Study of Cyber Crimes Using Data Science

Article *in* International Journal of Engineering Technology and Management Sciences · January 2023

DOI: 10.46647/ijetms.2023.v07i05.059

4 authors:

Ruta Vaidya
Haribhai V Desai College of Arts Science and Commerce
4 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Trupti S. Gaikwad
Haribhai V Desai College of Arts, Science and Commerce, Pune
3 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Snehal Jadhav
Savitribai Phule Pune University
6 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Jyoti Malusare
Haribhai V. Desai College of Arts Science and Commerce
1 PUBLICATION **0** CITATIONS
SEE PROFILE

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.059 ISSN: 2581-4621

Study of Cyber Crimes Using Data Science

Ruta Vaidya[1], Trupti Gaikwad[2], Snehal Jadhav[3] , Jyoti Malusare[4]
[1]*Assistant Professor – Haribhai V. Desai College of Arts Science and Commerce, Pune, Maharashtra*
[2]*Assistant Professor – Haribhai V. Desai College of Arts Science and Commerce, Pune, Maharashtra*
[3]*Assistant Professor – PVG's College of Science and Commerce, Pune, Maharashtra*
[4]*Assistant Professor – Haribhai V. Desai College of Arts Science and Commerce, Pune, Maharashtra*

ABSTRACT
In today's technological advancements in the digital world cybercrime has caused unprecedented challenges to individuals and organizations. Cybercrime is an acronym used to describe illicit activity where a computer or network serves as the source, tool of the crime. Cybercrime includes advanced unlawful activities like fraud, the trafficking of child pornography and other intellectual property, identity theft, and privacy violations. A rise in data has been correlated with a rise in incidents of data theft. Worldwide, organizations and individuals are increasingly concerned about hacking and system intrusion using various technologies. Data science is the study of data to extract useful business insights. It is a path of interdisciplinary access for analyzing enormous amounts of data that brings together information and techniques from the domains of artificial intelligence, statistics, mathematics, and computer engineering. In this paper, the data has been collected from Indian Government sites and studied for different states and metropolitan cities. Using different data science techniques along with the statistical methods like correlation, hypothesis testing, some conclusions have been drawn for various motives of cybercrimes.
Keywords—Cybercrimes, Fraud, Data science, Hypothesis, Statistics

1. Introduction

What is Cybercrime?
Cybercrime is the act of committing crimes using computers or other electronic devices. Basically, cybercrimes are unlawful activities which target computer networks and persons with the purpose of financial gain and disordering. There are different types of cybercrimes like cyberstalking, personal revenge, Phishing, Identity theft etc. Cybercrimes have increased as computers become a crucial part of day-to-day life. Cybercriminals employ a variety of attack vectors to carry out their cyberattacks and are always looking for new strategies to accomplish their objectives without being discovered and apprehended. Malware and other forms of software are commonly used by cybercriminals in their operations, but social engineering is frequently a crucial step in the implementation of the majority of cybercrimes. Another key component of several cybercrimes, particularly targeted cyberattacks, is phishing emails.

2. Experimental Methods or Methodology

In this research paper we have analyzed the data recorded by the government site during the years 2019 to 2021 i.e., before pandemic, during pandemic and after pandemic. We studied the data related to the crimes in different states of India along with metropolitan cities of India. Further we tried to compare the cybercrimes in the least developed states and the most developed states. Even we checked if there is a significant difference in persons arrested and persons chargesheeted in metropolitan cities in the year 2020. We tried to examine the correlation between the different motives of cybercrimes.

Analytical Avenue:
The different statistical methods have been used for the analysis.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.059 ISSN: 2581-4621

Chi-Square Test: Basically, Chi-square test is a non-parametric test used for goodness of fit, for independence and for proportions. It is used in determining if there is statistical difference between the observed and the expected frequencies. When the sample sizes are large, it is used in analysis of contingency tables. The null hypothesis is defined as there is no difference in the classes in the population and alternative hypothesis as there is difference in the classes. The observations are classified into mutually exclusive events. Chi- square distribution is applied when the observations are independent. The formula for the test is given by $\chi 2 = \sum (Oi - Ei)2/Ei$

The following steps are involved for conducting Chi-square test. First, we have to define Null and alternative hypotheses. Data is collected in the form of a contingency table. Chi-square statistics are calculated using observed and expected frequencies. Degrees of freedom and P value are calculated using statistical tables or software. By comparing the chi- square statistic and critical value, one can decide whether the null hypothesis is rejected or accepted.

T-Test: A T-test is a statistical hypothesis test used to determine if there is a significant difference between the means of two groups or samples. It is particularly useful when comparing means of small sample sizes or when the population standard deviation is unknown. The t-test is based on the t-distribution, which is similar to the normal distribution but with heavier tails, making it appropriate for small sample sizes. The general process for conducting a t-test involves the following steps: Formulation of Null and alternative hypothesis. Examine data for two groups. Compute t statistic. Determine degrees of freedom. Using the table or software, find critical value. By Comparing t statistic and critical value, the decision can be taken to accept or reject hypothesis.

Correlation: Correlation gives us an idea if the two variables are correlated or not. Correlation may be positive or negative.

3. Results and Discussion

State Wise Cybercrimes from 2019-2021:

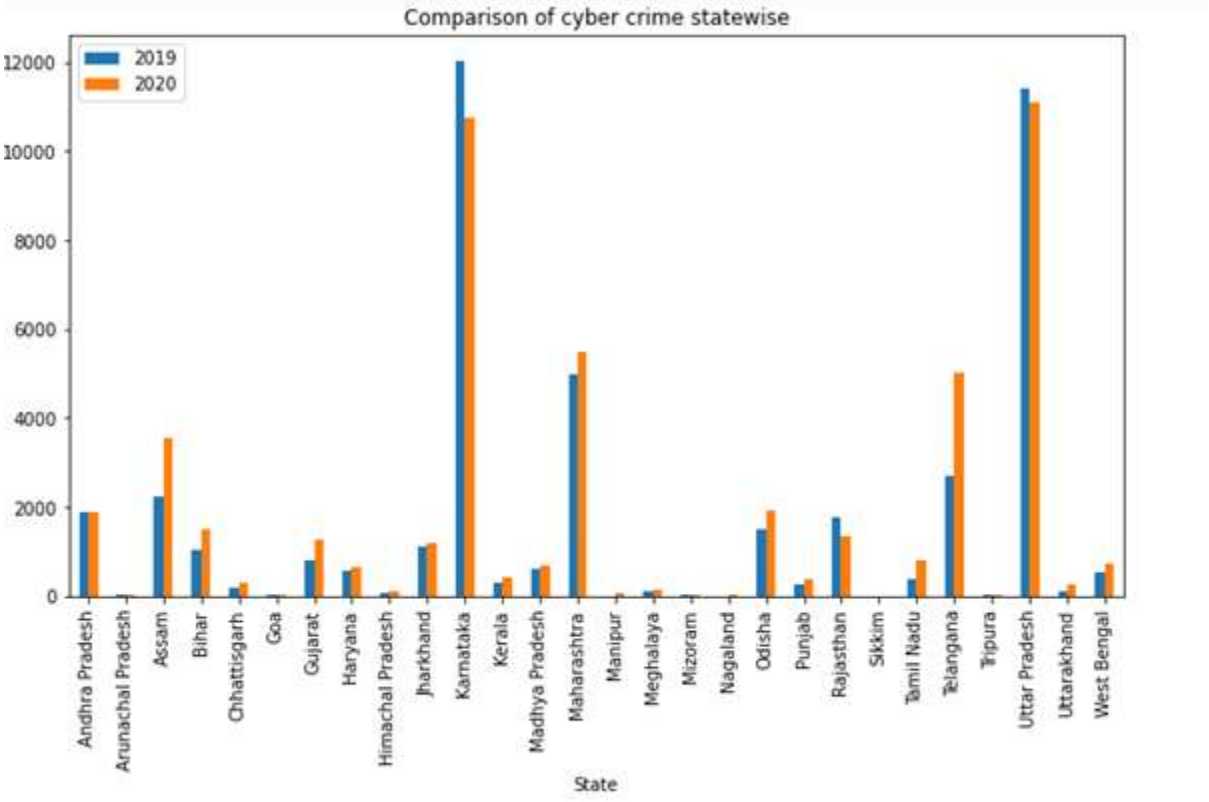


Fig. 1.

Fig 1 displays the chart of cybercrimes for the year 2019-2020. The graph clearly shows that there was a spike in cybercrimes in 2020 in many states, possibly due to a pandemic. The data shows that there is a decrease in offline crimes but rise in online crimes; one of the reasons for this is lockdown.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.059 ISSN: 2581-4621

The above bar graph also indicates for the year 2019 Karnataka followed by Uttar Pradesh, Maharashtra states has maximum cybercrimes while Nagaland and Sikkim have minimum number of cybercrimes. The reasons for this are the states like Karnataka, Maharashtra and Uttar Pradesh have a significant population with access to the internet, making them potential targets for cybercriminals. Another reason is individuals and organizations not taking adequate measures to protect their online activities and data. Nagaland and Sikkim have relatively low population densities compared to many other Indian states. Lower population densities can result in fewer potential targets for cybercriminals, reducing the overall incidence of cybercrime. The extent of internet penetration can affect the prevalence of cybercrimes. States with lower internet penetration may have fewer opportunities for cybercriminals to exploit.

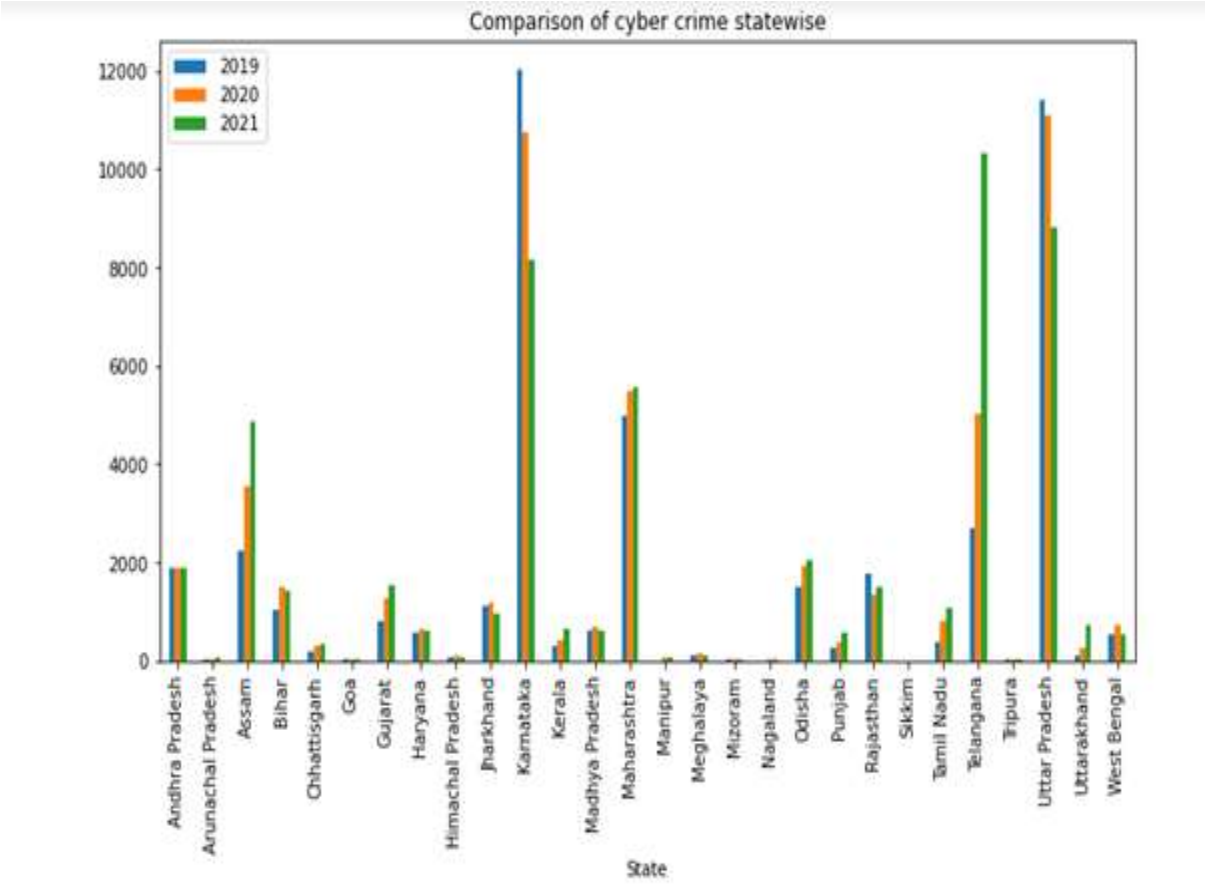


Fig. 2.

From Fig 2 we can check that the cybercrimes have jumped up in states like Telangana, Assam and Uttarakhand. Fraud, Personal Revenge and Bullying cases against women and children are the major motives for the rise in cybercrimes, as specified by NCRB data .

The 10 least developed states according to MDI (Major depression inventory score) are Odisha, Bihar, Madhya Pradesh, Chhattisgarh, Jharkhand, Arunachal Pradesh, Assam, Meghalaya, Uttar Pradesh and Rajasthan. On the other hand, seven most developed states are Goa, Kerala, Tamil Nadu, Punjab, Maharashtra, Uttarakhand and Haryana.

Following hypotheses are tested using Chi square Test:

1) Comparison between the cybercrimes in the most developed states and the least developed states in the years 2018-2020.

#H0: There is no significant difference in cybercrimes in the least developed states and most developed states

#H1: There is significant difference in cybercrimes in the least developed states and most developed states

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.059 ISSN: 2581-4621

OUTPUT

```
critical: 5.991464547107979
Dependent (reject H0) and so we can say that There is significant difference in cyber crime in the least developed states and m
ost deveeloped states
chi-square: 77.99355175398487
degree of freedom: 2
expected value table: [[12864.56922264 19458.43211833 21898.99865903]
 [ 4651.43077736  7035.56788167  7918.00134097]]
p value is 1.1585517154543455e-17
Dependent (reject H0) and so we can say that There is significant difference in cyber crime in the least developed states and m
ost deveeloped states
```

Here we added data of the most developed and the least developed states for years 2018-2020. The conclusion that "There is significant difference in cybercrimes in the least developed states and most developed states" has been drawn by using chi square test.

It is also very important that the person who has committed a cybercrime is punished. The persons arrested should be chargesheeted.

To check independence of two attributes chi square test has been used. For that purpose gender wise data has been collected and the relation between persons arrested and persons chargesheeted is studied.

#Ho: There is no significant difference in persons arrested and persons chargesheeted in metropolitan cities in the year 2020

#H1: There is significant difference in persons arrested and persons charge sheeted in metropolitan cities in the year 2020

OUTPUT

```
critical: 3.841458820694124
Independent (H0 holds true) and so  There is no significant difference in persons arrested and persons charge sheeted in metropolitan cities in the year 2020
degree of freedom: 1
p value is 0.2776683706954771
Independent (H0 holds true)and so  There is no significant difference in persons arrested and persons charge sheeted in metropolitan cities in the year 2020
```

It has been concluded that there is no significant difference in persons arrested and persons chargesheeted in metropolitan cities in the year 2020.

Pre pandemic and post pandemic changes between the cyber crimes are analyzed using T distribution.

#Ho: There is no significant difference in average cybercrimes during pandemic (2019) and after pandemic (2020)

#H1: There is significant difference in average cybercrimes during pandemic (2019) and after pandemic (2020)

OUTPUT

```
Test statistic is -0.011369
p-value for two tailed test is 0.990992
Conclusion n Since p-value(=0.990992) > alpha(=0.05) We accept null hypothesis  H0.There is no significant difference in averag
e cybercrimes during pandemic (2019) and after pandemic (2020)
```

This heat map shows that the diagonal values have perfect correlation and they can be seen in dark green color. The larger numbers and darker color indicate the higher correlation between the two variables.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.059 ISSN: 2581-4621

Output Correlation

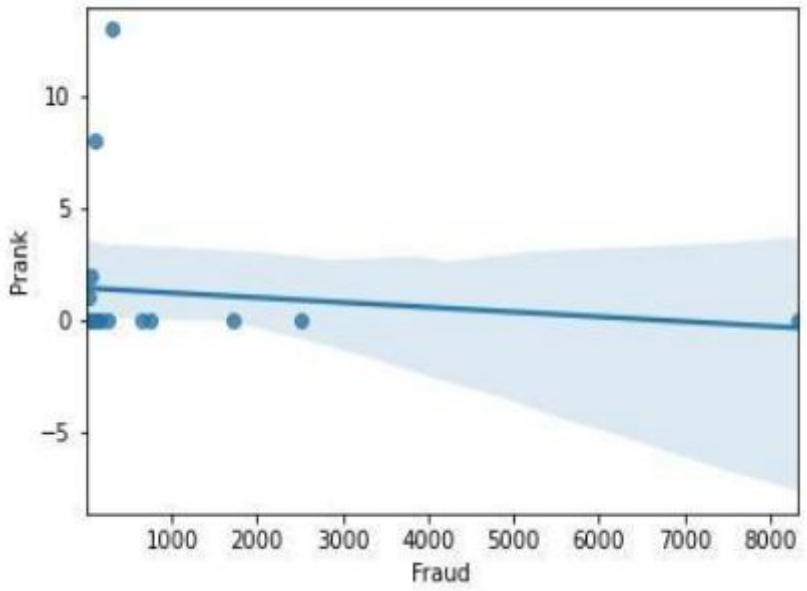
Following output table displays the correlation between the different attributes. If the correlation value is greater than 0.8 then there is a higher degree of correlation between the variables. While the value between 0.3 and 0.8 indicates moderate correlation. If the correlation value is less than 0.3 then there is negligible correlation between the two variables.

OUTPUT

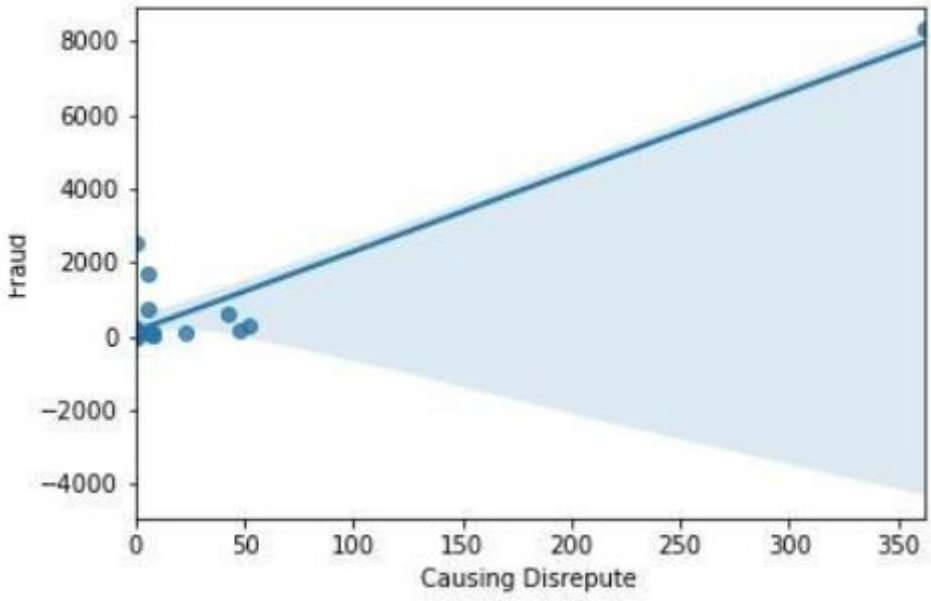
```
Causing Disrepute    Fraud                0.914747
                     Personal Revenge     0.780263
Personal Revenge     Fraud                0.737460
Prank                Anger                0.566576
Sexual Exploitation  Fraud                0.394998
                     Causing Disrepute    0.249774
                     Personal Revenge     0.154470
Extortion            Sexual Exploitation  0.103355
Sexual Exploitation  Anger                0.030107
Causing Disrepute    Extortion            0.028788
Prank                Causing Disrepute    0.004070
Fraud                Extortion           -0.013545
Prank                Sexual Exploitation -0.047801
                     Personal Revenge    -0.053899
Causing Disrepute    Anger               -0.073047
Anger                Personal Revenge    -0.084220
Extortion            Anger               -0.094457
                     Personal Revenge    -0.097394
                     Prank               -0.099941
Fraud                Anger               -0.114613
                     Prank               -0.121751
Personal Revenge     Personal Revenge          NaN
dtype: float64
```

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.059 ISSN: 2581-4621

It can be clearly seen that there is negative correlation between the Fraud and Prank with value -0.121751
OUTPUT



It can be observed from below graph that the Causing Disrepute and Fraud have high degree of positive correlation with value 0.914747
OUTPUT



CONCLUSION

This research work elaborates analysis of cybercrimes in states during 2019 to 2021 using statistical methods and data science tools. A state-wise study has been performed to check whether the pandemic period affected the rise of cybercrimes. It has been found that during a pandemic the internet is a crucial part of our day to day life. Graphical representation in the paper designates that

DOI:10.46647/ijetms.2023.v07i05.059 ISSN: 2581-4621

after the pandemic there is an increase in cyber crimes. Frequently observed cyber crimes are phishing, online scam, fake websites to download covid certificates and other documents, Ecommerce Fraud and Ransomware attacks etc. To protect from cybercrimes a multifaceted approach should involve collaboration between governments, law enforcement agencies, cybersecurity experts, and the private sector. Legislation and regulations must keep pace with technological advancements to deter cybercriminals and hold them accountable for their actions. Moreover, individuals and organizations should prioritize cybersecurity practices, such as strong password management, regular software updates, and employee training to reduce their vulnerability to cybercrimes. This study can be useful to cyber experts to find trends in cyber crime for the given years.

References

1. www.ncrb.gov.in
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8579416/#Sec4title
3. https://en.wikipedia.org/wiki/Chi-squared_test
4. https://en.wikipedia.org/wiki/Information_technology_in_India
5. https://ncrb.gov.in/sites/default/files/CII-2021/CII_2021Volume%202.pdf
6. https://timesofindia.indiatimes.com/city/hyderabad/telangana-revenge-and-fraud-main-motives-in-cyber-crimes/articleshow/93865258.cms
7. https://ncrb.gov.in/sites/default/files/CII%202020%20SNAPSHOTS%20STATES.pdf
8. https://ncrb.gov.in/sites/default/files/CII-2021/CII%202021%20SNAPSHOTS%20STATES.pdf
9. Lundy, Brandon D., Tyler L. Collette, and J. Taylor Downs. "The effectiveness of indigenous conflict management strategies in localized contexts." Cross-Cultural Research 56.1 (2022): 3-28.
10. Rupa Ch , Thippa Reddy Gadekallu , Mustufa Haider Abidi , Abdulrahman Al-Ahmari "Computational System to Classify Cyber Crime Offenses Using Machine Learning"
11. Ravichandran Kamalakannan "CYBER CRIME AND MEDIA AWARENESS IN INDIA (QUANTITATIVE ANALYSIS METHOD)" The Online Journal of Distance Education and e-Learning, October 2018 Volume 6, Issue 4