

**EDA of Rape Crime In India**

“UN Women states that 1 in 3 women face some kind of sexual assault at least once in their lifetime.” A rape occurs every 20 mins. Yet most women and girls do not talk about this abuse for multiple reasons such as fear of society, culture, victim-blaming, fear of the police, tedious formal procedures, etc. Here we are analysing the rape crimes in different states of India from 2001-2010.

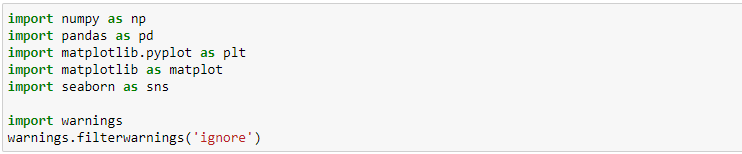
**PROBLEM STATEMENT:**

Perform complete EDA with observations and insights of Rape victim’s data in india.

**ABOUT THE DATASET:**

This dataset contains State-wise data of Rape Victims from 2001 to 2010 with their age in different area of India.

**IMPORTING LIBRARIES:**



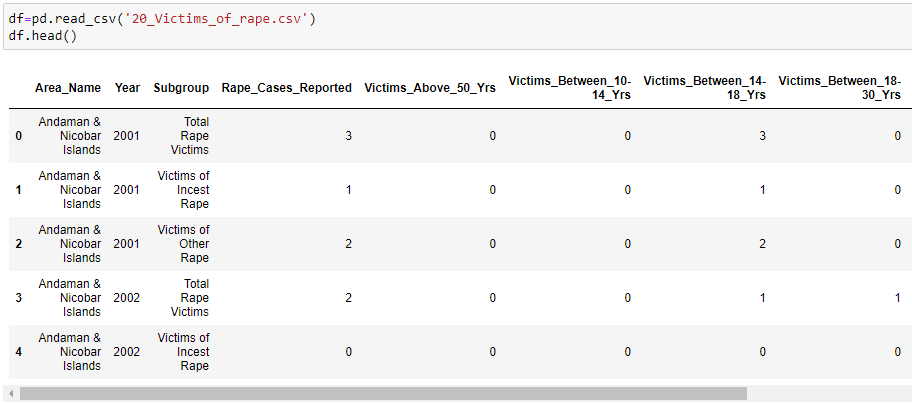
**Numpy** module is mainly used for working with numerical data. It provides us with a powerful object known as an Array. With Arrays, we can perform mathematical operations on multiple values in the Arrays at the same time, and also perform operations between different Arrays, similar to matrix operations.

**Pandas** module is used for working with tabular data. It allows us to work with data in table form, such as in CSV or SQL database formats. We can also create tables of our own, and edit or add columns or rows to tables. Pandas provides us with some powerful objects like DataFrames and Series which are very useful for working with and analyzing data.

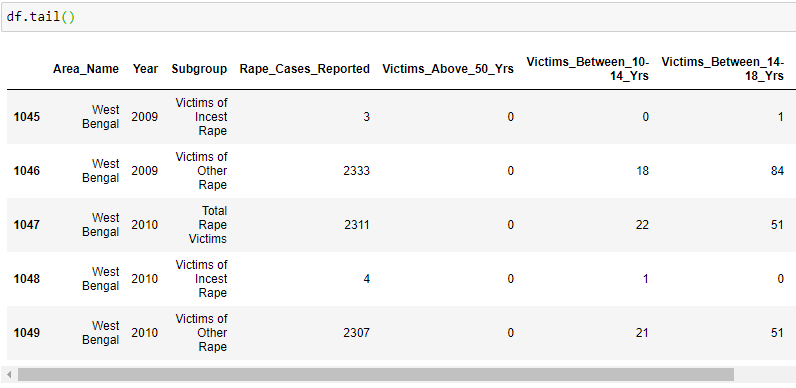
**Matplotlib** module is used for data visualization. It provides functionality for us to draw charts and graphs, so that we can better understand and present the data visually.

**Seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

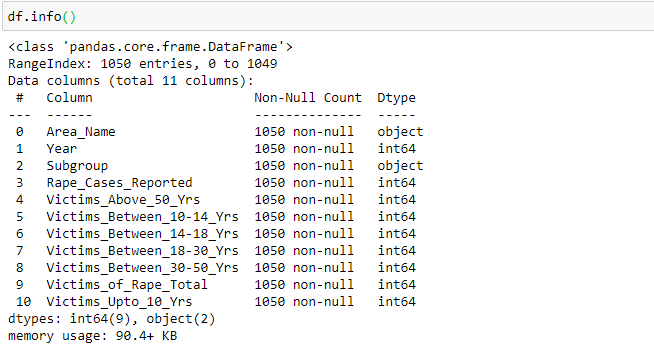
**READING AND EXPLORING THE DATASET:**

****

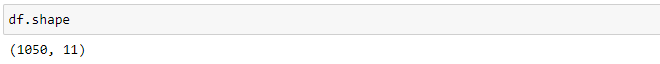
*#It gives the value of first 5 rows for each column.*



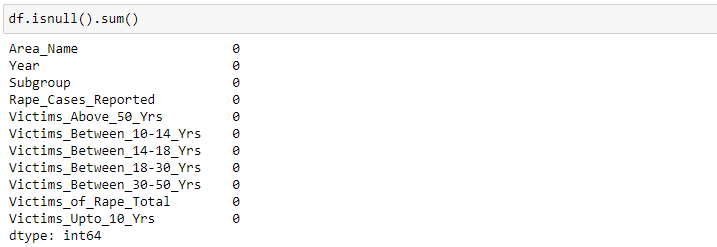
*#It gives the value of last 5 rows for each column.*



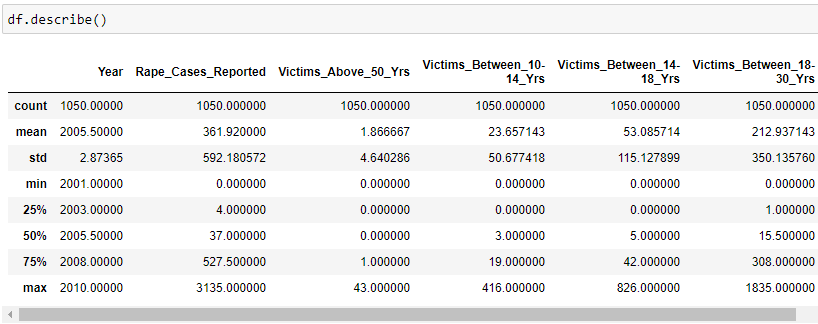
*#It gives the information of datatype.*



*#It gives total number of ROWS and COLUMNS present in the dataset.*



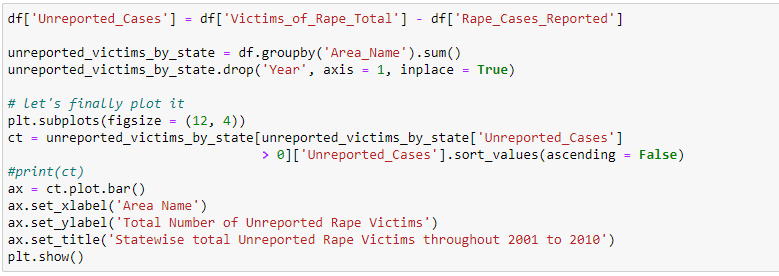
*#There is no null value present in the dataset.*

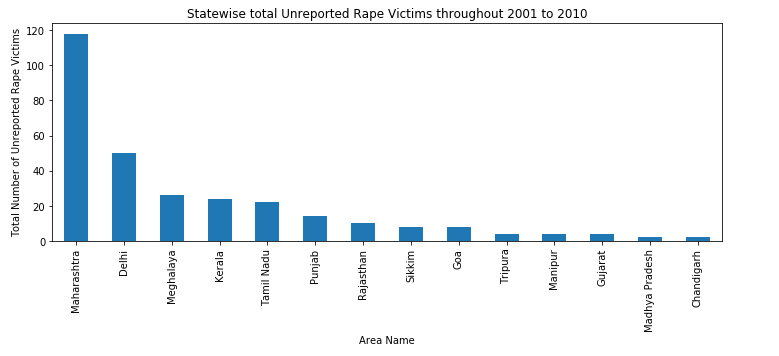


*#It gives some basic statistical details like count, percentile, mean, std etc.*

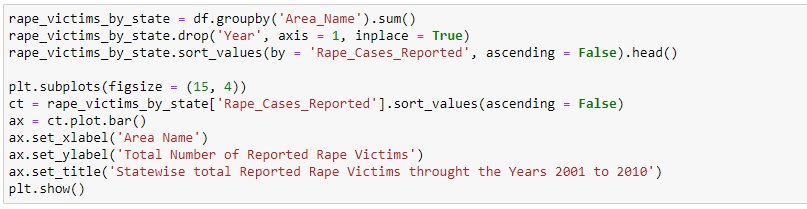
**EXPLORATORY DATA ANALYSIS:**

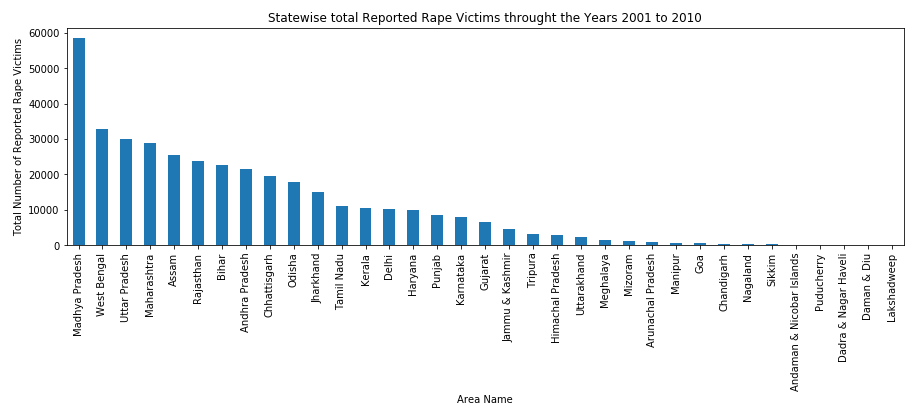
Exploratory Data Analysis (EDA) is an approach to analysing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modelling task.





*#Firstly, This Bar Chart shows no. of unreported cases of all states in descending order.*

**



*#From this bar chart we can say that these are the top five states in country with number of cases.*

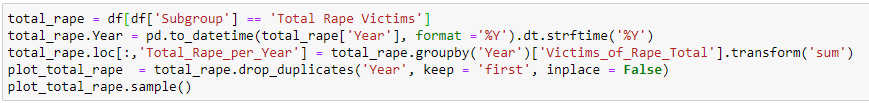
*Madhya Pradesh-58512*

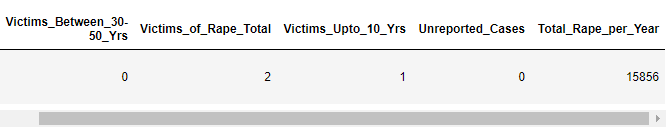
*West Bengal-32756*

*Uttar Pradesh-30106*

*Maharashtra-28892*

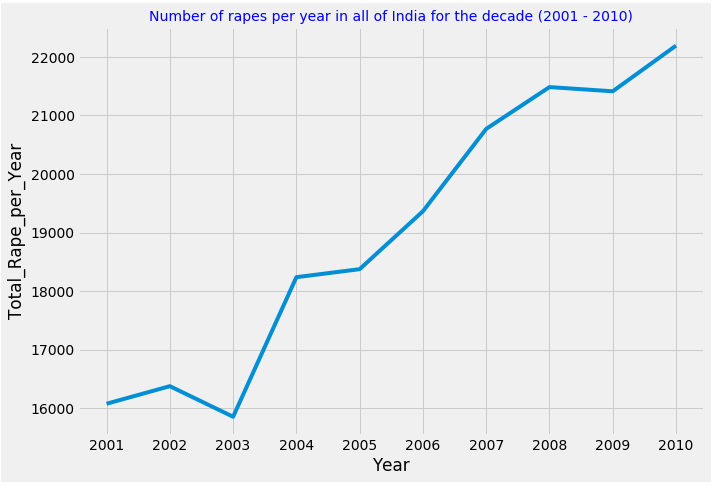
*Assam-25524*

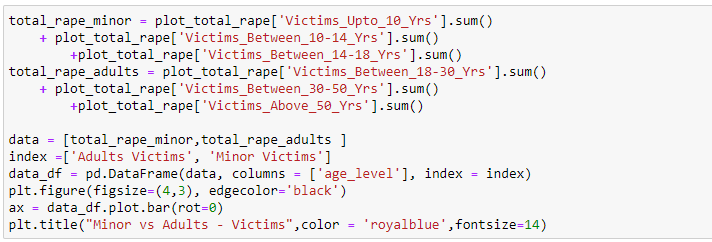




*#Here we are adding ‘Total\_Rape\_per\_Year’ column for better understanding of dataset across all categories.*

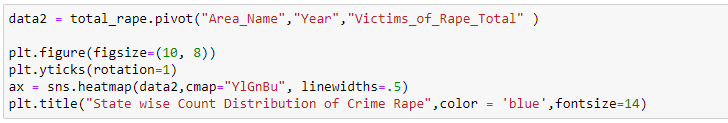


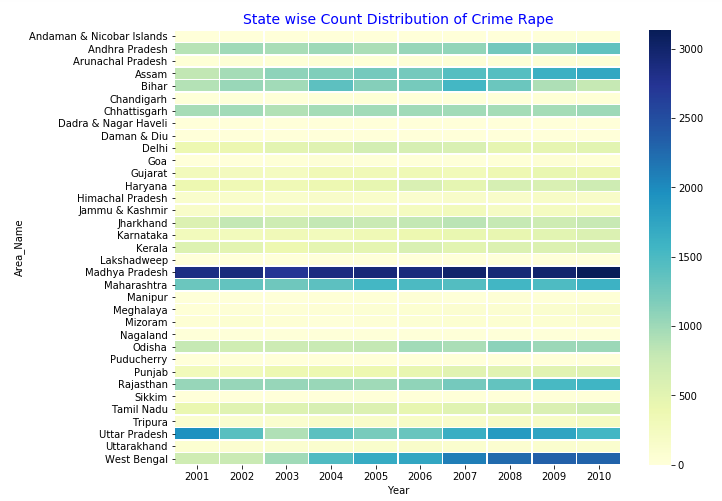




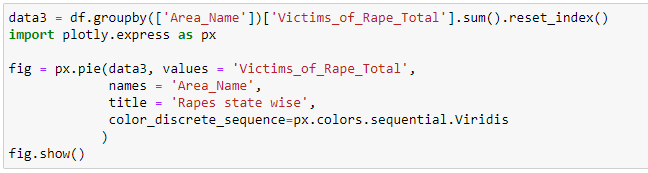


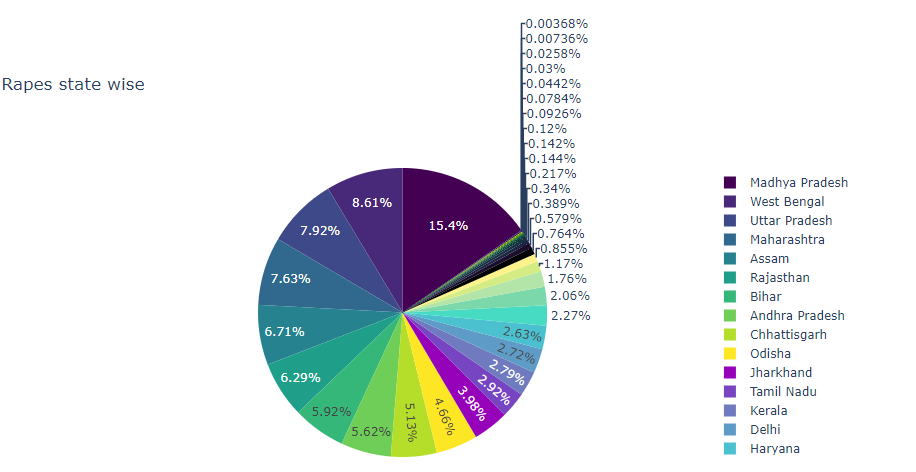
*#we can observe that number of Adults Victims are almost double compared to Minor Victims.*



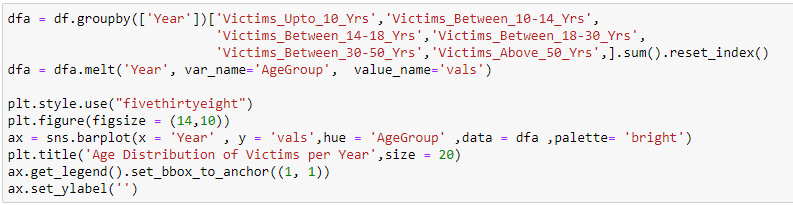


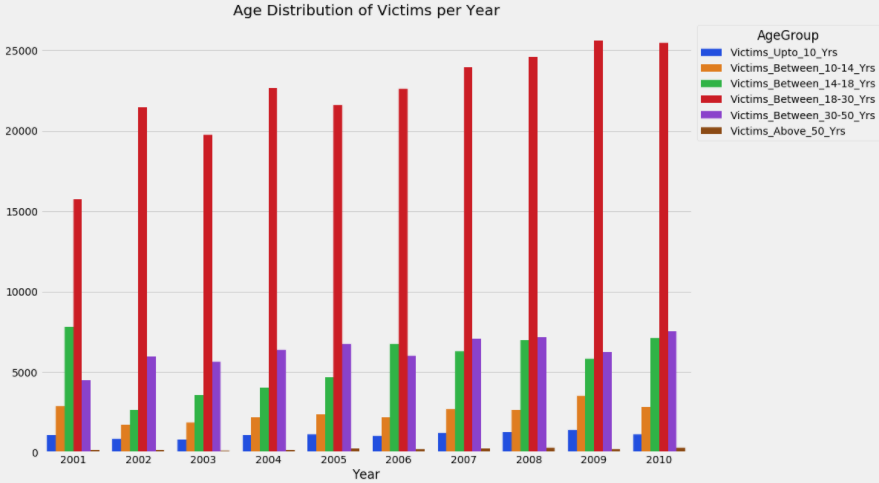
*#States of "Madhya Pradesh, Maharashtra, Uttar Pradesh, Assam, Bihar and West Bengal" were most impacted during this period. While Madhya Pradesh didnt improve its law and order situation, West Bengal situation went from bad to worse the most. Situation in states of "Odisha, Tamil Nadu, Punjab, Jharkhand, Chhattisgarh and Andhra Pradesh" worsend during this period. Baring state of Assam, other North Eastern states don't report Rape related crime in excess. The states like Bihar and UP have in-consistent graph in cases of rape. Few years it improved while few years it was really dark times.*



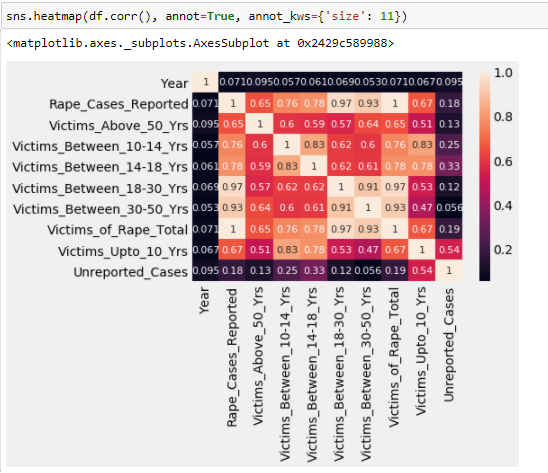


*# This Pie-Chart gives idea about total percentage of rape cases across all states.*





*#Age group of 18 - 30 years is more prone to be victim of the rape. We can observe that rape cases increase over the years.*



*#A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colours. Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are; that is as one increases so does the other and the closer to 1 the stronger this relationship is. A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases.*

*Here we can observe that victims of age group of 18 - 30 years and 30 – 50 years are highly corelated with rape cases reported and same victims are less corelated with unreported cases.*

***#SOURCE\_CODE:***

*import numpy as np*

*import pandas as pd*

*import matplotlib.pyplot as plt*

*import matplotlib as matplot*

*import seaborn as sns*

*import warnings*

*warnings.filterwarnings('ignore')*

*df=pd.read\_csv('20\_Victims\_of\_rape.csv')*

*df.head()*

*df.tail()*

*df.info()*

*df.shape*

*df.isnull().sum()*

*df.describe()*

*df['Unreported\_Cases'] = df['Victims\_of\_Rape\_Total'] - df['Rape\_Cases\_Reported']*

*unreported\_victims\_by\_state = df.groupby('Area\_Name').sum()*

*unreported\_victims\_by\_state.drop('Year', axis = 1, inplace = True)*

*plt.subplots(figsize = (12, 4))*

*ct = unreported\_victims\_by\_state[unreported\_victims\_by\_state['Unreported\_Cases']*

*> 0]['Unreported\_Cases'].sort\_values(ascending = False)*

*ax = ct.plot.bar()*

*ax.set\_xlabel('Area Name')*

*ax.set\_ylabel('Total Number of Unreported Rape Victims')*

*ax.set\_title('Statewise total Unreported Rape Victims throughout 2001 to 2010')*

*plt.show()*

*rape\_victims\_by\_state = df.groupby('Area\_Name').sum()*

*rape\_victims\_by\_state.drop('Year', axis = 1, inplace = True)*

*rape\_victims\_by\_state.sort\_values(by = 'Rape\_Cases\_Reported', ascending = False).head()*

*plt.subplots(figsize = (15, 4))*

*ct = rape\_victims\_by\_state['Rape\_Cases\_Reported'].sort\_values(ascending = False)*

*ax = ct.plot.bar()*

*ax.set\_xlabel('Area Name')*

*ax.set\_ylabel('Total Number of Reported Rape Victims')*

*ax.set\_title('Statewise total Reported Rape Victims throught the Years 2001 to 2010')*

*plt.show()*

*total\_rape = df[df['Subgroup'] == 'Total Rape Victims']*

*total\_rape.Year = pd.to\_datetime(total\_rape['Year'], format ='%Y').dt.strftime('%Y')*

*total\_rape.loc[:,'Total\_Rape\_per\_Year'] = total\_rape.groupby('Year')['Victims\_of\_Rape\_Total'].transform('sum')*

*plot\_total\_rape = total\_rape.drop\_duplicates('Year', keep = 'first', inplace = False)*

*plot\_total\_rape.sample()*

*plot\_total\_rape = total\_rape.drop\_duplicates('Year', keep = 'first', inplace = False)*

*total\_rape.Year = pd.to\_datetime(total\_rape['Year'], format ='%Y').dt.strftime('%Y')*

*total\_rape.loc[:,'Total\_Rape\_per\_Year'] = total\_rape.groupby('Year')['Victims\_of\_Rape\_Total'].transform('sum')*

*plot\_total\_rape = total\_rape.drop\_duplicates('Year', keep = 'first', inplace = False)*

*plt.figure(figsize=(10,7))*

*x = plot\_total\_rape['Year']*

*y = plot\_total\_rape['Total\_Rape\_per\_Year']*

*plt.plot(x,y)*

*plt.title('Number of rapes per year in all of India for the decade (2001 - 2010)',color = 'blue',fontsize=14)*

*plt.xlabel('Year')*

*plt.ylabel('Total\_Rape\_per\_Year')*

*plt.grid(True)*

*plt.show()*

*total\_rape\_minor = plot\_total\_rape['Victims\_Upto\_10\_Yrs'].sum()*

*+ plot\_total\_rape['Victims\_Between\_10-14\_Yrs'].sum()*

*+plot\_total\_rape['Victims\_Between\_14-18\_Yrs'].sum()*

*total\_rape\_adults = plot\_total\_rape['Victims\_Between\_18-30\_Yrs'].sum()*

*+ plot\_total\_rape['Victims\_Between\_30-50\_Yrs'].sum()*

*+plot\_total\_rape['Victims\_Above\_50\_Yrs'].sum()*

*data = [total\_rape\_minor,total\_rape\_adults ]*

*index =['Adults Victims', 'Minor Victims']*

*data\_df = pd.DataFrame(data, columns = ['age\_level'], index = index)*

*plt.figure(figsize=(4,3), edgecolor='black')*

*ax = data\_df.plot.bar(rot=0)*

*plt.title("Minor vs Adults - Victims",color = 'royalblue',fontsize=14)*

*data2 = total\_rape.pivot("Area\_Name","Year","Victims\_of\_Rape\_Total" )*

*plt.figure(figsize=(10, 8))*

*plt.yticks(rotation=1)*

*ax = sns.heatmap(data2,cmap="YlGnBu", linewidths=.5)*

*plt.title("State wise Count Distribution of Crime Rape",color = 'blue',fontsize=14)*

*data3 = df.groupby(['Area\_Name'])['Victims\_of\_Rape\_Total'].sum().reset\_index()*

*import plotly.express as px*

*fig = px.pie(data3, values = 'Victims\_of\_Rape\_Total',*

*names = 'Area\_Name',*

*title = 'Rapes state wise',*

*color\_discrete\_sequence=px.colors.sequential.Viridis*

*)*

*fig.show()*

*dfa = df.groupby(['Year'])['Victims\_Upto\_10\_Yrs','Victims\_Between\_10-14\_Yrs',*

*'Victims\_Between\_14-18\_Yrs','Victims\_Between\_18-30\_Yrs',*

*'Victims\_Between\_30-50\_Yrs','Victims\_Above\_50\_Yrs',].sum().reset\_index()*

*dfa = dfa.melt('Year', var\_name='AgeGroup', value\_name='vals')*

*plt.style.use("fivethirtyeight")*

*plt.figure(figsize = (14,10))*

*ax = sns.barplot(x = 'Year' , y = 'vals',hue = 'AgeGroup' ,data = dfa ,palette= 'bright')*

*plt.title('Age Distribution of Victims per Year',size = 20)*

*ax.get\_legend().set\_bbox\_to\_anchor((1, 1))*

*ax.set\_ylabel('')*

*sns.heatmap(df.corr(), annot=True, annot\_kws={'size': 11})*

**CONCLUSION:**

I Explored rape crime data set and performed an Exploratory Data Analysis (EDA) for given features. Moreover, we can compare each and every state’s data with each other and analyse the same thing .

**REFERENCE:**

* *https://www.kaggle.com/rajanand/crime-in-india*
* [*https://www.kaggle.com/udayanguha/eda-in-crime-of-rape-in-india/*](https://www.kaggle.com/udayanguha/eda-in-crime-of-rape-in-india/)
* *https://www.kaggle.com/aurghyadip/eda-of-rape-crimes-in-india?select=20\_Victims\_of\_rape.csv*