Analysis of Crime Against Women in India

Tanaya Kavathekar and Madhuri Yadav

Data Science Department

Columbian College of Arts and Science

George Washington University

Washington, DC, 20052

Table of Contents

Purpose 3

Summary 3

Introduction 3

Smart Question **Error! Bookmark not defined.**

Method 5

- **Data** 5
- **Data Cleaning** 6
- **Exploratory Data Analysis** 6
- **Variation across crime types** 7
- **Variation across states** 9
- **Variation across age group and gender** 10
- **Variation across population** 11
- **Variation across illiteracy rate** 11

Results 13

- **Continuous Factors** 13
- **Categorical Factors** 15

Conclusion 16

Limitations 18

Recommendation 19

Future scope 19

References 19

Purpose

The purpose of this report is to investigate statistics and analysis of the crime that happened against women in India. The analysis identifies and proves numerous factors affecting the number of increasing crimes.

Summary

In the era of big data and emerging machine learning algorithms, there is a growing interest in studying crime records data to take proactive instead of reactive measures. The goal of a crime department is not just to reduce the number of crimes but to avoid crimes. This is because crime is an integral part of society. It affects the overall growth of the economy and even the safety of the country. And hence, there is a need for leveraging big data and applying machine learning algorithms to gain an in-depth understanding of what factors play an influential role in the increase in crime rate. This project provides a thorough study of the crimes against women in the year 2010. There are numerous factors that may affect crime but as per the data availability, this paper explores the region, population, age, gender, literacy, and employment rate as the factors. In order to validate a relationship between the factors, hypothesis testing techniques such as ANOVA and Pearson Correlation are performed on the data. The identified factors can further be used for predicting the number of crimes and also to devise optimal strategies to reduce crimes.

Introduction

Crime affects everyone - rich and poor, young and old, men and women. It is related to the overall growth and socio-economic development. Crime takes place in myriad form namely crimes against safety, property, body, cybercrime, economic crimes. According to the report published by Mukherjee (2019, p. 1) states that along with the other crimes in general, crimes against women have been increasing over the decades in India.

Crime against women creates a huge social obstruction in many developing countries. It acts as a hindrance in the participation of women in society and the labor economy. The relationship between the increase in crime and socio-economic factors is expected to be recurrent. According to Siwach (2017) in crime against women in India, shows that increasing participation of women in the workforce and politics is making them more exposed to crime. Moreover, attempts to control and intimidate women associated with decision-making processes are also leading to violence (Ganguli 1990). As crimes against women are increasing each passing day, it is important for women to live freely, safely and participate equally in all activities (be it political, social, or economic). However, no study examines the relationship between this increasing number of crimes and various influencing factors. Hence this study intends to find the factors that lead to the crime against women. In the scope of this paper, we are analyzing the factors such as the age of the criminal, region, population, illiteracy rate, and gender. We attempt to find a relationship between the aforementioned variables and crime against women.

Smart Question

What is the relationship between variables such as region, population, age, gender, illiteracy rates and an increasing number of crimes against women in India?

Method

Data

In this study, we intend to analyze variables such as the age of the criminal, region, population, illiteracy rate, and gender. The scope of the study is crimes against women in India. National Crime Records Bureau has put together data on different types of crimes. This data is being published annually since 1953. NCRB has also started publishing data on punishable homicide and rape. Indian Penal Code (IPC) has issued the following crimes against women: (1) rape (Sec 376 IPC); (2) kidnapping and abduction for different purposes (Sec 363-373 IPC); (3) homicide for dowry, dowry deaths or attempts to commit such crimes (Sec 302/304B IPC); (4) torture, both mental and physical (Sec 498-A IPC); (5) molestation (Sec 354 IPC); (6) sexual harassment (Sec 509 IPC) (referred to as 'eve-teasing' in the past);3 and (7) importation of girls (up to 21 years of age, Sec 366-B IPC) (Rai 2019).

This large set of data has been compiled and stored on a platform called Kaggle. Kaggle is an open-source community for data scientists. Kaggle allows a user to find and publish data sets. This data is usually collated and preprocessed to use. Here is the link to the Kaggle dataset being used for our purpose.

The Dataset contains 10 years of data starting from 2001. The level of the data is state-year-crime types. In this data, we are focusing on variables such as age, gender, reported crimes, regions, and types of crimes. The shortcoming of this dataset is that it has not been updated in recent years. There are a lot of missing data when it comes to the crimes reported by the local police department. Apart from this, we have also used census data. This data is also available on Kaggle, which is originally owned by the Registrar General and Census Commissioner of India under the Ministry of Home Affairs, Government of India. The data includes demographic information and housing data at a district level. In this analysis, we are using columns indicating the illiteracy rate in India. The data is aggregated at a state level and later merged with crime data.

Data Cleaning

We looked for missing values and outliers in the data set. This data is already processed and published on Kaggle. Hence, we did not find any major inaccuracies and inconsistencies in the data.

Exploratory Data Analysis

We begin with exploratory data analysis to find key patterns and insights. Below are the trends we observed in the data.

Figure 1. represents the variation of total crime cases reported against women across ten years (2001-2010). From the graph, it is observed that the crime rates overall are increasing, which is a critical situation because the variation in rate has to be the other way around.

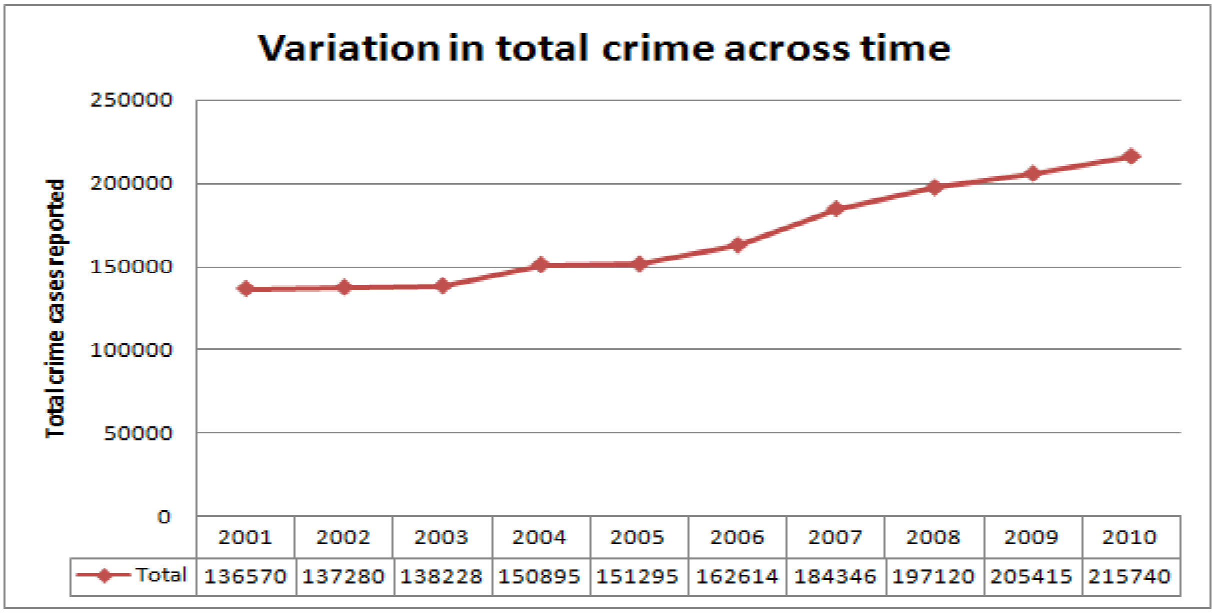

Figure 1. Crime Cases Reported Over 10 Years (2001-2010)

Figure 1. further indicates that the crime rate does not change from 2001 to 2003, and there is a plateau observed in 3 consecutive years. There was a 13% increase from 151295 (in 2005) to 162614 (in 2006). This has been the maximum increase for that decade. Below is a closer look at each of the factors which might be affecting the crime rate.

1. **variation across crime types.**

Below is the pie-chart showing the variation of the crime rate across different types of crime reported against women.

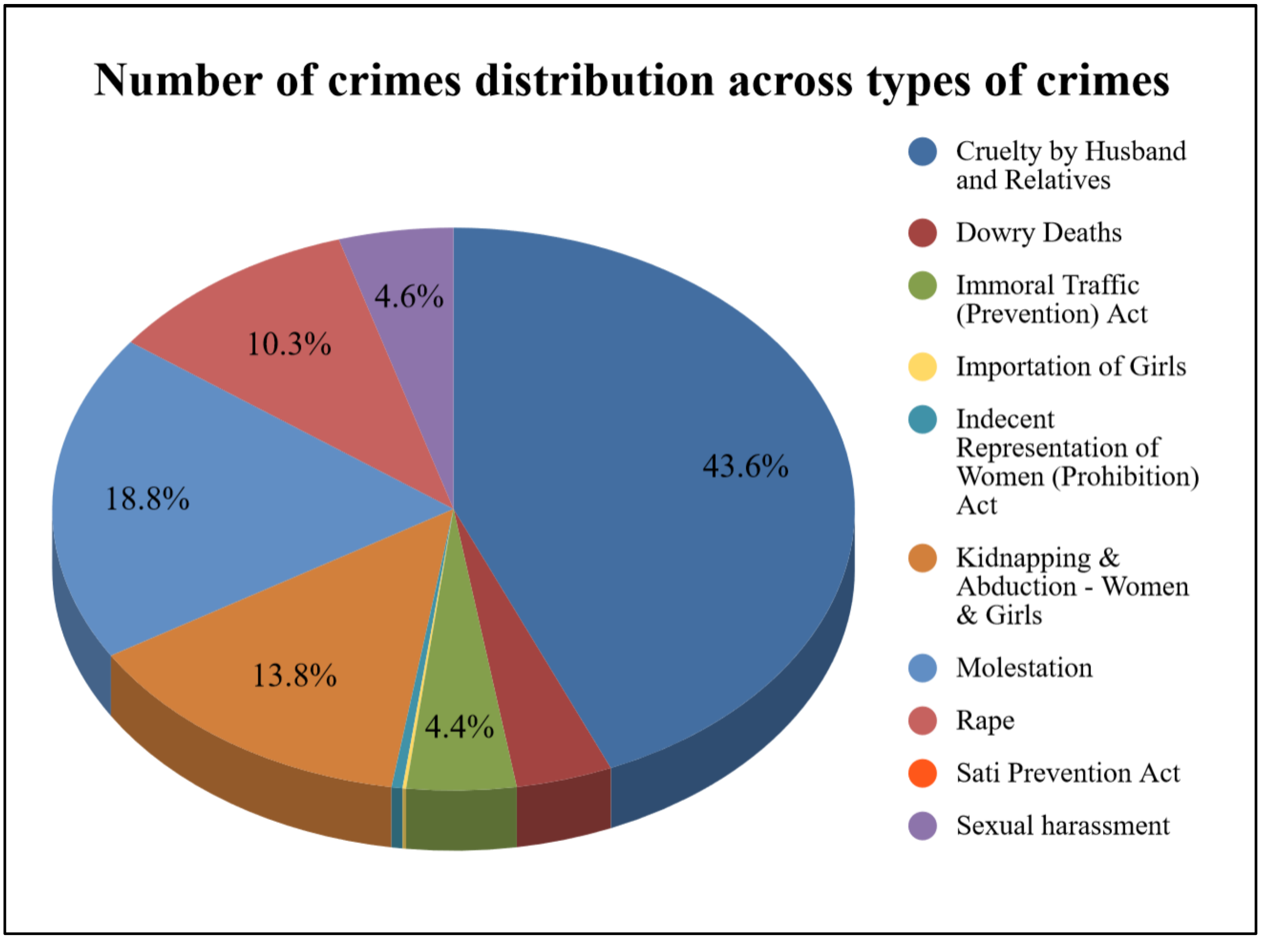

Figure 2. Crime Rate Across Different Types of Crime Reported

Figure 2. shows the distribution of the number of crimes across different types of crime. Cruelty by husband and relatives contributes the highest, which is 44%. ‘Sati’ is the ancient Hindu tradition, wherein a widow would burn herself to death on her husband’s pyre. Sati Prevention Act was enacted by_the Parliament of India in 1988. Since the Sati practice is 0% it cannot be seen in the graph. Out of 215740 cases reported against women, only 1 was reported for, Sati practice. The practice seems to be decreasing in the modern period. The importation of girls is observed to be 0% however, immoral trafficking is 4.4%, so more attention should be given to these areas.

These implications are as per the descriptive analysis, to get statistically significant results, hypothesis testing is performed.

2. variation across states.

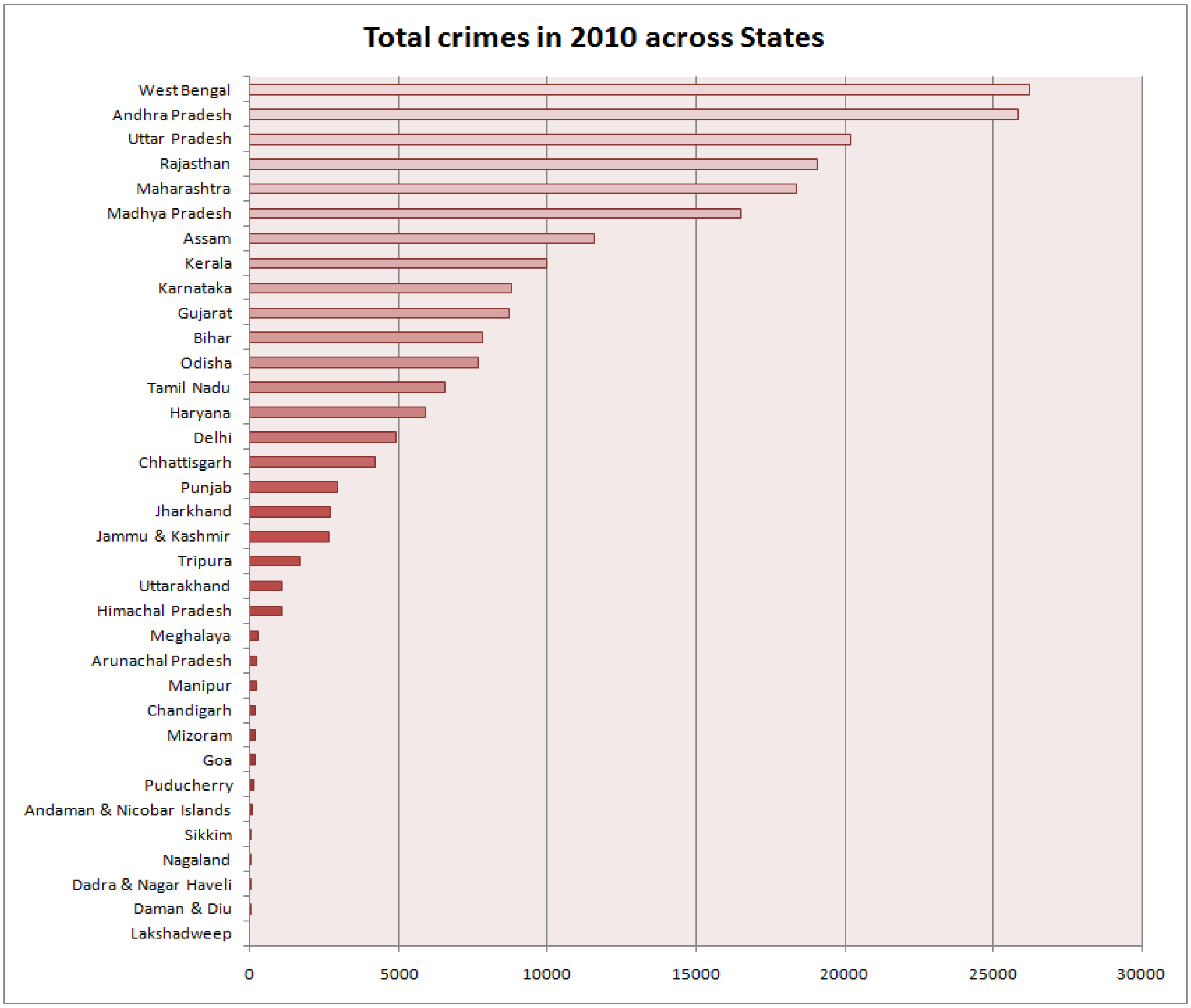

Figure 3. Crime Rate Across Different States and Union Territories

From Figure 3., it can be observed that the crime rate against women in the year 2010 was highest in West Bengal (26207) and second highest in Andhra Pradesh (25832). These states were the most populous state as well. This can be attributed to population size and illiteracy rate, which is further verified. Surprisingly, Lakshadweep has zero number of crimes, followed by Daman & Diu and Dadra & Nagar Haveli with the number of crimes 18 and 36 respectively.

3. variation across age groups and gender.

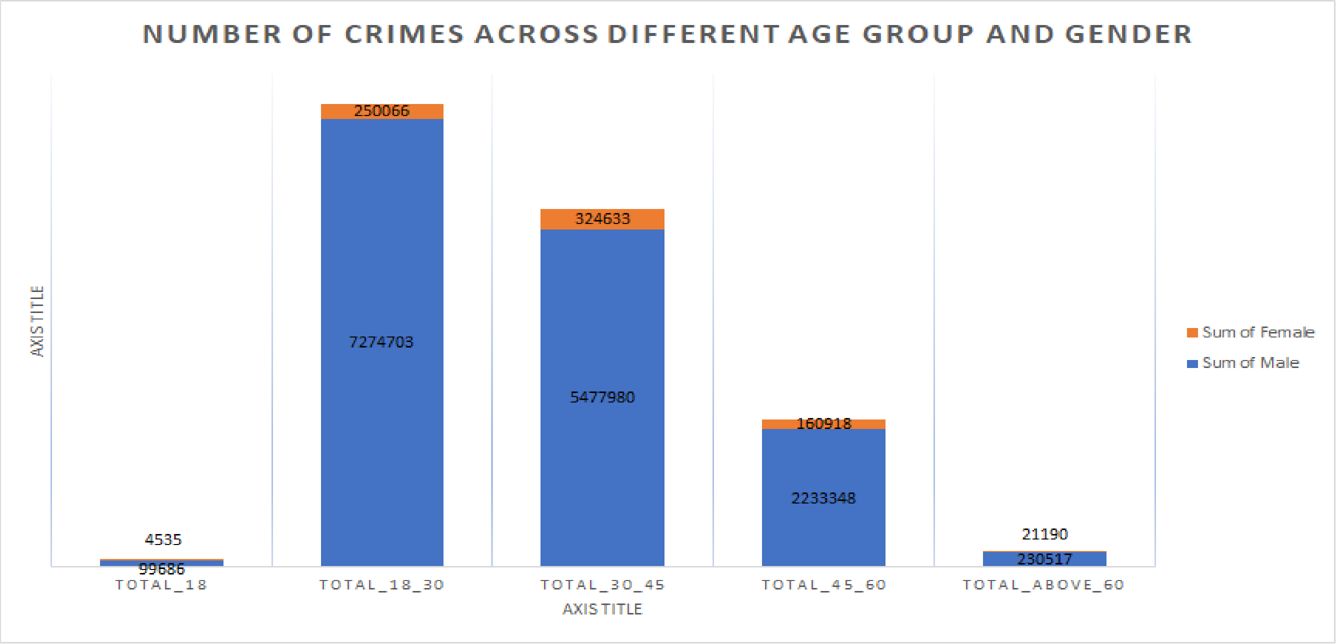

Figure 4. Crime Rate Across Gender and Various Age Groups

Figure 4. shows the number of crimes across various age groups and gender. The crime rate is observed to be less in children (0-18 years) and senior citizens (60 and above years). The number of crimes is observed to be maximum in young people between the ages of 18 to 30 and gradually decreasing from middle age (30 to 34 and 45 to 60) to old age (above 60). It is implied from the graph that age is a significant factor for crime rates. It is also implied that men are more likely to be involved in criminal activities against women.

4. variation across population.

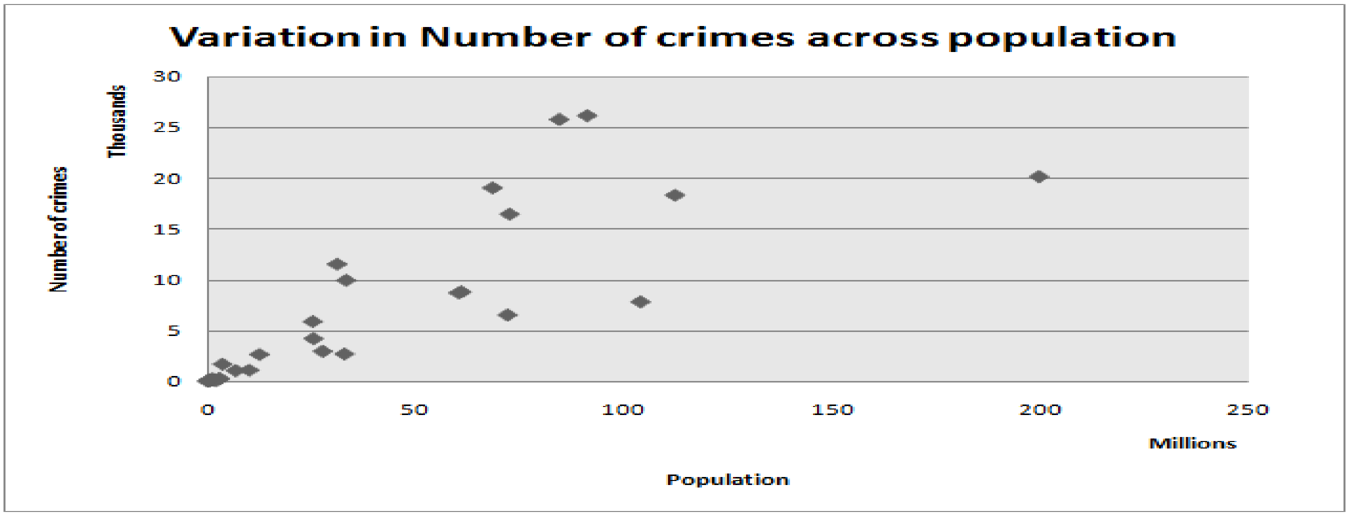

Figure 5. Crime Rate Across Population Rate.

Figure 5. is the scatter plot of the population vs crime rate. From the graph, it can be observed that an increase in population also increases the crime rate. For 0 population crime rate observed is 0 and as population increases to around 100M crime rate also increases to around 27,000. Further for populations higher than 100M crime rate seems to be constant at around 20,000 crimes. As a part of the descriptive analysis, we observe the linear relation between the two variables up to 100M.

5. variation across illiteracy rate.

As per the Census, a person can be considered as literate if aged above seven and knows both read and write in any language. In the year 1991, it was decided as a child below age group 6 will be treated as illiterate and only the age group above 6 will be used to determine the illiteracy/literacy rate. The average illiteracy rate in India for the year 2010 across all the states is

25%. Rajasthan has a high illiteracy rate of 43% and Lakshadweep has the lowest 9% illiteracy rate. From Figure 3. It can be observed that the crime rates are observed more in Rajasthan and least in Lakshadweep. However, even though crime observed illiteracy rate is less in Kerala(9%) observed number of crimes is more(100,000) as per Figure 3.

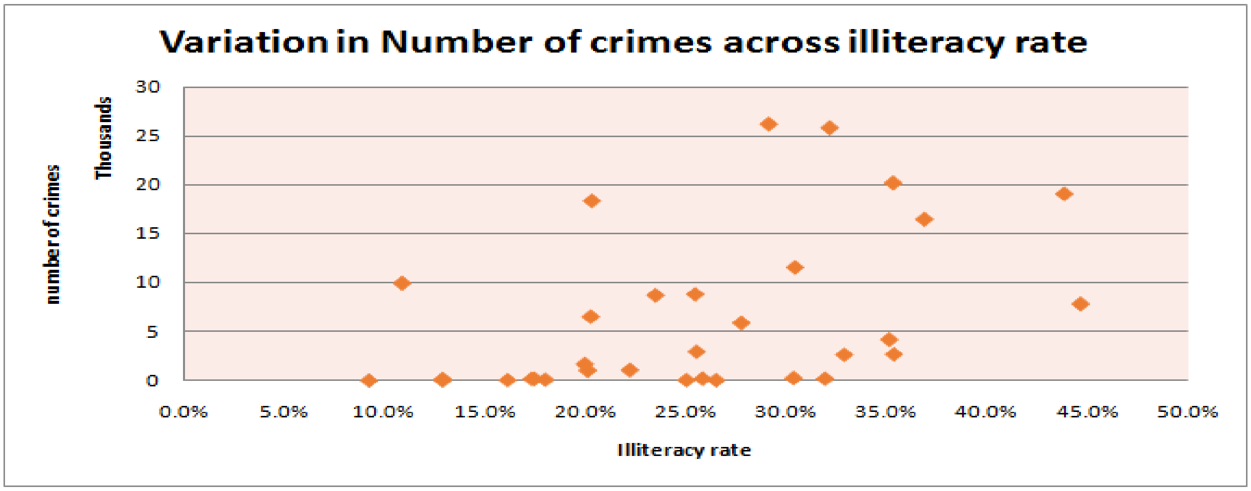

Figure 6. Crime Rate vs Illiteracy Rate Scatter Plot

Figure 6. is a scatter plot of crime rate vs illiteracy rate. From the graph, it can be observed that irrespective of illiteracy rate from 10% to 50%, the number of crime vary between 0 to 28,000. The graph does not imply any trend between the crime rate and the illiteracy rate. This can be further discussed with exploratory data analysis.

Results

As the purpose of this report is to investigate and analyze the factors such as region, type, age, gender, population and illiteracy rate, etc that may affect the crime rate against women in India**.** The following are the results of the EDA on these factors.

1. Continuous Factors

The socio-economic factors population, illiteracy rate, and employment rate are numeric and continuous variables. Hence the Pearson Correlation test is performed for these variables. According to the Statistics Solutions website:

> Pearson Correlation is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance The Pearson Correlation is a parametric measure. This measure is also known as Pearson's correlation.
>
> (https://www.statisticssolutions.com/pearsons-correlation-coefficient/)

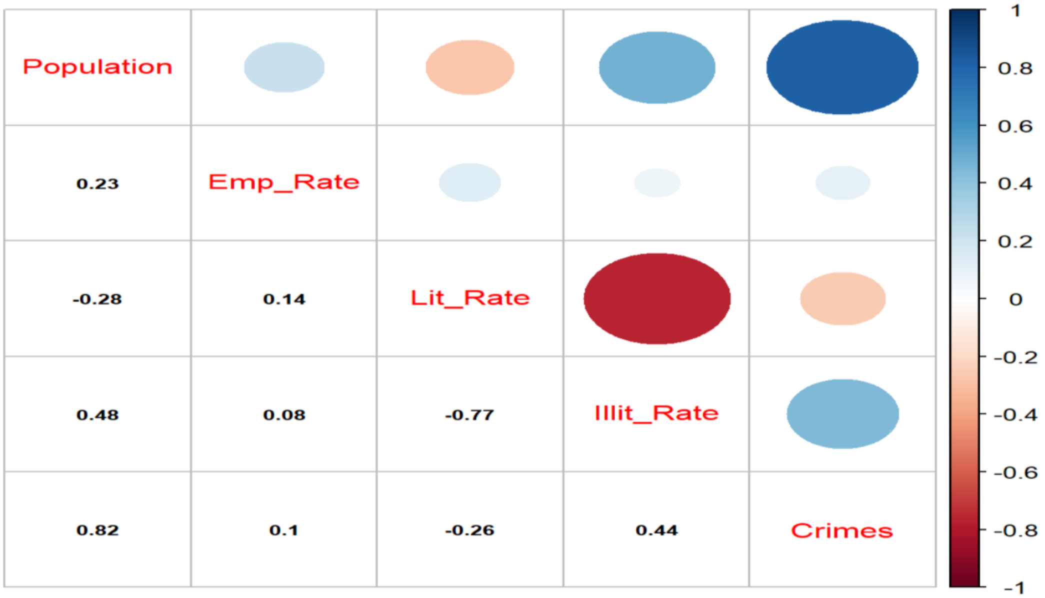

Figure 7. Heatmap Showing Pearson Correlation of Multiple Variables

Figure 7. is the heatmap showing the correlation between the socio-economic variable and the number of crimes.

- The population is directly and strongly correlated with the crimes with the Pearson Correlation Coefficient (ρ) = 0.82. Here the correlation is positive, meaning both variables move in the same direction, as population increases, the number of crimes increase with the population.
- The illiteracy rate is moderately correlated with crimes with Pearson Correlation Coefficient (ρ) = 0.44. Here the correlation is positive, meaning both variables move in the same direction however since the correlation is not strong(44%) implies there is a lower likelihood of there being a relationship with each other.
- Surprisingly the Pearson Correlation Coefficient (ρ) = 0.1 for the employment rate indicating employment rate is weakly related to crime rates. This implies an increase/decrease in employment rate does not affect the crime rate.

2. Categorical Factors

The categorical factors are analyzed by using ANOVA. ANOVA is Analysis Of Variance. This method is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups. ANOVA is performed on types of crimes and the location of the crime.

I. Anova for types of crimes: Anova is performed at a significance level of 5% (α = 0.05).

Null hypothesis: Means between different types of crime are the same.

Alternate hypothesis: Means between different types of crime are not the same.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group_Name	10	$2.93 * 10^{11}$	$2.93 * 10^{10}$	29.73	$2 * 10^{-16}$ ***
Residuals	87	$8.58 * 10^{10}$	$9.86 * 10^{8}$		
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 1. ANOVA results for Type of crime

The one-way ANOVA results of the comparison between the effect of the type of crime and crime rate for various crime types are, there was a significant effect of the crime type of crime at the p<.05 level for the three conditions [F(10,87) = 29.73, p = 0.00001].

II. Anova for regions of crimes: Anova is performed at a significance level of 5% (α = 0.05).

Null hypothesis: Means between different regions of crime are the same.

Alternate hypothesis: Means between different regions of crime are not the same.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Area_Name	31	576.4	18.593	6.152	$2 * 10^{-16}$ ***
Residuals	288	870.4	3.022		
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Table 2. ANOVA results for Region of crime

The one-way ANOVA results of the comparison between the effect of the region or state of crime and crime rate for various regions are, there was a significant effect of the region of crime at the p<.05 level for the three conditions [F(31,288) = 7.668, p = 0.00001].

For both type and region of crime, P-value = 2e^-16, which is between 0 and 0.001. Since the value is less than 0.05 (significance level), the null hypothesis is rejected, meaning the means are not the same for different crime types and regions. Since the means for the crime types and regions are statistically different at a confidence interval of 95%, it can be concluded that the crime rate varies with the type and region of crime and these are statistically significant factors.

Conclusion

Thus, the report successfully concludes that the categorical factors region and type of crime play a significant role in the increase/decrease of crime. Also, socio-economic factors such as age, population, and illiteracy rate highly correlated with the increase in crime rate against women. Surprisingly, the employment rate does not seem to be weakly correlated with the crime rate as per our analysis with the data. Table 3. Gives the summary of our variables and their significance coefficient.

Factors	Test	Values(P/ρ)
No of Crime Vs Types of Crime	Anova	P = 0.00001
No of Crime Vs Location of Crime	Anova	P = 0.00001
No of Crime Vs Population Rate	Pearson Correlation	ρ = 0.82
No of Crime Vs Illiteracy Rate	Pearson Correlation	ρ = 0.44
No of Crime Vs Employment Rate	Pearson Correlation	ρ = 0.1

Table 3. Results of Analysis (p: the probability of obtaining the observed results of a test, ρ: Pearson Correlation), ANOVA is performed at the 0.05 significance level.

- The crime rate differs from the type of crime against women ex. Dowry deaths, Importation of girls, etc. This could be due to the various law enforcements - Dowry deaths have been reduced due to Dowry Prevention Act of 1961, Education and increasing participation of women in the workforce - might be reasons for the increase in the kidnapping. Hence more attention could be given to types of crimes with increasing rates.
- The crime rate differs at various regions ie crime rate is different in different states and it is statistically significant with the region of crime. This may be due to the varying population, literacy and employment rate. As these are observed to be different for each state.
- It is quite obvious that the number of crimes increases with an increase in the population. The key reason behind this could be the employment rate and literacy. Hence more

attention should be given to reduce crimes against women for the states with a bigger population.

- It is observed that the Illiteracy rate is moderately related to crimes against women. This could be because irrespective of the education people still ask for dowry. Also due to education to women, they have increased participation in the workforce and increased exposure to various crimes that happen outside which women in the older era did not face. Hence new laws could be established with a better concentration in these areas.
- The weak correlation between the employment rate and crime rate indicates that the. The employment rate does not contribute to crimes against women.

Limitations

Like every other data set, this data has some limitations that may impact the results of our study.

1. **Unavailability of latest data:** The data is available only from 2001 to 2010, which is a decade-old data. The newer version might provide more reliable factor significance.
2. **Unreported crime:** The data includes only reported crimes to the police department. So the analysis misses out the unreported crimes. More than half the crimes in metropolitan cities go unreported.
3. **More features:** For the factors age and gender, if multiple years of data were available, better significance tests could have been conducted.
4. **Human Errors in Data set:** Due to the manual entry of the data in the police department, the dataset may prone to human errors.

Recommendation

1. **Social awareness campaigns:** As we have identified age and gender affect crime rate, there is a need to drive social awareness campaigns to increase knowledge and empower and lift women's rights.
2. **Strong laws and punishment:** Severe punishment for heinous crimes and strong law should be passed to combat crimes and make the country a safe place for women

Future scope

1. A similar analysis could be performed with larger and latest datasets to get a better and more reliable analysis for various factors.
2. The identified factors can be used to build a model to predict the number of crimes. This will be helpful to devise strategies for avoiding crimes and make the country a safe place for women.

References

Crime in India - 2016. Retrieved from https://data.gov.in/catalog/crime-india-2016?filters%5Bfield_catalog_reference%5D=4223681&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc

EDA in Crime of Rape In India | Kaggle. (2019). Retrieved from https://www.kaggle.com/udayanguha/eda-in-crime-of-rape-in-india/data

Ganguli, A. (1990): 'Objective Study on the Nature of Violence Committed on the Rural Women in West Bengal' in Sushma Sood (ed), Violence Against Women. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4462781/

Mukherjee, C., Rustagi, P., & Krishnaji, N. (2001). Crimes against Women in India. Economic and Political Weekly, 36(43), 1–3. Retrieved from https://www.jstor.org/stable/pdf/4411293.pdf?refreqid=excelsior:bc2f24f35e34baa66060f3a205e59779

National Crime Records Bureau - 2016. Retrieved from http://ncrb.gov.in/StatPublications/CII/CII2016/pdfs/NEWPDFs/Crime%20in%20India%20-%202016%20Complete%20PDF%20291117.pdf

Pearson's Correlation Coefficient - Statistics Solutions. (2019). Retrieved from https://www.statisticssolutions.com/pearsons-correlation-coefficient/

Rai, D. (2019).Offences Against Women. Retrieved from https://blog.ipleaders.in/offences-against-women/

Siwach G (2018), ‘Crimes against women in India: Evaluating the role of a gender representative police force, SSRN Electronic Journal, Retrieved from https://ssrn.com/abstract=316553