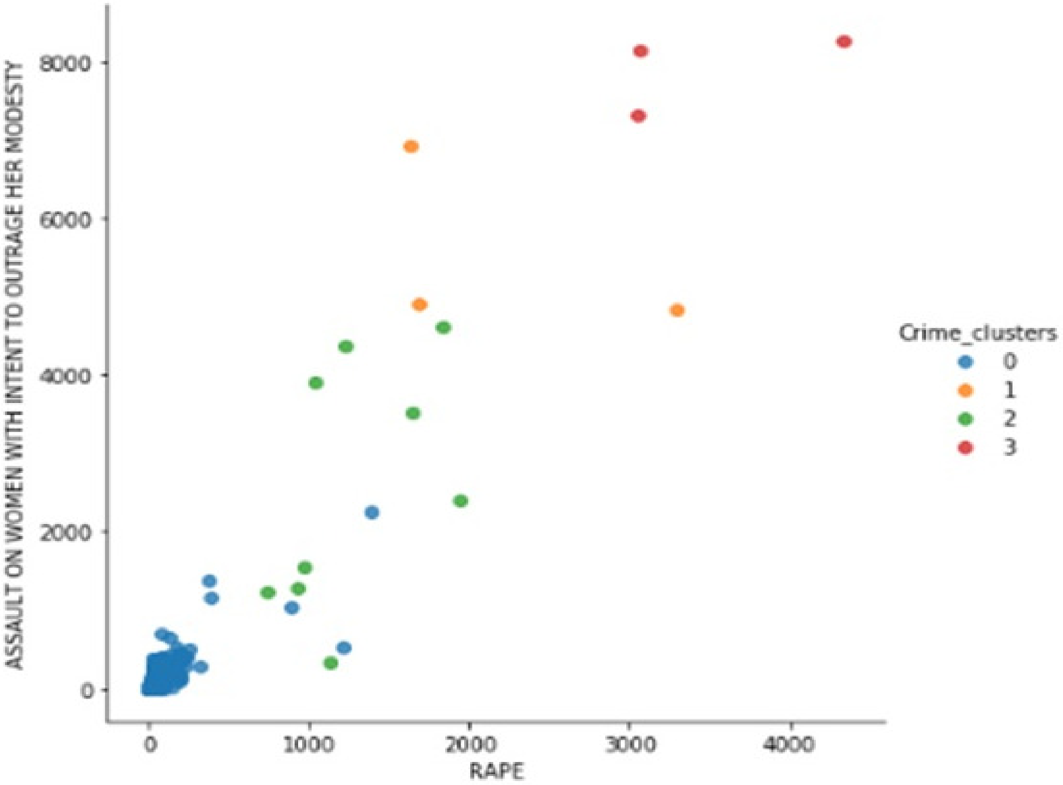


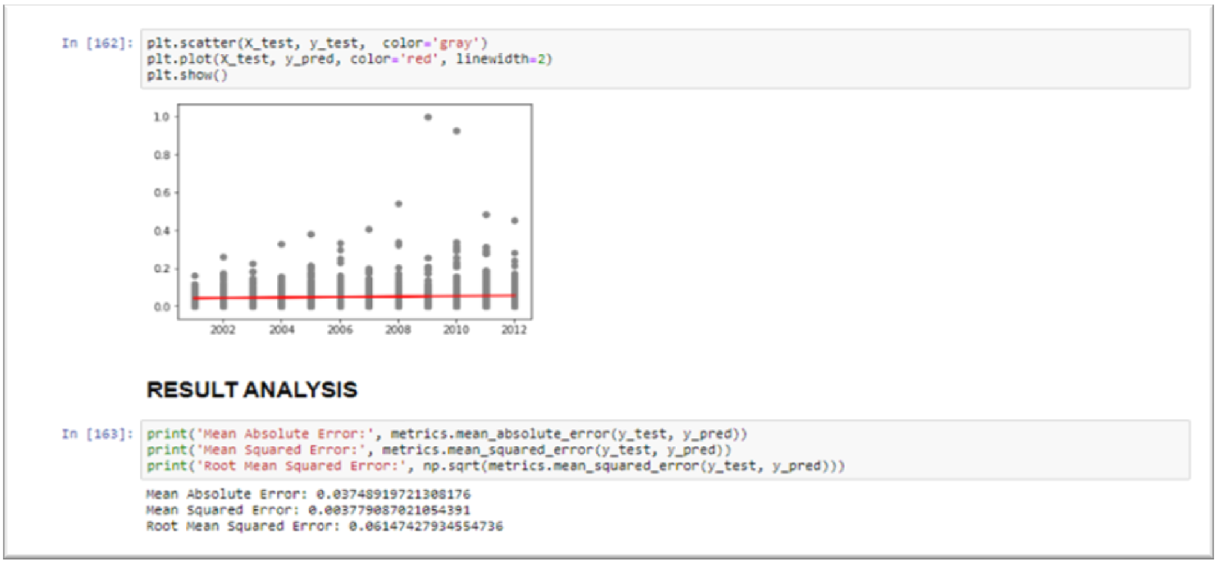
Figure 38

We look at the four clusters formed on the different states in India when we bring in the two variables RAPE and ASSAULT ON WOMEN WITH THE INTENT TO OUTRAGE HER MODESTY to better understand the safer and most dangerous states in India. These seem to create the most well defined clusters. This will be proven when you find out that these two variables have the highest correlation as well.

Chapter 5

Results and Discussion

5.1 RESULT ANALYSIS



```
In [162]: plt.scatter(X_test, y_test,  color='gray')
          plt.plot(X_test, y_pred, color='red', linewidth=2)
          plt.show()
```

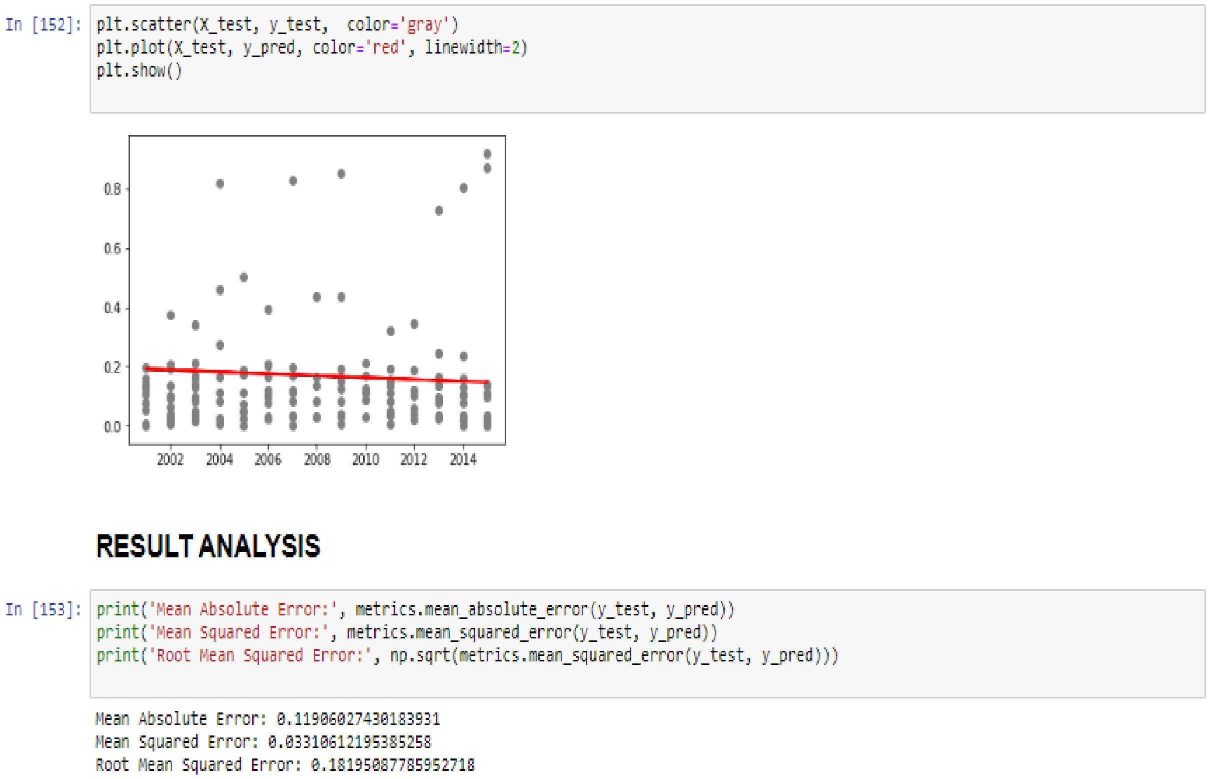
RESULT ANALYSIS

```
In [163]: print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
          print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
          print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 0.03748919721308176
Mean Squared Error: 0.003779087021054391
Root Mean Squared Error: 0.06147427934554736
```

Figure 39

From the analysis now we can find that the algorithm will give an average of 35 cases of rape in India.



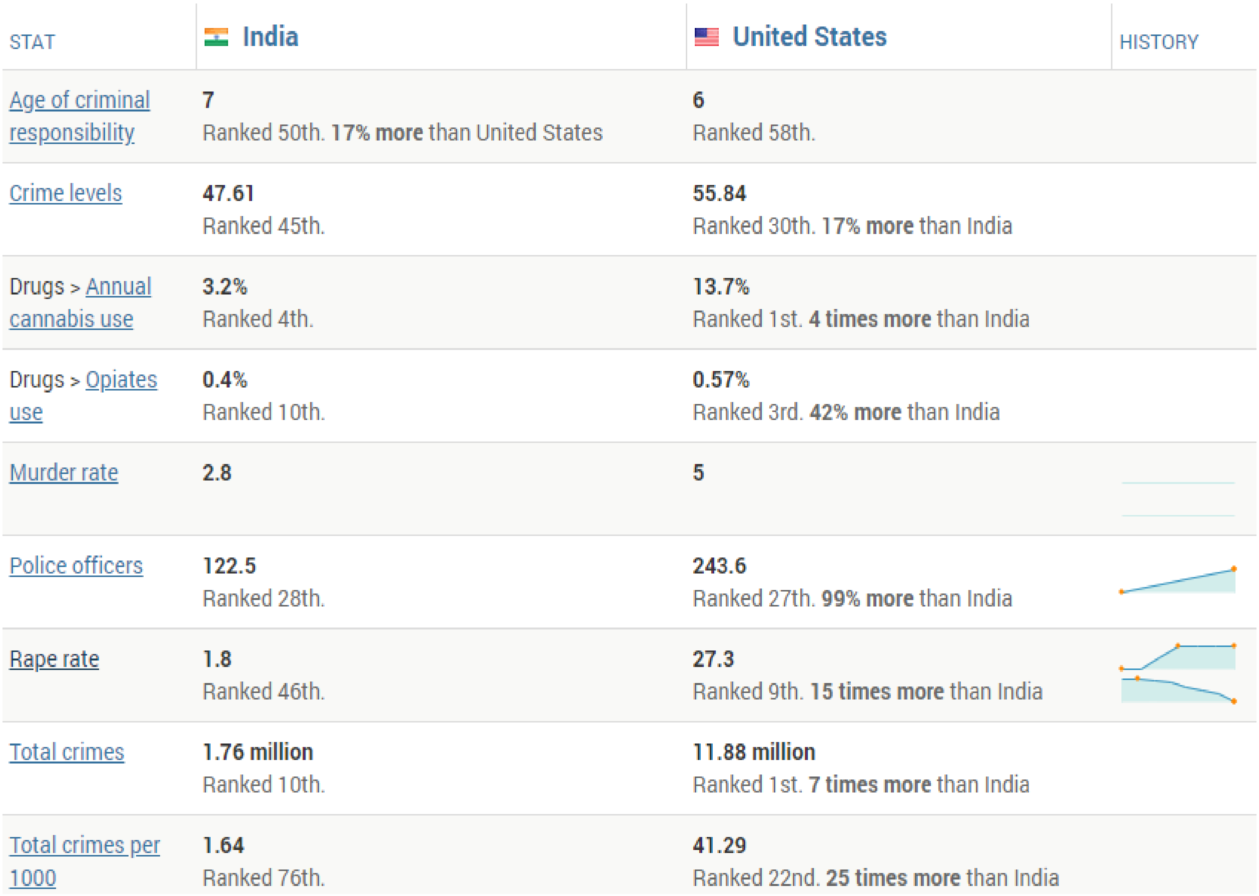
```
In [152]: plt.scatter(X_test, y_test,  color='gray')
          plt.plot(X_test, y_pred, color='red', linewidth=2)
          plt.show()
```

RESULT ANALYSIS

```
In [153]: print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
          print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
          print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 0.11906027430183931
Mean Squared Error: 0.03310612195385258
Root Mean Squared Error: 0.18195087785952718
```

Figure 40



STAT	India	United States	HISTORY
Age of criminal responsibility	**7** Ranked 50th. **17% more** than United States	**6** Ranked 58th.	
Crime levels	**47.61** Ranked 45th.	**55.84** Ranked 30th. **17% more** than India	
Drugs > Annual cannabis use	**3.2%** Ranked 4th.	**13.7%** Ranked 1st. **4 times more** than India	
Drugs > Opiates use	**0.4%** Ranked 10th.	**0.57%** Ranked 3rd. **42% more** than India	
Murder rate	**2.8**	**5**	
Police officers	**122.5** Ranked 28th.	**243.6** Ranked 27th. **99% more** than India	
Rape rate	**1.8** Ranked 46th.	**27.3** Ranked 9th. **15 times more** than India	
Total crimes	**1.76 million** Ranked 10th.	**11.88 million** Ranked 1st. **7 times more** than India	
Total crimes per 1000	**1.64** Ranked 76th.	**41.29** Ranked 22nd. **25 times more** than India	

Well let us see some statistics:

India has been characterised as one of the "countries with the lowest per capita rates of rape.

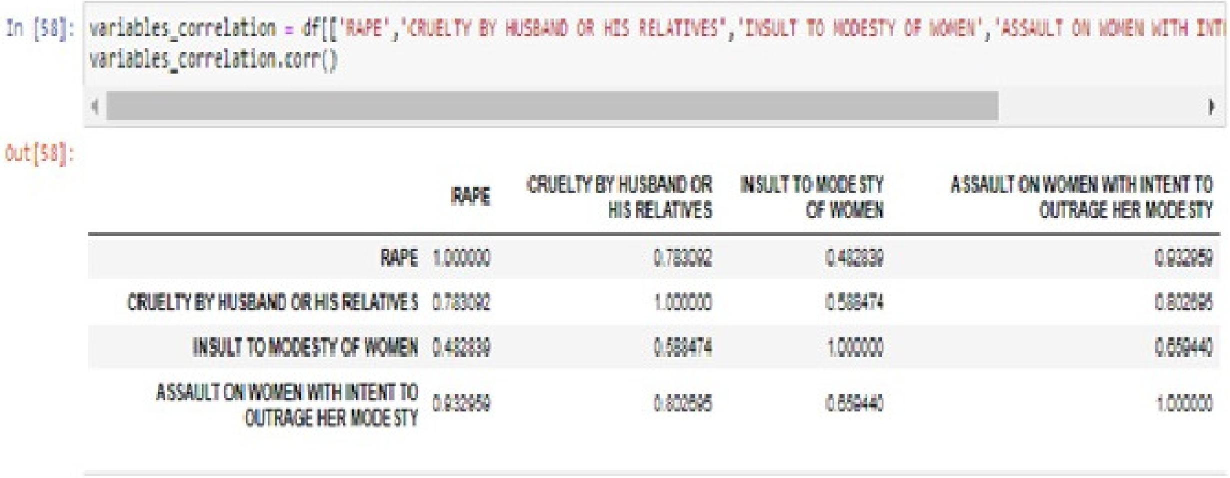
- Delhi has lower crime rate than NewYork [2]
- USA has 16 times more rapes than in India. Even if 50% of the rape cases in India are not reported, and all in USA are reported, they are still 8 times unsafe.
- 2 rapes per 100,000 people for India compared to
- 28.6 rapes/100,000 people for US and
- 24.1 rapes / 100,000 people for UK
- It is estimated that only 53% of rape crimes are reported in India.
- But a UN study estimates just **11%** of rape and sexual assault cases **worldwide** are ever reported.
- 10% in France, 18.3% in USA, and 15% in UK and 13% in Europe are ever reported, rest remain unreported.
- Also, in UK, 26% of all sexual offences (including rape) reported to police are not even recorded as crimes.
- In some places in USA, the victims were forced to take a polygraph test before rape is reported.

CORRELATION

Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it’s a measure of how things are related. The study of how variables are correlated is called correlation analysis.The main result of a correlation is called the **correlation coefficient** (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).

While correlation coefficients are normally reported as r = (a value between -1 and +1), squaring them makes then easier to understand. The square of the coefficient (or r square) is equal to the percent of the variation in one variable that is related to the variation in the other. After squaring r, ignore the decimal point. An r of .5 means 25% of the variation is related (.5 squared =.25). An r value of .7 means 49% of the variance is related (.7 squared = .49).

A correlation report can also show a second result of each test - statistical significance. In this case, the significance level will tell you how likely it is that the correlations reported may be due to chance in the form of random sampling error. If you are working with small sample sizes, choose a report format that includes the significance level. This format also reports the sample size.

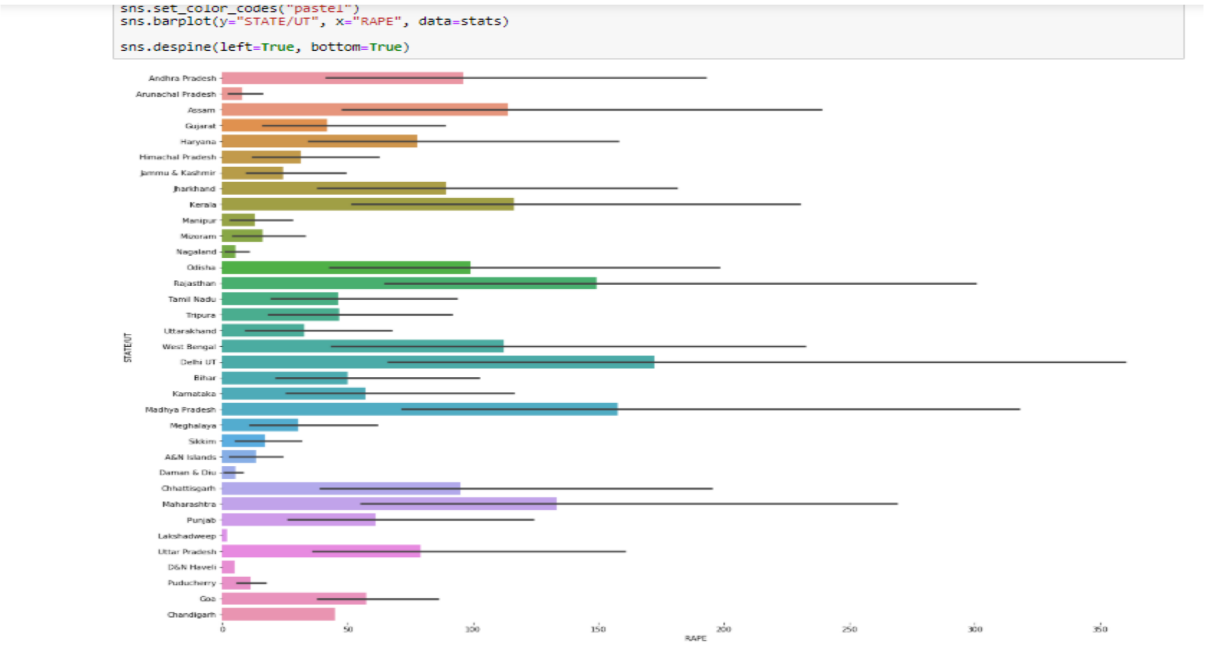


```
In [58]: variables_correlation = df[['RAPE','CRUELTY BY HUSBAND OR HIS RELATIVES','INSULT TO MODESTY OF WOMEN','ASSAULT ON WOMEN WITH INT
         variables_correlation.corr()
```

Out[58]:

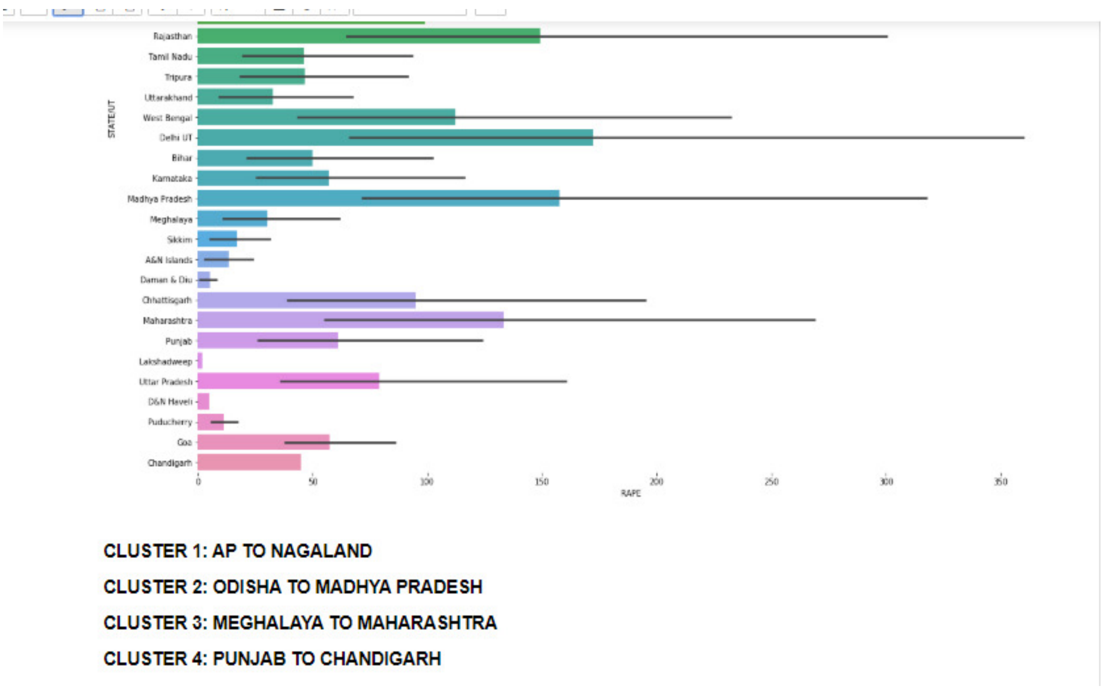
	RAPE	CRUELTY BY HUSBAND OR HIS RELATIVES	INSULT TO MODE STY OF WOMEN	A SSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODE STY
RAPE	1.000000	0.783092	0.482839	0.932959
CRUELTY BY HUSBAND OR HIS RELATIVE S	0.783092	1.000000	0.588474	0.802595
INSULT TO MODESTY OF WOMEN	0.482839	0.588474	1.000000	0.659440
ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODE STY	0.932959	0.802595	0.659440	1.000000

Figure41



```
sns.set_color_codes("pastel")
sns.barplot(y="STATE/UT", x="RAPE", data=stats)

sns.despine(left=True, bottom=True)
```

Figure 42



CLUSTER 1: AP TO NAGALAND

CLUSTER 2: ODISHA TO MADHYA PRADESH

CLUSTER 3: MEGHALAYA TO MAHARASHTRA

CLUSTER 4: PUNJAB TO CHANDIGARH

Figure 43

From the following, we understand that these are the safest to the most dangerous sates in our country. This study shows the various characteristics that affect the conditions of rape in

India and does an in depth study of the comparisons of both the countries, India and USA. India is known as a third world country and is said to be a ‘developing country’. USA is considered to be a first world country and economically developed. However, we see that even though India is considered to be a third world country, women are much safer. A recent study showed that ~50 of rapes go unreported while in the US ~10 percent only get reported. So even if we were to double the amount of rape committed per capita, you can see that India is considerably safer than the US. This can be because of a number of reasons.

In our predictions, we can see that the y intercept for USA has a value of 67881.2 and the slope has a value of -32.90874403. When we apply this to the equation, y=mx+c, to find the value of the dependable variable y. This will help us predict the rape rates per state for the future years. This can help us understand and better prepare for the future law enforcement. The predicted value of the average value of rapes per state in USA, is found to be the value of ~ 1537.18 per state for a single year.

From the calculations we made for India, we can see that the y intercept has a value of -2.585 and the slope has a value of 0.0001. When we apply this to the equation, y=mx+c, to find the value of the dependable variable y. This will help us predict the rape rates per state for the future years. This can help us understand and better prepare for the future law enforcement. Let us calculate the predicted value of the average value of rapes per state in India. Once, this is done the value of the predicted variable is ~34.7 per state for a single year.

India has a rape rate per capita of 1.8 and is ranked 46th in the world among crimes of rape committed. USA has a rape rate per capita of 27.3 and is ranked 9th. 15 times more than India.

Chapter 6

Conclusion and Scope for Future work

6.1 CONCLUSION:

From conducting exploratory data analysis we have better understood the past data. To summarise, let’s look at the various interesting findings we have made.

- Firstly, we find that the state with the highest number of rapes is Madhya Pradesh with an approximate number of 3000
- The one with the lowest number of rapes is Lakshadweep with the approximate value of two per year.
- It is important to note that when you calculate the mean number of rape cases in India, you can see that over the decade it has been approximately ~50
- We then wanted to know the age group for women mostly targeted by rapists and we have found that the women who are at the prime of their age and are usually in colleges or work targeted. The average of age group targeted is from the age of 18-30.
- From the correlogram , you can see that the highest correlation exists between rapes and assault against a woman with the intent to outrage her modesty. This can lead one to believe that rape is conducted with the intent to shame the women or to cause question of her modesty.
- From the world map, you can notice that the number of rapes per capita in India is quite low. This is very contradictory to popular belief that women are not at all safe in India. However, this can also be due to a number of reasons, the main one being that the many number of rapes do not get justice or in many cases, do not even get reported in India due to the various cultural and religious constraints and judgements cast on a woman subject to such abuse.
- We find that the state with the highest number of rapes is California with an approximate number of 11000.
- It is important to note that when you calculate the mean number of rape cases in India, you can see that over the decade it has been approximately ~ 1800
- We then wanted to know the age group for women mostly targeted by rapists and we have found that the women who are at the prime of their age and are usually in colleges or work targeted. The average of age group targeted is from the age of 12-30..
- From the world map, you can notice that the number of rapes per capita in USA is very high. This is very contradictory to popular belief that women are all safe in the USA

From the following, we understand that these are the safest to the most dangerous sates in our country. This study shows the various characteristics that affect the conditions of rape in India and does an in depth study of the comparisons of both the countries, India and USA.

India is known as a third world country and is said to be a ‘developing country’. USA is considered to be a first world country and economically developed. However, we see that even though India is considered to be a third world country, women are much safer. A recent study showed that ~50 of rapes go unreported while in the US ~10 percent only get reported. So even if we were to double the amount of rape committed per capita, you can see that India is considerably safer than the US. This can be because of a number of reasons.

In our predictions, we can see that the y intercept for USA has a value of 67881.2 and the slope has a value of -32.90874403. When we apply this to the equation, y=mx+c, to find the value of the dependable variable y. This will help us predict the rape rates per state for the future years. This can help us understand and better prepare for the future law enforcement. The predicted value of the average value of rapes per state in USA, is found to be the value of ~ 1537.18 per state for a single year.

From the calculations we made for India, we can see that the y intercept has a value of -2.585 and the slope has a value of 0.0001. When we apply this to the equation, y=mx+c, to find the value of the dependable variable y. This will help us predict the rape rates per state for the future years. This can help us understand and better prepare for the future law enforcement. Let us calculate the predicted value of the average value of rapes per state in India. Once, this is done the value of the predicted variable is ~34.7 per state for a single year.

India has a rape rate per capita of 1.8 and is ranked 46th in the world among crimes of rape committed. USA has a rape rate per capita of 27.3 and is ranked 9th, 15 times more than India.

6.2 SCOPE FOR FUTURE WORK

Our study sheds light on the state of the safety of women for the two first and third world countries. However, one cannot narrow down the complexities of the reasons behind rape to the economic and social conditions. This is definitely part of a much greater venture to understand the state of various countries that fall under different economic backgrounds. In the future, one must take into account the various cases that go unreported in the countries for a better assessment. It has been reported that ~50 percent of the cases in India go unreported and `~12 percent of cases in the US are actually reported. Finding the actual datasets can help produce better and accurate predictions. This study provided us with a surprising fact that the number of rapes per capita in the USA is 15 times more than that in India. This begs the question whether feminism, social and economic standards have a large role to play as assumed by the general public. This study should be done on a number of different countries to better understand the possible reasons behind crime against women.

REFERENCES

1. Approach of Predictive Modeling on Crime Against Women Problem ,Priya Gandhi, Shayog Sharma (Department of Computer Science & Engineering, Geeta Engineering college, panipat, india .)
2. Crime rate prediction using data clustering algorithms , Omkar Vaidya, Sayak Mitra, Raj Kumbhar, Suraj Chavan, Mrs. Rohini Patil (Department of Computer Engineering, Terna Engineering College, Mumbai University, India)
3. Crime against indian women –women crime, susceptibility indexes (wcsi): a principal component analysis , prarthna agarwal goel , vandana yadav , delhi university, india.
4. Prediction of Crime Rate Using Data , Clustering Technique ,A. Anitha (VIT University)
5. Research Corner , Regression Analysis for Prediction: Understanding the Process , Phillip, Associate Professor, Hardin-Simmons University, Department of Physical Therapy, Abilene, TX Professor & Shelton-Lacewell Endowed Chair, Hardin-Simmons University, Department of Physical Therapy, Abilene, TX
6. Yamuna and Sudha Bhuva Neswa ,DataMining techniques to Analyze and predict crimes
7. Predictive Analytics: Opportunities, Challenges and Use Cases by Lityx.
8. Applied Predictive Analytics "Principles and Techniques for the Professional Data Analyst" by Dean Abbott
9. Understanding Environmental Factors that Affect Violence in Salinas, California Clarke, Jason A.; Onufer, Tracy L. Dec,2009, Malathi, Dr. S. Santhosh Baboo , 2011.
10. Poisson-Based Regression Analysis of Aggregate Crime Rates D. Wayne Osgood,2000.
11. Forecasting Crime: A City Level Analysis John V. Pepper Department of Economics University of Virginia.Ch-6.
12. Analysis of Criminal Behaviour Using a Logistic Regression Model Maria Jofre Department of Industrial Engineering, University of Chile.
13. Jiawei Han, Micheline Kamber, Jian Pei, "DATA MINING Concepts and Techniques", Publisher: Morgan Kaufmann, Third Edition 2012, ISBN 978- 0-12-381479-1
14. R. Kiani, S. Mahdavi, A. Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", *(IJARAI)* International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.8, 2015, PP. 11-17.
15. V. Jain, Y. Sharma, A. Bhatia, V. Arora, "Crime Prediction using K-means Algorithm", Global Research And Development Journal for Engineering, Volume 2 Issue 5, April 2017, PP. 206–209.
16. T. Sonawanev, S. Shaikh, S. Shaikh, R. Shinde, A. Sayyad, "Crime Pattern Analysis, Visualization And Prediction Using Data Mining", IJARIIE, Vol-1 Issue-4 2015, PP. 681-686.
17. J. Agarwal, R. Nagpal, R. Sehgal, "Crime Analysis using K-means Clustering", International Journal of Computer Applications, Volume 83 – No4, December 2013.
18. Malathi A., Dr. S. Santhosh Baboo, "An Enhanced Algorithm to Predict a Future Crime using Data Mining", International Journal of Computer Applications, Volume 21-No.1, May 2011, PP. 1-6.
19. M. Gupta, B. Chandra and M. P. Gupta, "Crime Data Mining for Indian Police Information System", Computer Society of India, 2008, PP. 388-397.
20. Grady Booch, James Rumbaugh, Ivar Jacobson, "The Unified Modeling Language User Guide", Publisher: Addison Wesley, First Edition October 20, 1998, ISBN 0-201-57168-4
21. *https://www.kaggle.com/c/sf-crime/data*

22. L. Ding et al., “PerpSearch: an integrated crime detection system”, 2009 *IEEE* 161-163 ISI 2009, June 8-11, 2009, Richardson, TX, USA.
23. D.E. Brown 1998, ”The regional crime analysis program (RECAP): A frame work for mining data to catch criminals", In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, Pp. 2848-2853
24. Yamuna.S, N.Sudha Bhauvaneswari D, Data mining Techniques to Analyze and Predict Crimes, International Journal of Engineering And Science(IJES) Vol-1,Issue-2,PP 243-247.
25. Sylvia Walby, Improving the statistics on violence against women, Statistical Journal of the united nation ECE 22(2005)193-216.
26. S.Bewley, J.Friend and G.Mezey, Eds, Violence against Women, London: Royal College of Obestricians and Gynaecologists, 1997.
27. Malathi.A and Dr.S.Santhosh Baboo. Article: an enhanced algorithm to predict a future crime using data mining. International Journal of Computer Applications,21(1):1-6, May2011.Published by foundation of Computer Science.
28. A.Buczak and C.Gifford, ‘Fuzzy association rule mining for community crime pattern discovery’, in ACM SIGKDD workshop on intelligence and security Informatics, Washington, D.C., 2010, PP.1-10.
29. Crimereports.com,2015.[online].Available:http://www.crimereports.com.[Accessed:20-May-2015].
30. Anshu Sharma,Shilpa Sharma 2012 An Intelligent Analysis of Web Crime Data using Data Mining, International Journal of Engineering And Innovative Technology(Ijeit)2(3)
31. Devendra Kumar Tayal et al., Crime detection and criminal identification in India using data mining techniques,AI & Soc(2015) 30, pp.117-127.
32. Roslin V.Hausk and Hsinchun Chen., Coplink: A Case of Intelligent Analysis and Knowledge Management, Proceedings of International Conference on Information Systems, 1999, pp.15-28.
33. Rasoul kiani, Siamak mahdavi, Amin Keshavarzi, Analysis and Prediction of Crimes by clustering and Classification.(IJARAI) International Journal of Advanced Research in Artificial Intelligence,Vol-4,No:8,2015
34. Amarnathan, L.C.(2003) Technological Advancement: Implications for the Crime, the Indian Police Journal, April-June.

TEAM NUMBER – 1

Name: PREETI RACHEL JASPER

Register Number: 17BCS0003

Date of Birth: 19/12/1997

Blood Group: B+ve

Phone Number: 9003306860

Email: preetirachel.jasper2017@vitstudent.ac.in

Address: BETHELS, TRA C6, TC 26/2139,

STATUE ROAD, TRIVANDRUM-1

Pin Code: 695001

TEAM NUMBER – 2

Name : DINESH S

Register Number: 17BCS0123

Date of Birth: 15-09-1999

Blood Group: B+

Phone Number: 7871114428

Email: dineshs15091999@gmail.com

Address: No.9,pillaiyar kovil street arunagiripettai,
Mothakal post village,mottupalayam,kammavanpettai,
Vellore-632319

TEAM NUMBER – 3

Name: KAMESHWARAN E

Register Number:17BCS0074

Date of Birth:16-02-2000

Blood Group: O+

Phone Number: 9486625185

Email:kamesh1620@gmail.com

Address: No.40/47 Mettu street Pazhampet,Chetpet
Tiruvannamalai District,606801,Tamil Nadu, India