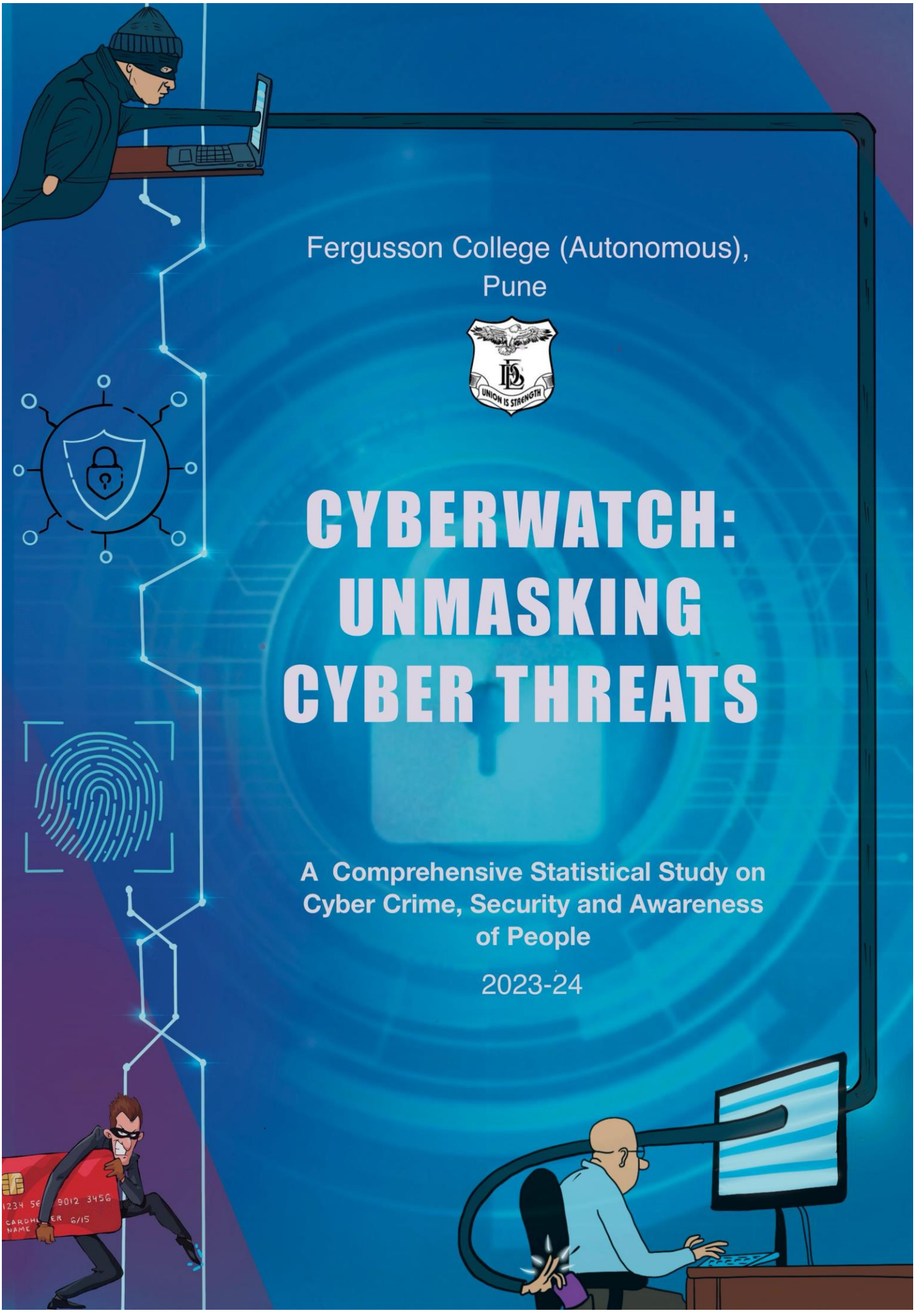


Fergusson College (Autonomous),
Pune

CYBERWATCH: UNMASKING CYBER THREATS

A Comprehensive Statistical Study on Cyber Crime, Security and Awareness of People

2023-24

Deccan Education Society's Fergusson College (Autonomous), Pune

Department of Statistics STS3609:

Statistics Practical III 2023-2024

CERTIFICATE

This is to certify that Mr./Ms._____with roll no._____ ____has satisfactorily completed the project entitled *'Cyber Watch: Unmasking Cyber Threats: A Comprehensive Statistical Study on Cyber Crime, Security & Awareness of people'* towards the partial fulfillment of B.Sc. (Statistics), Semester VI during the academic year 2023-24.

Place: Pune

Date:___/03/2024

Mrs. Deepa Kulkarni
Assistant Professor,
Department of Statistics,
Fergusson College
(Autonomous), Pune

Dr S.S Shende,
Head of Department,
Department of Statistics,
Fergusson College
(Autonomous), Pune

EXTERNAL EXAMINER

GROUP MEMBERS

SERIAL NO.	NAME	ROLL NUMBER
1	ADITI JOSHI	212232
2	DRISHTI SHAH	212263
3	ADITI PRATAPWAR	212234
4	MAYURI WAGHMARE	212245
5	NIKHIL MATKAR	212236
6	ABHIJIT MANE	212226
7	AKASH GIRDE	213157

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who contributed to the completion of this project. Active participation, enormous thoughts, criticisms, suggestions and guidance of people helped us to complete it efficiently.

First and foremost, we extend our deepest appreciation to our Project guide **Prof. Deepa Kulkarni** for their invaluable guidance, support, and encouragement throughout the duration of this project.Her expertise and insightful feedback were instrumental in shaping the direction of our work.

We would also like to thank **Dr. S. S. Shende**, Head of the Department of Statistics and the rest of the faculty for their insightful inputs on our project. In addition, we would also like to thank the Department of Statistics of Fergusson College, Pune for providing the necessary assistance, infrastructure and facilities required.

Of course, we would also like to extend our gratitude to all the respondents of our survey, without whose enthusiasm and participation, our project would not have been possible.

Lastly, we acknowledge the contributions of all the authors, researchers, and practitioners whose work served as a foundation for our project. Thank you all for your unwavering support.

INDEX

INTRODUCTION

Now-a-days, computers and the Internet are becoming essential tools in several aspects of our lives, including academic study, professional work, entertainment, and communication. The internet, and various mobile technologies have altogether transformed our ways of living. Digital environment has taken the place of the physical environment for shopping, entertainment, communication, banking and sharing information. Certainly each development also creates a place to be exploited for criminal purposes.

The ease of information sharing and accessibility makes digital data a prime target for theft, disruption, and misuse. The abuse of photos shared on the internet, online banking fraud, ATM fraud, stalking and harassment via email and SMS, and copyright infringement due to the ease of sharing digital media are just a few examples of how our growing dependence on computers and digital networks has made technology itself an alluring target. These crimes in which the offender uses special knowledge of cyberspace are referred to as **Cyber Crimes**.

In general, **Cyber Crime** may be defined as "Any unlawful act where a computer or communication device or computer network is used to commit or facilitate the commission of crime". All over the world the number of cybercrimes are rising exponentially with time causing huge financial and personal losses. Therefore, computer and network security are a concern not only for traditional security awareness organizations, for example, military, bank, or financial institutions, but also for every individual and government official who uses computers. Besides, nowadays, more and more organizations' valuable assets are stored in the computerized information system; security has become an essential and urgent issue.

Cyber Security refers to the practice of protecting computer systems, networks, devices, and data from unauthorized access, cyberattacks, theft, damage, or disruption. Early detection and reporting of suspicious activity can help mitigate the impact of cyberattacks. Educated users are less likely to fall prey to cyberattacks. Therefore, promoting **Cyber Awareness** is vital for building a safer and more secure digital environment for everyone. Thus through this project we want to highlight the large scale and dangerous impacts of cybercrimes and make people aware about such crimes and also the importance of being secure online.

MOTIVATION

It is estimated that Cybercrimes will cost the world $10.5 trillion annually by 2025! Also, every 39 seconds, there's a Cyberattack somewhere in the world.

- A few years back, NASA's computers were hacked and shut down for 21 days! There were around 1.7 million software downloads during the attack, which cost the space giant around $41,000 in repairs.
- 2018 started with a massive data breach of personal records of 1.1 Billion Indian Aadhaar cardholders. UIDAI revealed that around 210 Indian Government websites had leaked the Aadhaar details of people online.
- Large scale organizations like India's space agency, ISRO deals with 100 cyber hacking attempts daily.

Such incidents highlight the fact that Cyber Crime is a major global issue. If such large organizations are being affected by cybercrimes, then we as individuals are an easy target for cyber criminals.

On an individual level we always hear about incidents like someone's social media account being hacked, financial losses due to clicking on some malicious link, spam calls trying to get someone's personal information and trying to exploit the person, e-commerce frauds, etc. Since such crimes have become predominant in today's world, this made us students think that are people really aware about such crimes? And if they are aware, are they practicing any security measures to deal with them? Hence, we wanted to find out if there are any trends in cyber awareness and security practices among different age groups, professions, streams of study and level of expertise through a survey. These findings might help us in understanding the impact of cybercrimes and help us in deciding what measures need to be taken to be safe in the digital world.

OBJECTIVES

1. To understand the relationship between different cyber indexes and the overall human development and economy of countries.
2. To check the significance of losses due to Cyber Crime in different sectors.
3. To understand the trends and pattern of losses and victims in Cyber Crime.
4. To check if cyber awareness implies more cyber security practices.
5. To find how cyber awareness and cyber security vary with age, gender, educational qualification and level of expertise.
6. To spread awareness amongst the generation about the measures to be taken to stay safe in the digital world.

DEFINITIONS

- **Virus**- A computer virus is a type of malicious software, or malware, that spreads between computers and causes damage to data and software. Computer viruses aim to disrupt systems, cause major operational issues, and result in data loss and leakage.
- **Firewall**- A firewall is a network security device that monitors incoming and outgoing network traffic and decides whether to allow or block specific traffic based on a defined set of security rules.
- **Antivirus**- Antivirus is a kind of software designed to detect and remove viruses and other kinds of malicious software from your computer or laptop. Malicious software - known as malware - is code that can harm your computers and laptops, and the data on them.
- **Authentication**- Authentication is the process of verifying a user or device before allowing access to a system or resources. This ensures only those with authorized credentials gain access to secure systems.
- **Encryption**- Encryption is the process of protecting information or data by using mathematical models to scramble it in such a way that only the parties who have the key to unscramble it can access it.
- **Software update**- A software update (also known as patch) is a set of changes to a software to update, fix or improve it. Changes to the software will usually either fix bugs, fix security vulnerabilities, provide new features or improve performances and usability.
- **Cookie** - A cookie is a piece of data from a website that is stored within a web browser that the website can retrieve at a later time. Cookies are used to tell the server that the users have returned to the particular website
- **Types of cookies:**
  - **First party cookies** - First-party cookies are directly stored by the website (or domain) you visit. These cookies allow website owners to collect analytics data, remember language settings, and perform other useful functions that provide a good user experience. First-party cookies will probably remain a core part of website infrastructure for a long time to come. They work well, improve performance, and pose limited threats to user safety and privacy.

- **Third party cookies**-A third-party cookie is a cookie that's placed on a user's device computer, cell phone or tablet -- by a website from a domain other than the one the user is visiting. Third-party cookies are most frequently used for online advertising. Allowing third-party cookies can pose privacy risks, as they can track your online behavior and collect data across websites. Some users choose to block them for added privacy.

❖ **Types of Cyber Crimes:**

- **Identity theft**-Identity theft, identity piracy or identity infringement occurs when someone uses another's personal identifying information, like their name, identifying number, or credit card number, without their permission, to commit fraud or other crimes.
- **Phishing**- A technique for wanting to acquire sensitive data, such as bank account no.s, through a fraudulent solicitation in email or on a website in which the perpetrator masquerades as a legitimate business or a reliable person. It can occur through embedded links and attachments and takes away sensitive information.
- **Malware**- A malicious type of software with the intent of damaging devices.
- **Ransomware**- a type of malicious software designed to block access to a computer system until a sum of money is paid.
- **Data breach** - A data breach is a security violation, in which sensitive, protected or confidential data is copied, transmitted, viewed, stolen, altered or used by an individual unauthorized to do so.

DATA COLLECTION & METHODOLOGY

In this Project, we mainly focus on spreading the awareness of different types of Cyber Crime and how one can avoid being the victim of Cyber Crime by taking different Cyber Security measures. Hence to serve the objectives of our project we decided to go for both primary and secondary data.

We collected the crime type data through various authentic sources such as Surfshark, CERT, GCI, NCSI, IBM, APWG, AAG etc

Primary data collection is done with the help of google form. We created a google form with 30 questions which are based on Cyber Awareness and Cyber Security. We collected responses from all age groups starting from 11 to 60 .To ensure that there should not be any biased we collected data from people working in different sectors and also we considered the region they belong to urban , semi urban and rural. In total we received 623 responses, complete data is collected through online mode only.

With the help of primary data collected we defined two new variables Cyber Awareness Score and Cyber Security Score. Based on different cyber security practices and cyber awareness among the individuals we calculated these scores. For this study, we considered 10 cyber security measures that an individual should practice regularly to reduce the risk of cyberattacks. Similarly for calculating the Cyber Awareness Score we considered 9 questions. Each question could be answered with any one of the following options:

1. Always
2. Often
3. Sometimes
4. Rarely
5. Never

For each question, we allocated a score out of 5 based on the response appropriately.

Example: Do you use strong, unique passwords for all your online accounts?

Option	Always	Often	Sometimes	Rarely	Never
Score	5	4	3	2	1

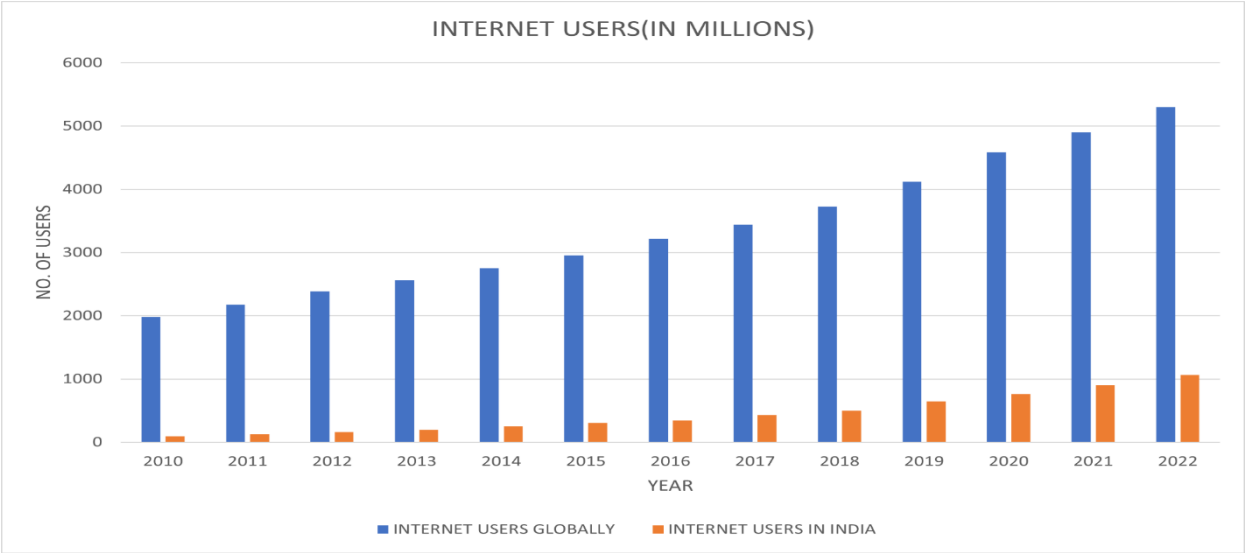
Based on 10 questions we calculated a total cybersecurity score out of 50. Similarly, based on 9 questions we calculated a total cybersecurity score out of 45. Higher cyber security score and higher awareness score signify more secure and aware about cyberattacks.

A Cyber security score greater than 37 is considered as a good score and a cyber awareness score greater than 33 is considered as a good score, which is used in further primary analysis.

EXPLORATORY DATA ANALYSIS

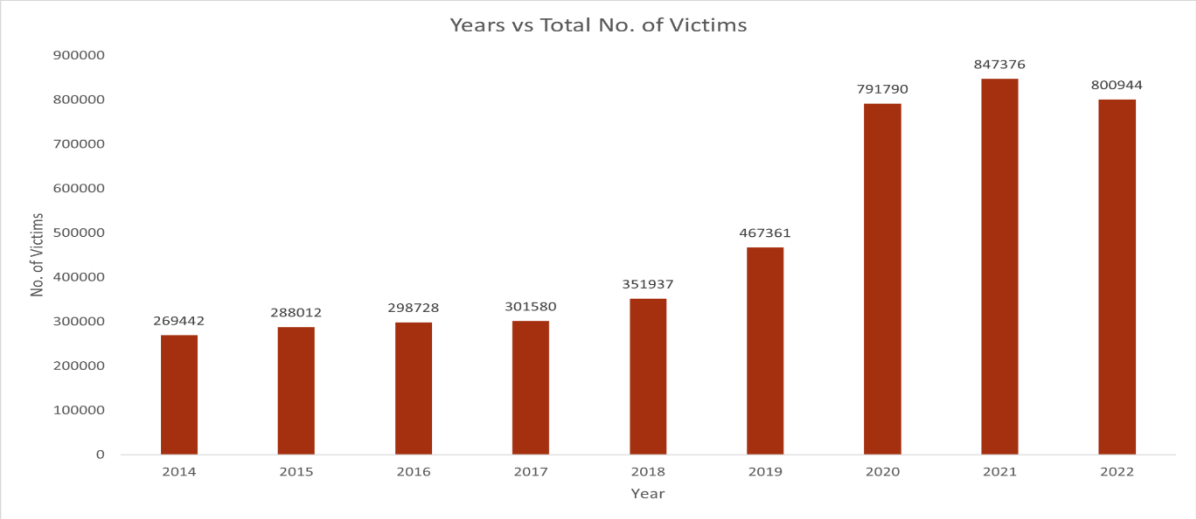
SECONDARY

1. Total Internet Users (Globally & India)



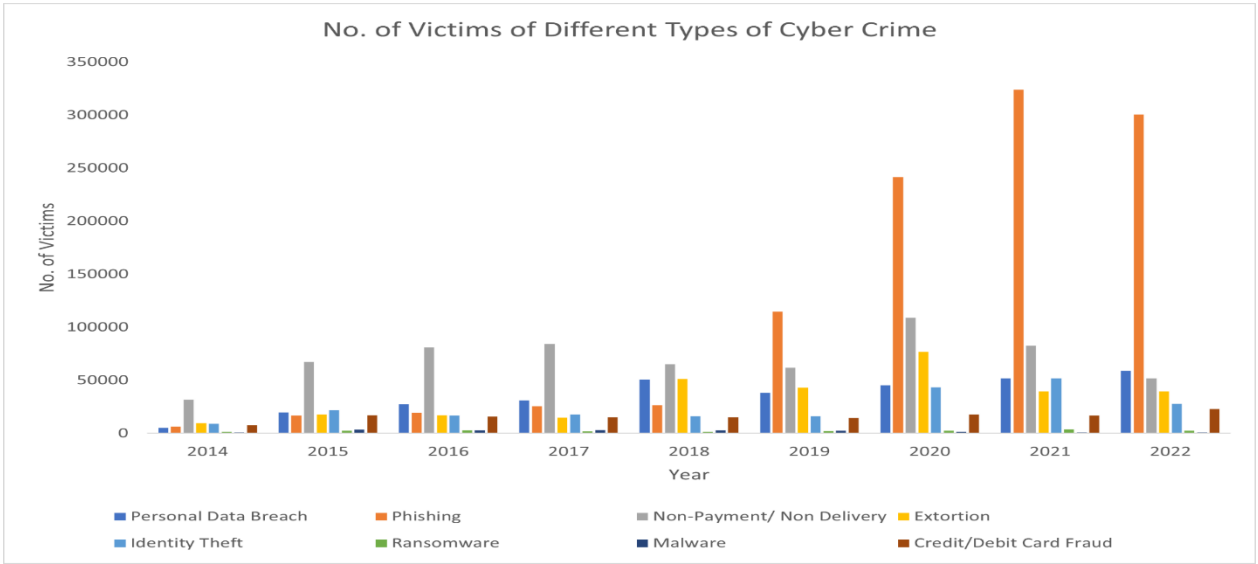
As we can see from the graph the total number of internet users have increased drastically over the years. Looking ahead, Cybersecurity Ventures predicts that there will be more than 7.5 billion Internet users by 2030.

2.Total no. of Cyber Crime victims (India)



The increase in internet users has lead to an increase in Cyber Crime rate over the years as seen in the graph above, especially after 2019 as we were forced to sit back at home due to COVID-19 leading to an increase in internet usage and hence cyber crimes

3. No. of victims in different types of Cyber Crime (India)



Increasing usage of technology has lead to an increase in different types of Cyber Crimes committed as seen in the graph above. We can see that crimes due to Non- payment / Non delivery are consistent over all the years. There is a rise in phishing attacks after 2019.

SECONDARY DATA ANALYSIS

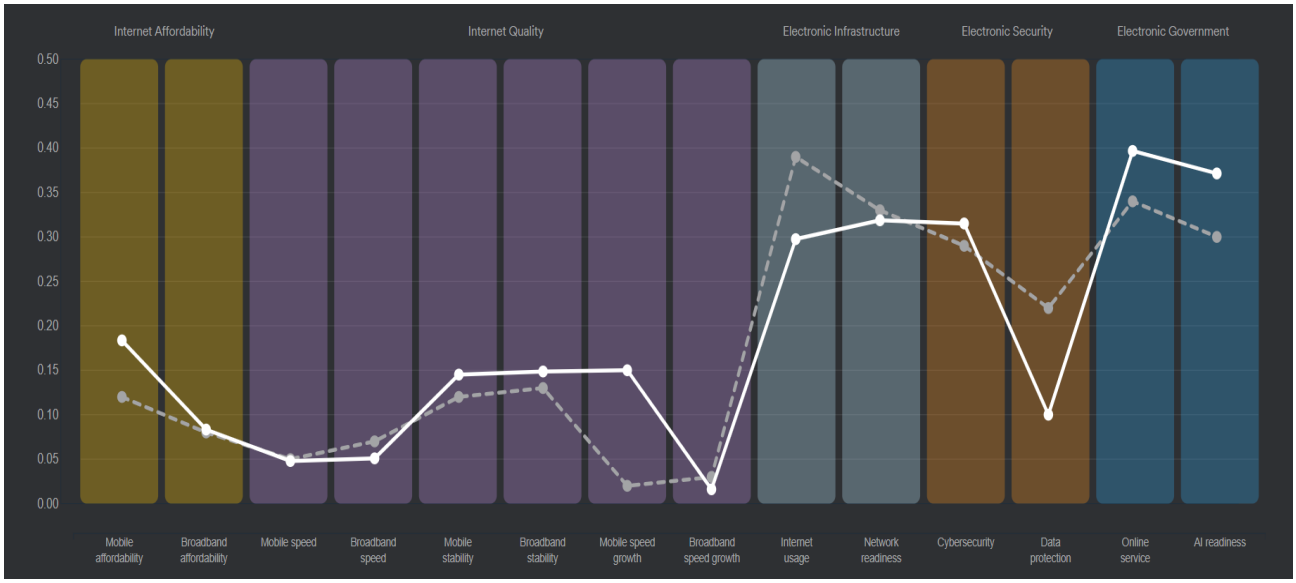
CYBER SECURITY INDEX

To see the preparedness of countries globally to tackle Cyberattacks and to judge the digital quality & electronic infrastructure of countries globally, different Cybersecurity indexes have been designed by organizations around the world.

I) DIGITAL QUALITY OF LIFE INDEX (DQL)

The Digital Quality of Life (DQL) Index reveals insights into factors that impact a country's digital wellbeing and areas that should be prioritized for future improvement. A country's susceptibility to cyberattacks depends on factors such as the robustness of its cybersecurity measures, the level of preparedness against cyber threats, investment in cybersecurity infrastructure, and the effectiveness of legal and regulatory frameworks related to cybersecurity. The relationship between the DQL index and cyberattacks lies in the fact that a higher DQL index often implies a more digitally connected and dependent society. The DQL Index offers a unique perspective into a country's digital quality of life according to five pillars:

1. Internet Affordability-Internet affordability determines how much time people have to work to afford a stable internet connection.
2. Internet Quality-Internet quality measures how fast and stable internet connectivity in a country is and if it's improving.
3. Electronic Infrastructure-Electronic infrastructure determines how well-developed and inclusive a country's existing electronic infrastructure is.
4. Electronic Security-Electronic security measures how safe people are online. E-security shows a country's readiness to counter cybercrimes and its commitment to protecting online privacy.
5. Electronic Government-Electronic government determines how advanced and digitized a country's government services are.



This line graph shows the index of different factors that are significant in calculating Digital Quality life of a country. The dotted line shows the global average index and the main line shows the index of India. We can see that there is a scope of development for India in certain factors such as broadband speed , broadband speed growth , internet usage , data protection and network readiness .

REGRESSION ANALYSIS

Linear regression is a statistical method that is used to predict the value of unknown data using Other related data values. Linear regression is used to study the relationship between a dependent variable and independent variable.

The equation for a simple linear regression model is:

$$Y = \beta o + \beta 1 X + \varepsilon$$

- *Y is the dependent variable*
- *X is the independent variable*
- *β0 is the intercept, which represents the value of y when x=0.*
- $\beta 1$ *is the slope, which represents the change in y for a one-unit change in x*
- ε *is the error term, which represents the difference between the observed value of y and the value predicted by the model. It captures the variability in y that cannot be explained by the linear relationship with x.*

Assumptions of simple linear regression include:

1. *Linearity: The relationship between x and y is linear.*
2. *Independence: Observations are independent of each other.*
3. *Homoscedasticity: The variability of y is constant across all values of x.*
4. *Normality: The residuals are normally distributed.*

A) Simple Linear Regression to study the relationship between Digital Quality Index & Human Development Index.

Y=Human Development Index (HDI)

X=Digital Quality Index (DQL)

The fitted model is:

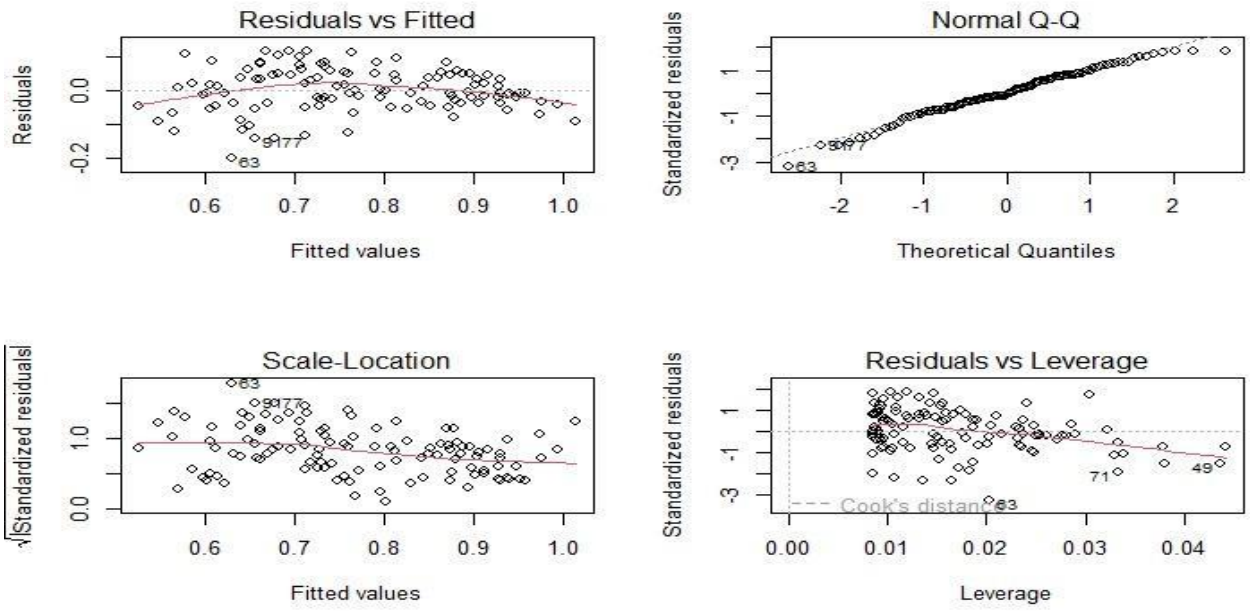
```
Call:
lm(formula = h ~ d)

Coefficients:
(Intercept)            d
     0.4033       0.8022
```

Therefore α=0.4033, β=0.8022

Y=0.4033+0.8022 X

Interpretation: For unit change in DQL there is a 0.8022 change in HDI

Residual Analysis:



We can see from the above plots that all the assumptions have been satisfied

To test the significance of Regression

Hypothesis:

Ho: $\beta 1=0$ ($\beta 1$ is not significant) vs H1: $\beta 1 \neq 0$ ($\beta 1$ is significant)

We reject the null hypothesis if p-value is less than 0.05

```
            Df Sum Sq Mean Sq F value Pr(>F)
d            1 1.6824   1.682   421.1 <2e-16 ***
Residuals  115 0.4594   0.004
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Decision: As we can see the p-value is less than (α=0.05) level of significance, we reject Ho (null hypothesis)

Conclusion: $\beta 1$ is significant i.e The regressor DQL plays a significant role on HDI .

Interpretation : As the regression coefficient for DQL is positive and statistically significant it suggest that countries with high DQL tend to have a high HDI .

Showing the summary statistics of the regression model:

```
Call:
lm(formula = h ~ d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.201499 -0.038611 -0.005559  0.045262  0.117223

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.40327    0.01884   21.41   <2e-16 ***
d            0.80224    0.03909   20.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06321 on 115 degrees of freedom
Multiple R-squared:  0.7855,    Adjusted R-squared:  0.7836
F-statistic: 421.1 on 1 and 115 DF,  p-value: < 2.2e-16
```

As the adjusted R-squared=0.7836, we can say that the independent variable (DQL) explains 78% of the variation in the target variable (HDI) .

IB) Correlation Matrix

To show the correlation between different indicators of the Digital Quality Index (DQL) a correlation matrix has been plotted.

Indicators: internet affordability, internet quality, electronic infrastructure, electronic security, and electronic government

Countries that invest in improving their e-governance and e-infrastructure are most likely to improve their whole DQL index. This is because out of the 5-pillars e-governance & e-infrastructure have the highest correlation with DQL index (0.92). In contrast, internet affordability demonstrated the lowest correlation (0.58). Focusing on these parameters for developing the DQL index one can reduce the Cyber Crimes in country.

II) GLOBAL CYBERSECURITY INDEX (GCI)

The Global Cybersecurity Index (GCI) is a trusted reference that measures the commitment of countries to cybersecurity at a global level – to raise awareness of the importance and different dimensions of the issue. As cybersecurity has a broad field of application, cutting across many industries and various sectors, each country's level of development or engagement is assessed along five pillars –

(i) Legal Measures,

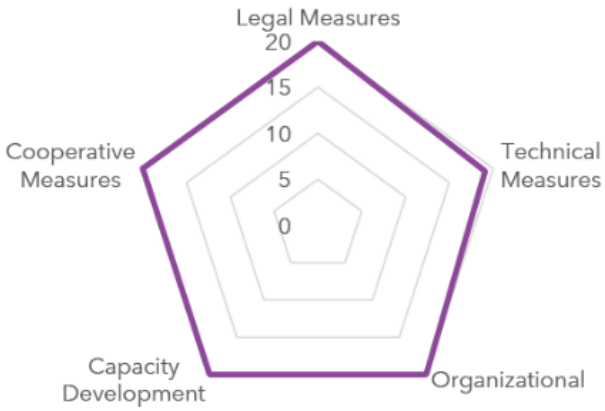
(ii) Technical Measures,

(iii) Organizational Measures,

(iv) Capacity Development, and

(v) Cooperation – and then aggregated into an overall score.

India (Republic of)



Development Level:
Developing Country

Area(s) of Relative Strength
Legal, Cooperative, Capacity Development Measures

Area(s) of Potential Growth
Organizational Measures

Overall Score	Legal Measures	Technical Measures	Organizational Measures	Capacity Development	Cooperative Measures
97.49	20.00	19.08	18.41	20.00	20.00

Source: ITU Global Cybersecurity Index v4, 2020

Across the five pillars India has achieved certain scores leading to an overall score of 97.49 and placing it in the 10th position of the Global Cybersecurity Index

KRUSKAL WALLIS TEST

A) To compare the GCI indexes of Low, Medium & High GDP countries

Group 1: GCI Index of countries with Low GDP

Group 2: GCI Index of countries with Medium GDP

Group 3: GCI Index of countries with High GDP

Low GDP Range: Below 69 billion

Medium GDP Range: 69 billion- 504 billion USD

High GDP Range- Above 504 billion USD

Hypothesis

Ho: the median value of the GCI index of all three groups is same

H1: the median value of the GCI index of all three groups is different

Result:

```
Kruskal-Wallis rank sum test

data:  A by B
Kruskal-Wallis chi-squared = 78.476, df = 2, p-value < 2.2e-16
```

Decision: As p-value is less than level of significance ($\alpha = 0.05$), we reject null hypothesis

Conclusion: The median value of the GCI index of all three groups is different i.e the median value of the GCI Index for Low GDP, Medium GDP & High GDP countries is different

III) NATIONAL CYBERSECURITY INDEX (NCSI)

The National Cyber Security Index is a global live index, which measures the preparedness of countries to prevent cyber threats and manage cyber incidents. The indicators of the NCSI have been developed according to the national cyber security framework.

The NCSI focuses on measurable aspects of cyber security implemented by the central government:

1. Legislation in force – legal acts, regulations, orders, etc.
2. Established units –existing organizations, departments, etc.
3. Cooperation formats – committees, working groups, etc.
4. Outcomes – policies, exercises, technologies, websites, programmes, etc.

Indicators:

General cyber security indicators

Baseline cyber security indicators

Incident and crisis management indicators

WILCOXON SIGN RANK TEST

B) Comparing NCSI & GCI (Wilcoxon Sign Rank Test)

To check if there's a statistically significant difference between the NCSI and GCI scores within your paired data set

Hypothesis

Ho: The median value of the two scores are same vs

H1: The median value of the two scores are different

Result:

```
Wilcoxon signed rank test with continuity correction

data:  X1 and X2
V = 13568, p-value = 6.152e-08
alternative hypothesis: true location shift is not equal to 0
```

Decision: As the p-value is less than the (α=0.05) level of significance, we reject Ho

Conclusion: The median value of the two scores are different .

Interpretation: The rankings given by GCI & NCSI differ for the same country indicating different cybersecurity preparedness for the country .

TIME SERIES ANALYSIS

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. Time series data is generally composed of different components that characterize the patterns and behavior of the data over time. By analyzing these components, we can better understand the dynamics of the time series and create more accurate models. Four main elements that make up a time series data set are trend , seasonality ,cycle and irregularity.

***Trends** show the general direction of the data, and whether it is increasing, decreasing, or remaining stationary over an extended period of time.*

***Seasonality** refers to predictable patterns that recur regularly, like yearly retail spikes during the holiday season. Seasonal components exhibit fluctuations fixed in timing, direction, and magnitude.*

***Cycles** demonstrate fluctuations that do not have a fixed period, such as economic expansions and recessions. These longer-term patterns last longer than a year and do not have consistent amplitudes or durations.*

***Irregularity** includes unpredictable, erratic deviations after accounting for trends, seasonality, and cycles.*

***Single exponential smoothing (SES) :** It is the method of time series forecasting used with univariate data with no trend and no seasonal pattern. It needs a single parameter called alpha (a), also known as the smoothing factor. Alpha controls the rate at which the influence of past observations decreases exponentially. The parameter is often set to a value between 0 and 1.*

Mathematical Model:

$\hat{y}_{t+1} = \alpha y_t + (1- \alpha)\hat{y}_t$

here,

$\hat{y}_{t+1}$ *+ the forecast for the next time period t+1.*

α *= smoothing factor of data;* $0 < \alpha < 1$ *, t = time period*

y_t *= the observed value at time t.*

$\hat{y}_t$ *=the forecast for time t.*

Double Exponential Smoothing

Double exponential smoothing is a time series forecasting method that adds a trend component to the simple exponential smoothing method. It is used for short-term forecasting and has two smoothing parameters, alpha and beta, to smooth the level and trend components, respectively. This technique is commonly used in business forecasting and inventory management.

Mathematical Model:

Level equation: $L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1})$

Trend equation: $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$

where,

L_t= smoothed level at time t

T_t= smoothed trend at time t

Y_t = actual observation at time t

α = smoothing parameter for level ($0 < \alpha < 1$)

β = smoothing parameter for trend ($0 < \beta < 1$)

L_{t-1} = smoothed level at time t-1

T_{t-1}= smoothed trend at time t-1

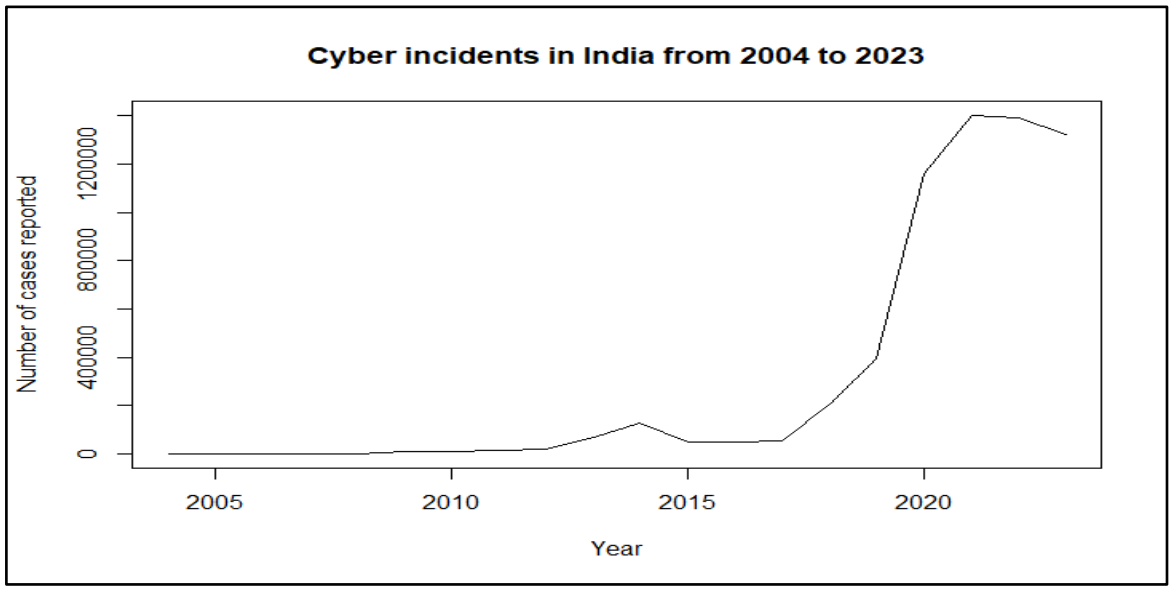
Forecast equation: $Y_{t+1} = L_t + T_t$

$Y_{t+1} = L_t + T_t$

Where, Y_{t+1} is the forecast for the next period.

A) Time Series Analysis for forecasting future values of cyber incidents in India:

1) Time Series Plot



By looking at the above graph it is clear that cybercrime incidents are increasing very rapidly in India. The sudden spike in number of incidents near 2020 can be a result of COVID 19 pandemic as lockdowns and social distancing measures led to a surge in online activities such as shopping, banking, and socializing. Cybercriminals took advantage of this increased online presence to launch cybercrimes.

Time Series after Smoothing

Single exponential smoothing:

Smoothing parameters: α: 0.9999225

SSE=719674401679

Double Exponential Smoothing:

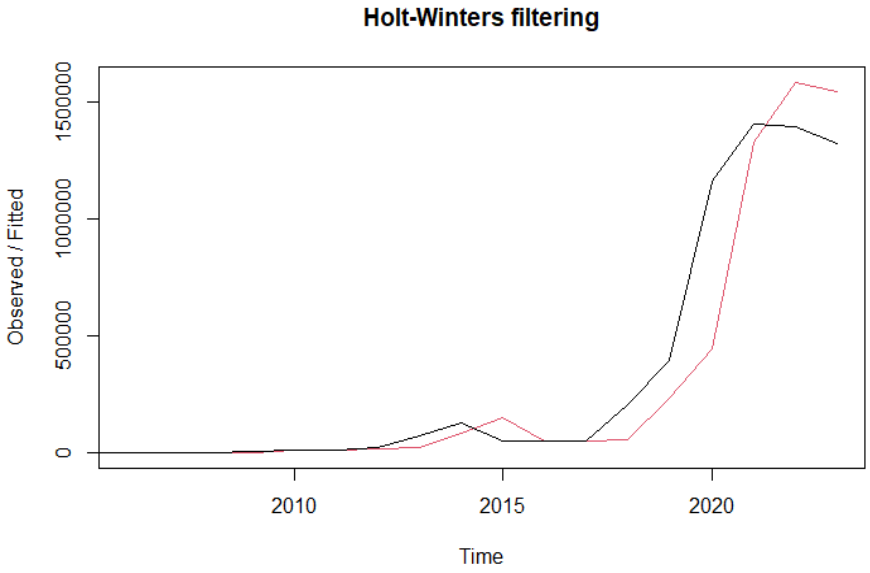
Smoothing parameters:

α: 1

β: 0.1622423

SSE=660050827911

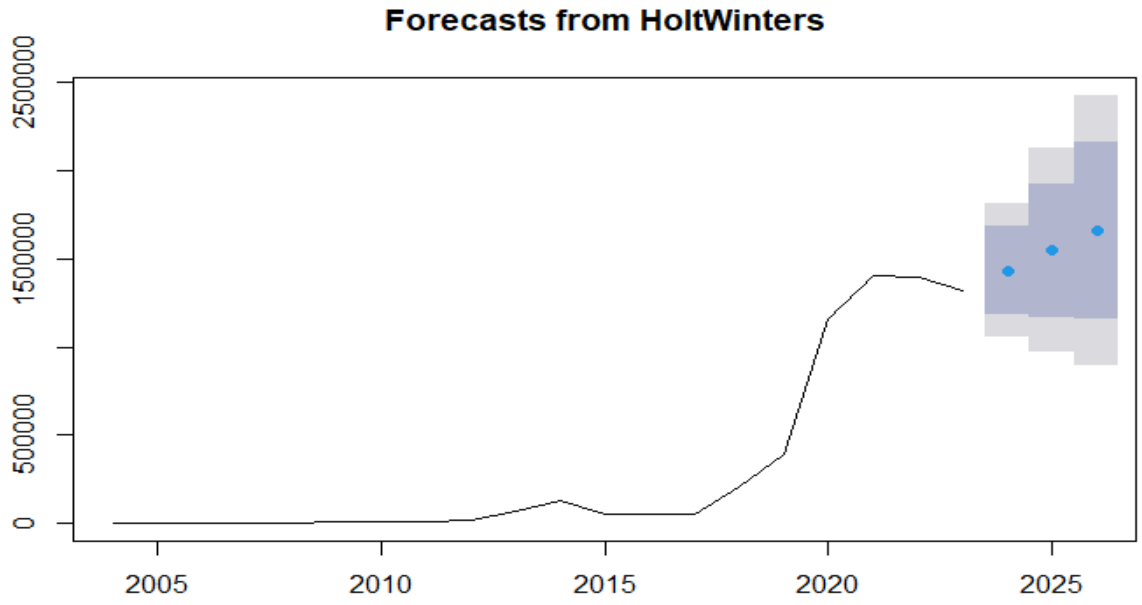
As SSE (Sum of square of errors) for double exponential smoothing is less, hence we use the same model for forecasting the future values.



Forecasting values for the years 2024, 2025 and 2026

2024	1433683
2025	1547260
2026	1660838

Conclusion**:** Estimated number of cyber incidents in India for the year 2024,2025 and 2026 are 1433683,1547260,1660838 respectively



Phishing, Ransomware & Data Breach are the major Cyber Crime attacks

Phishing:

Nearly 1 billion emails were exposed in a single year, affecting roughly 1 in 5 internet users globally Data breaches cost businesses an average of $4.35 million in 2022, and phishing is a major access point for these breaches. Phishing attacks themselves are estimated to have stolen $44.2 million in 2021, with an average loss of $136 per attack. Phishing attempts often impersonate trusted brands, with Yahoo currently being the most impersonated at 20%

Ransomware:

Around 236.1 million ransomware attacks occurred globally in the first half of 2022 alone. A significant portion of data breaches (24%) involve a ransomware component. These attacks can disrupt operations and cause significant financial losses for businesses

Data Breaches:

Data breaches are a concerningly common occurrence. Nearly 1 billion emails were exposed in a single year, and businesses experienced an average of one breach per year according to a 2022 report

CHECKING NORMALITY OF VARIABLES

Hypothesis:

Ho : Given variable is normally distributed

H1: Given variable is not normally distributed

Decision : For p-value less than 0.05 we reject Ho otherwise we accept Ho

Variable	P-value	Decision
Number of phishing attacks	0.5143	Accept Ho
Losses due to ransomware attacks	0.2401	Accept Ho
Average cost of a data breach	0.00704	Reject Ho

As the No of phishing attacks and losses due to ransomware attacks are normally distributed we use ANOVA for further analysis . As average cost of a data breach does not follow normality we use Kruskal Wallis test for analysis .

B 1.) PHISHING

A) Comparing number of Phishing attacks over different sectors

ANALYSIS OF VARIANCE (ANOVA):

Response variable**:** Number of phishing attacks

Treatments**:** Different sectors/industries

X1: Financial institution , X2: Webmail , X3: Social media , X4: Logistics , X5: Payment

X6: E-commerce

Blocks: Different Years (2020 to 2023)

Hypothesis**:**

For Treatments:

Ho: Average number of phishing attacks for different sectors is same

H1: Average number of phishing attacks for different sectors is significantly different

For Blocks:

Ho**:** Average number of phishing attacks for different years is same

H1**:** Average number of phishing attacks for different years is significantly different

ANOVA Table:

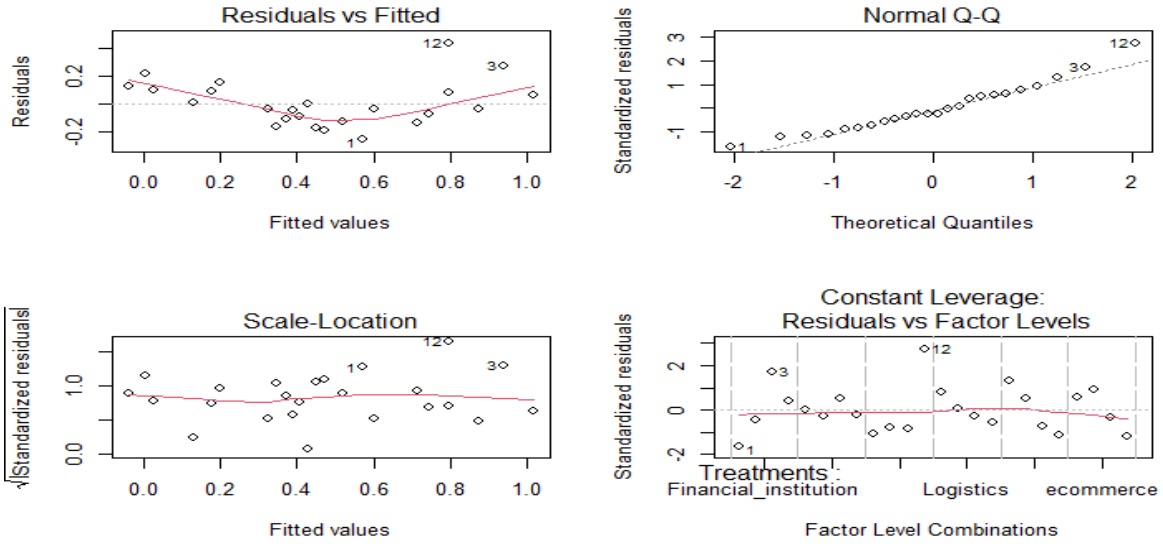
	df	**SS**	**MSS**	**F- ratio**	**P-value**
Treatments	5	1.3532	0.27064	6.615	0.00193
Blocks	3	0.7315	0.28384	5.960	0.00695
Residuals	15	0.6137	0.04091		

Decision: As p-value is less than the level of significance ($\alpha = 0.05$) for both Treatments and Blocks, we reject both the null hypothesis.

Conclusion: 1) Average number of phishing attacks for different sectors is significantly different.

2) Average number of phishing attacks for different years is significantly different

RESIDUAL ANALYSIS:



From the above residual plots, we can conclude that the residuals are identical & independent (iid) normal variates.

POST-HOC ANALYSIS

μi: Average number of phishing attacks encountered by i^{th} sector (for i=1,2,3,4,5,6)

Hypothesis:

Ho: $\mu_i = \mu_j$ ($\forall$ i, j =1,2,3,4,5,6), i≠j

H1: $\mu_i \neq \mu_j$ ($\forall$ i, j =1,2,3,4,5,6), i≠j

Post-Hoc Table:

The following is the table of p-values for different treatment mean combinations.

	μ_1	μ_2	μ_3	μ_4	μ_5
μ_2	0.4680				
μ_3	0.2617	0.6813			
μ_4	0.0052	0.0255	0.0588		
μ_5	0.0087	0.0408	0.0909	0.8197	
μ_6	0.0109	0.0505	0.1106	0.7382	0.9151

Decision: The mean combinations (μ_1,μ_4) , (μ_1,μ_5), (μ_1,μ_6), (μ_2,μ_4), (μ_2,μ_5) show inequality as their p-value is less than level of significance α = 0.05

Conclusion: There is significant difference in means due to μ_1 and μ_2 i.e. Financial Sectors and Web mails

B 2.) RANSOMWARE

2A) Comparing losses due to ransomware attacks in different sectors

ANALYSIS OF VARIANCE (ANOVA)

Response Variable: Losses due to ransomware attacks

Treatments: Different sectors / industries

X1: Health Care ,X2: Software Services ,X3: Financial Services ,X4: Consumer Services ,

X5: Food & Staples , X6: Public Sector ,X7: Professional Services

Blocks: Different Years (2020,2021,2022,2023)

Hypothesis:

For Treatments:

Ho: Average losses due to ransomware attacks for different sectors is the same

H1: Average losses due to ransomware attacks for different sectors is significantly different

For blocks:

Ho: Average losses due to ransomware attacks for different years is the same

H1: Average losses due to ransomware attacks for different years is significantly different

ANOVA TABLE:

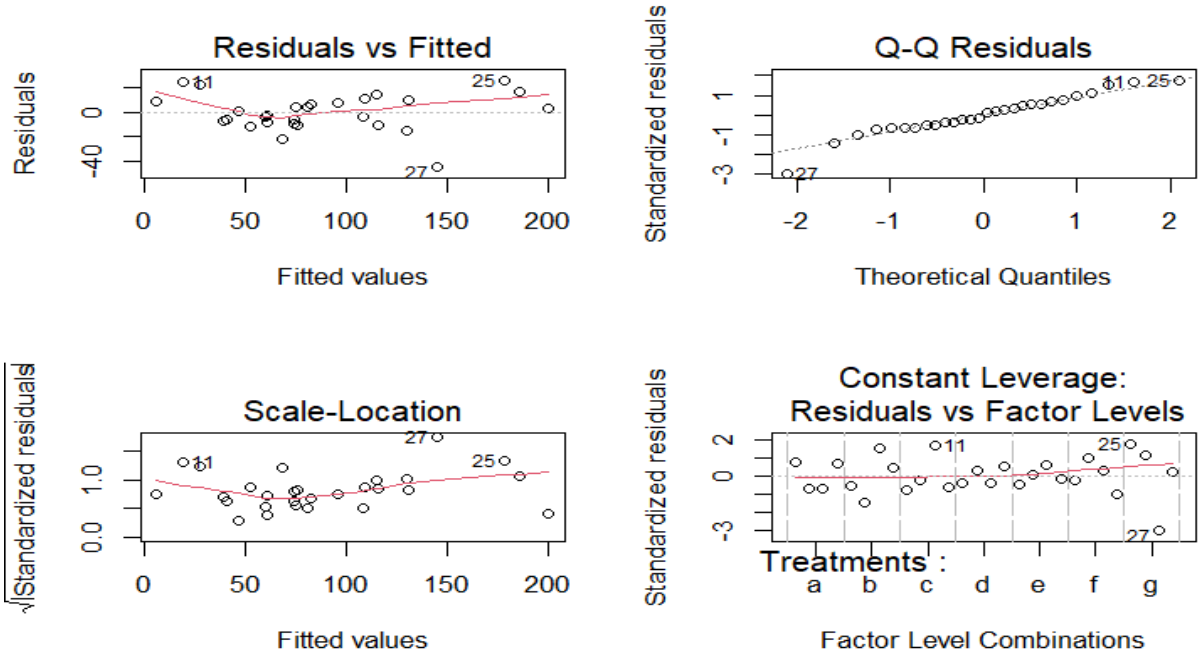
	df	SS	MSS	F-Ratio	P-value
Treatments	6	54375	9063	26.34	5.53 e-08
Blocks	3	11365	3788	11.01	0.000246
Residuals	18	6192	344		

Decision**:** As p-value is less than the level of significance (α = 0.05) for both Treatments and Blocks, we reject both the null hypothesis

Conclusion: 1) Average losses due to ransomware attacks for different sectors is significantly different

2)Average losses due to ransomware attacks for different years is significantly different

RESIDUAL ANALYSIS:



From the above residual plots, we can conclude that residuals are iid normal variates.

POST-HOC ANALYSIS

μi: $Average$ losses due to ransomware in $i\ th$ sector (for i=1,2,3,4,5,6,7)

Hypothesis:

Ho: $\mu_i = \mu_j$ ($\forall$ i, j=1,2,3,4,5,6), i≠j against

H1: $\mu_i \neq \mu_j$ ($\forall$ i, j=1,2,3,4,5,6,7), i≠j

Post - Hoc Table

The following is the table of p-values for different treatment mean combinations

	μ1	μ2	μ3	μ4	μ5	μ6
μ2	0.0281					
μ3	0.0117	0.6908				
μ4	0.1002	0.5300	0.3093			
μ5	0.0026	0.3044	0.5231	0.1056		
μ6	0.9627	0.0311	0.0130	0.1093	0.0029	
μ7	0.0027	1.0 e-05	4.1 e-06	4.6 e-05	9.9 e-07	0.0024

Decision: The mean combinations with μ1 and μ7 shows inequality as their p value is less than level of significance ($\alpha = 0.05$)

Conclusion: There is a significant difference in average losses due to ransomware attacks due to health care and professional services

B)Comparing losses due to ransomware based on company size

ANALYSIS OF VARIANCE (ANOVA)

Response Variable: Losses due to ransomware attacks

Treatments: Company size (small, medium, large)

Small- 1-1000 employees

Medium - 1001 – 10000 employees

Large - 10000 above employees

Blocks: Different Years (2020 to 2023)

Hypothesis:

For Treatments:

Ho: Average losses due to ransomware attacks based on company size is the same

H1: Average losses due to ransomware attacks based on company size is significantly different

For Blocks:

Ho: Average losses due to ransomware attacks for different years is the same

H1: Average losses due to ransomware attacks for different years is significantly different

ANOVA TABLE**:**

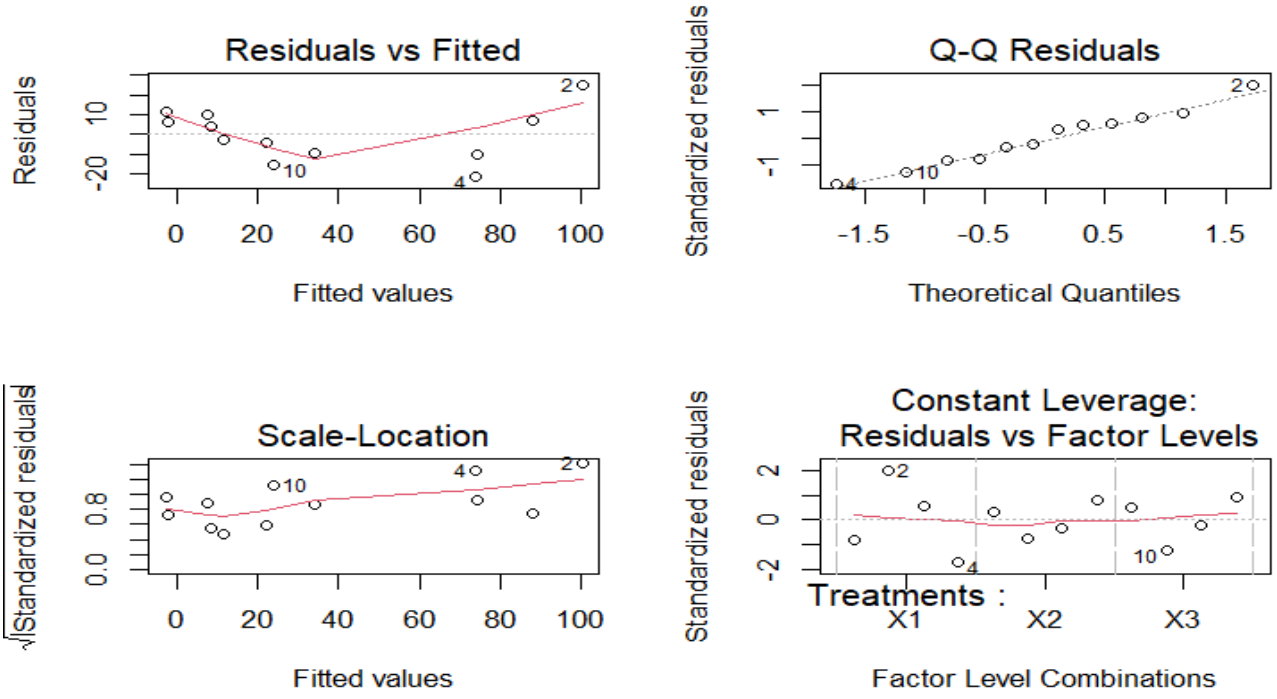
	df	SS	MSS	F-Ratio	P-value
Treatments	2	13737	6869	21.92	0.00174
Blocks	3	1429	476	1.52	0.30266
Residuals	6	1880	313	-	-

Decision**:** As p-value is less than the level of significance (α = 0.05) for Treatments, we reject null hypothesis and p- value is larger than the level of significance (α = 0.05) for blocks, we accept hypothesis.

Conclusion: 1) Average losses due to ransomware attacks based on company size is significantly different

2)Average losses due to ransomware attacks for different years are the same

RESIDUAL ANALYSIS:



From the above residual plots, we can conclude that the residuals are iid normal variates

POST-HOC ANALYSIS

μi: $Average$ losses due to ransomware for i th company size (for i=1,2,3)

Hypothesis:

Ho: $\mu_i = \mu_j$ ($\forall$ i, j= 1,2,3), i≠j against

H1**:** $\mu_i \neq \mu_j$ ($\forall$ i, j=1,2,3), i≠j

POST-HOC ANALYSIS:

The following is the table of p-values for different treatment mean combinations

Pairwise comparisons using t tests with pooled SD

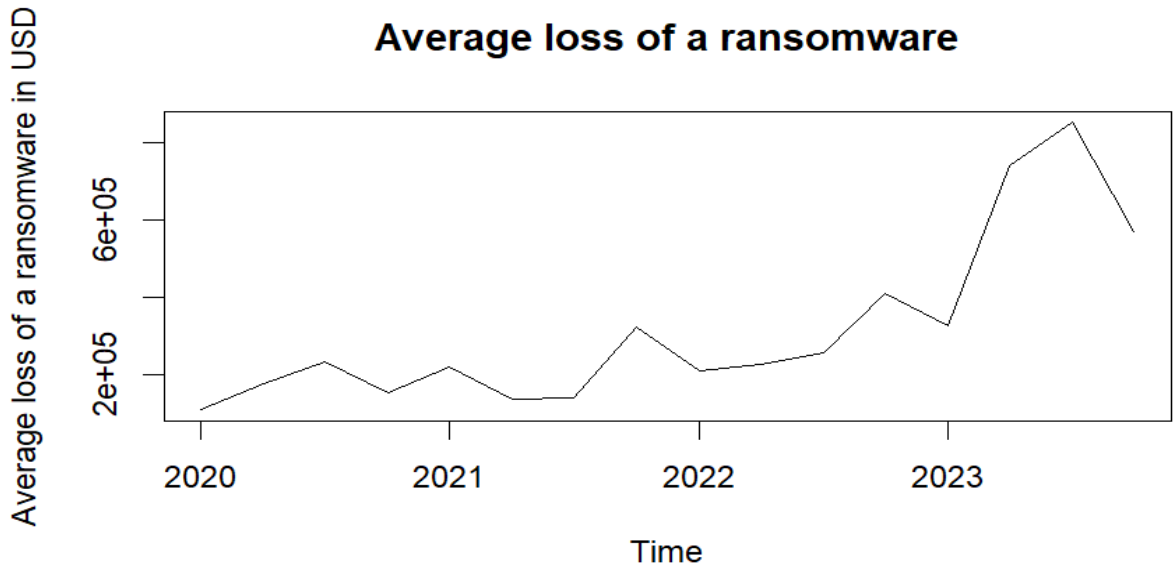
data: p and q

	X1	X2
X2	0.00089	-
X3	0.00032	0.46234

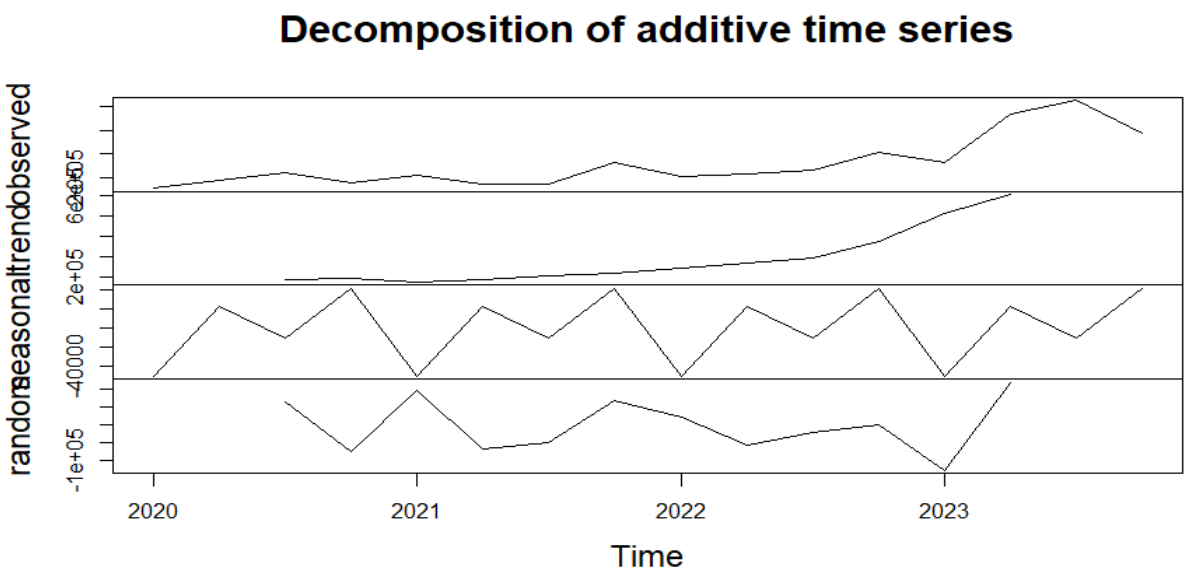
Decision: The mean combinations ($\boldsymbol{\mu}1$, $\boldsymbol{\mu}2$), ($\boldsymbol{\mu}1$, $\boldsymbol{\mu}3$) shows inequality as their P-value is less than level of significance α = 0.05

Conclusion: There is a significant difference in average losses due to ransomware attacks due to small company size

C) Time Series Analysis of Ransomware Loss



Decomposing the Time Series model we obtain an additive type model



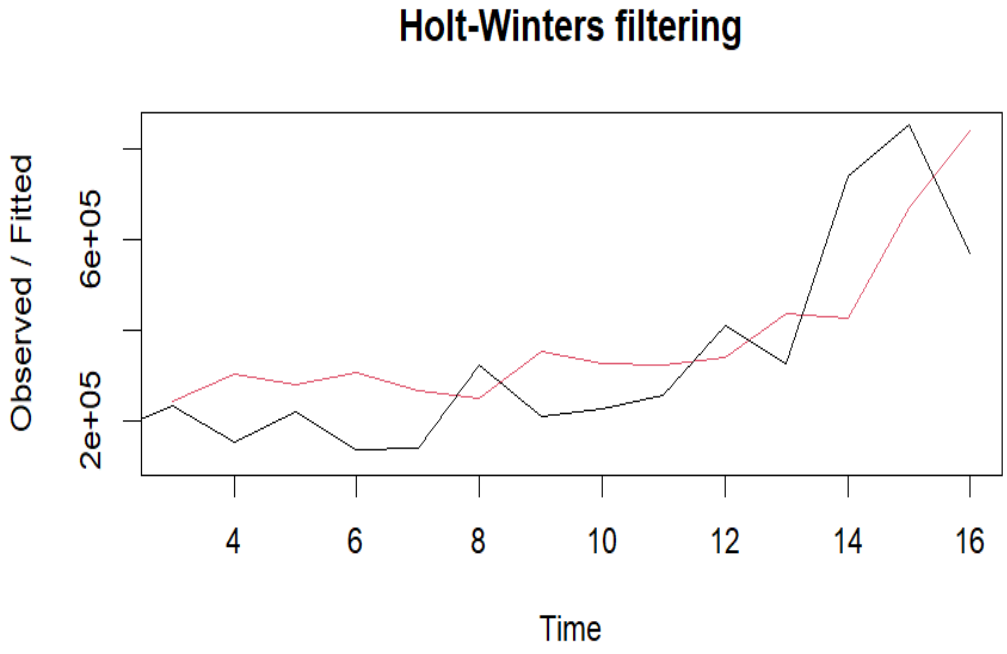
Time Series after smoothing

1. Single Exponential Smoothing

Smoothing parameter α=0.7168983 SSE=343466634481

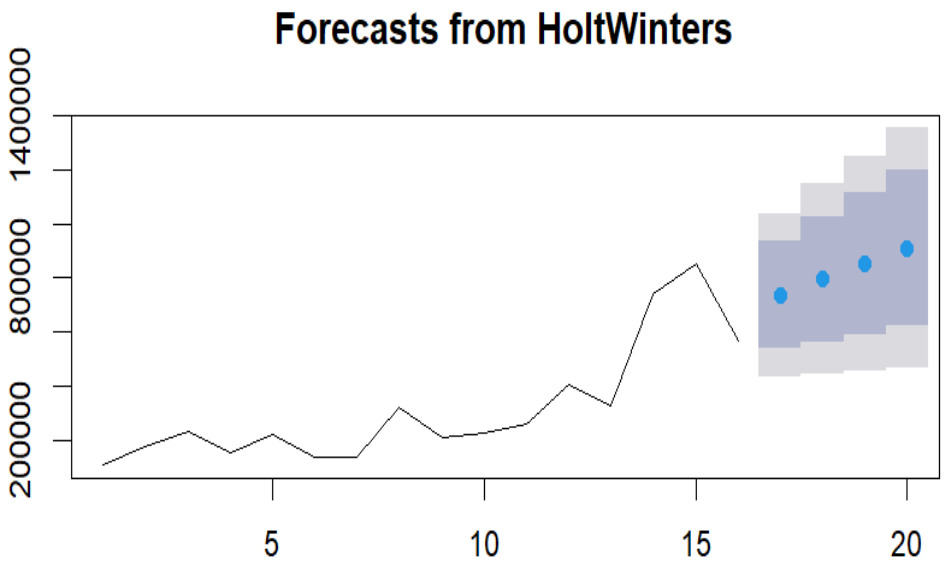
2. Double Exponential Smoothing

Smoothing parameters α= 0.5891321 β=0.0253871 SSE=330830301876



Since the Sum of square of errors for Double exponential smoothing is less we use it to forecast the future values for loss of ransomware

Forecast:

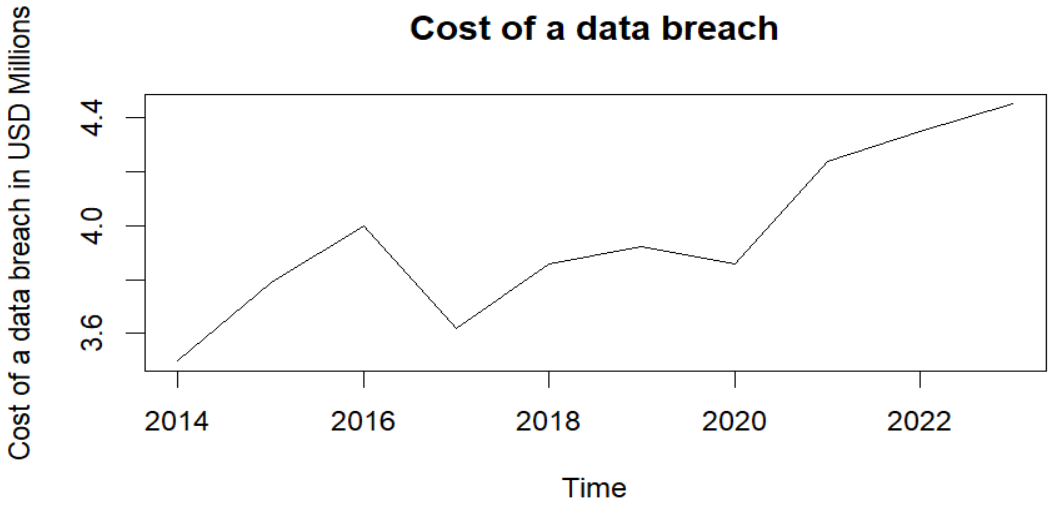


Predicted values of average ransomware loss in USD

2024 Q1	737431.6
2024 Q2	795260.6
2024 Q3	853089.5
2024 Q4	910918.5

B 3) DATA BREACH

3A)Time Series Analysis on average cost of a data breach globally in USD millions



We can see that there is a drastic increase in data breach after 2020

Time Series after Smoothing

1) Single exponential smoothing

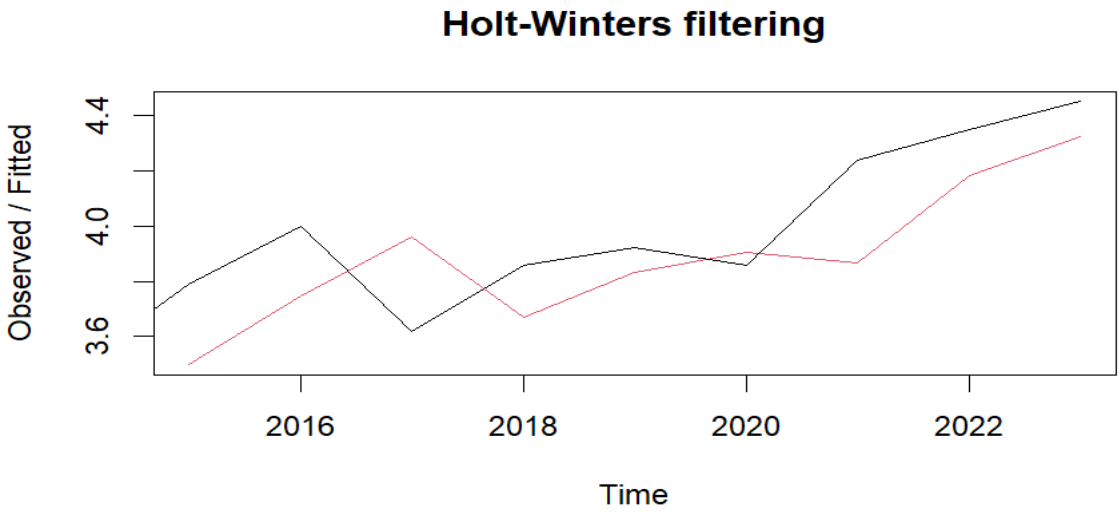
Smoothing factor α=0.8516664, SSE=0.4931566

2) Double Exponential Smoothing

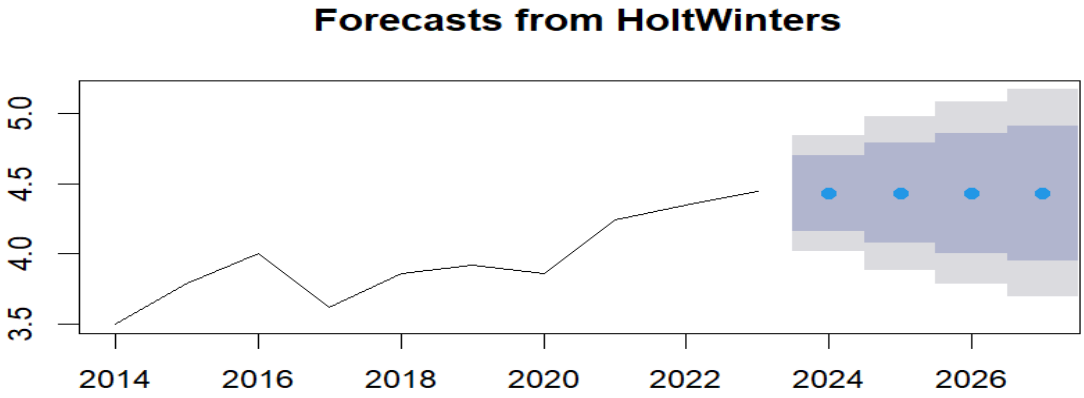
Smoothing parameter α=0.8583699 β=0.2709848

SSE=0.5653708

Since the SSE of single exponential smoothing is less, we will use it to forecast the future values of an average cost of a data breach

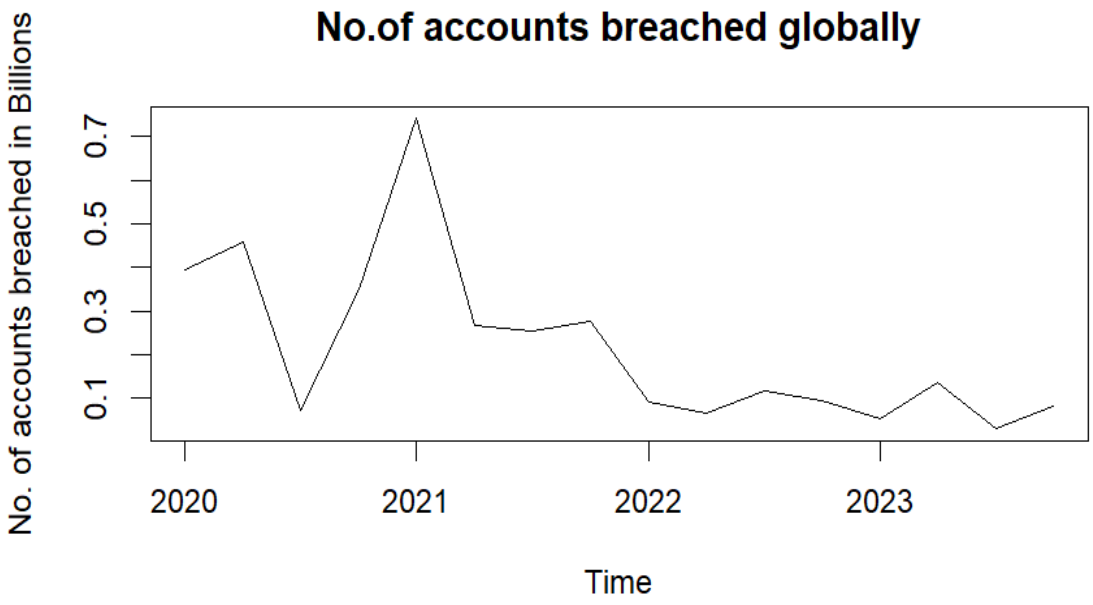


Forecast:



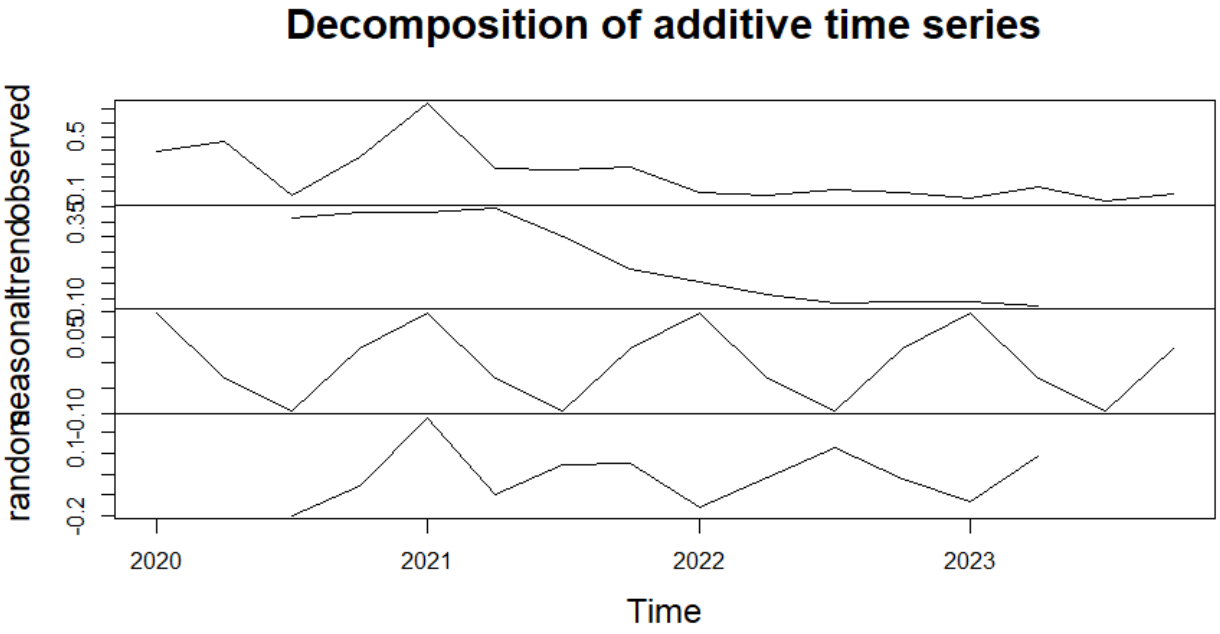
Hence the forecasted number of average cost of a data breach globally for the year 2024-2025 is 4.431529 Million USD

3B) Time Series Analysis on no. of accounts breached globally in billions

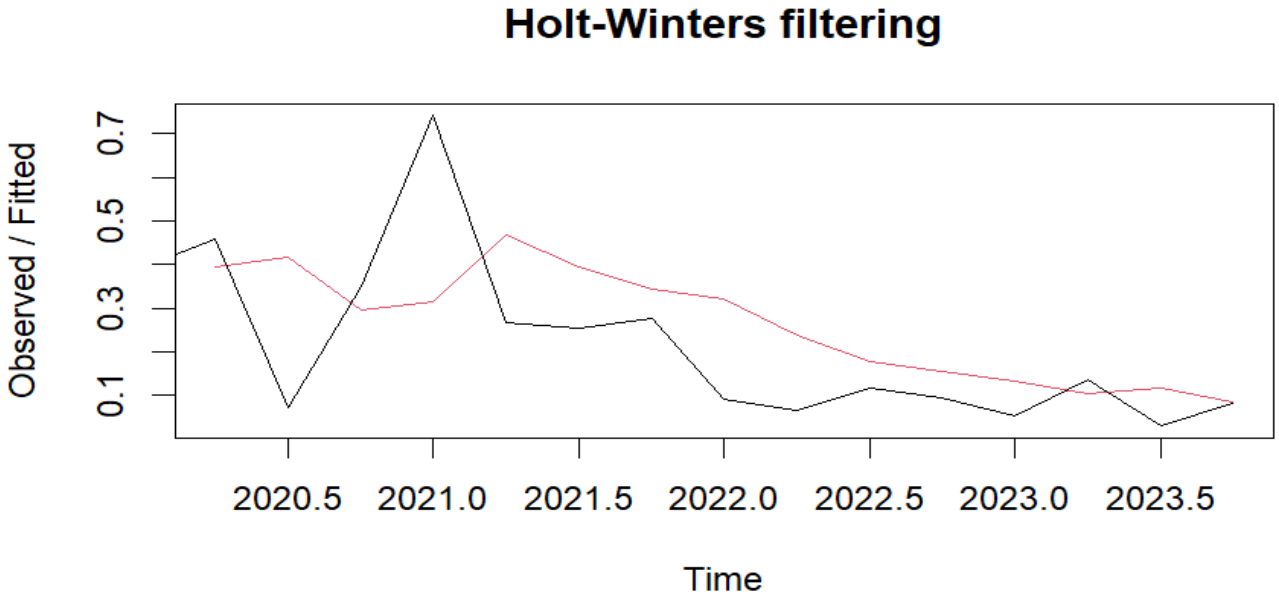


As you can see the number of accounts breached were highest during the COVID-19 period.

Decomposing the Time Series model we obtain an additive type model



Time Series after Smoothing



1. Single Exponential Smoothing

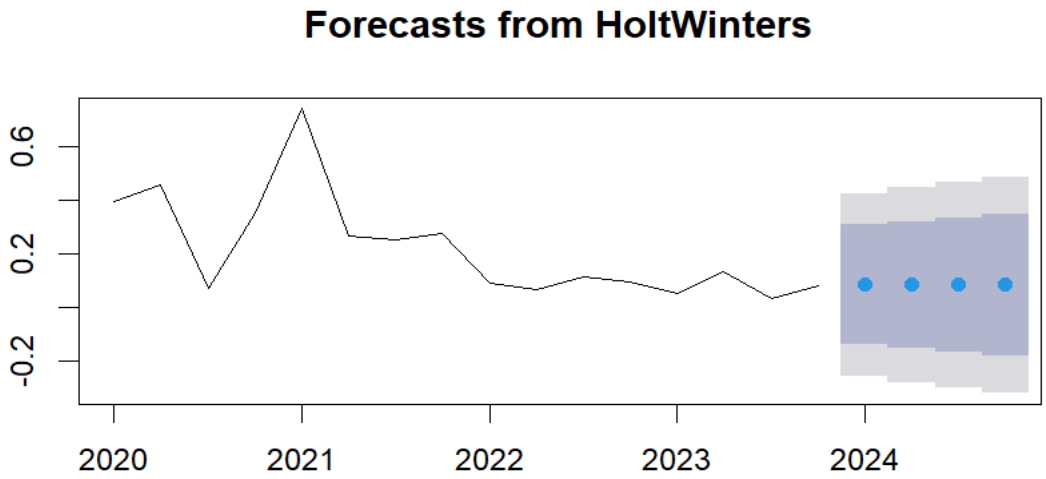
   Smoothing parameter α=0.3563913 SSE=0.4764287

2. Double Exponential Smoothing

   Smoothing parameters α=0.2203657 β=1 SSE=0.5862697

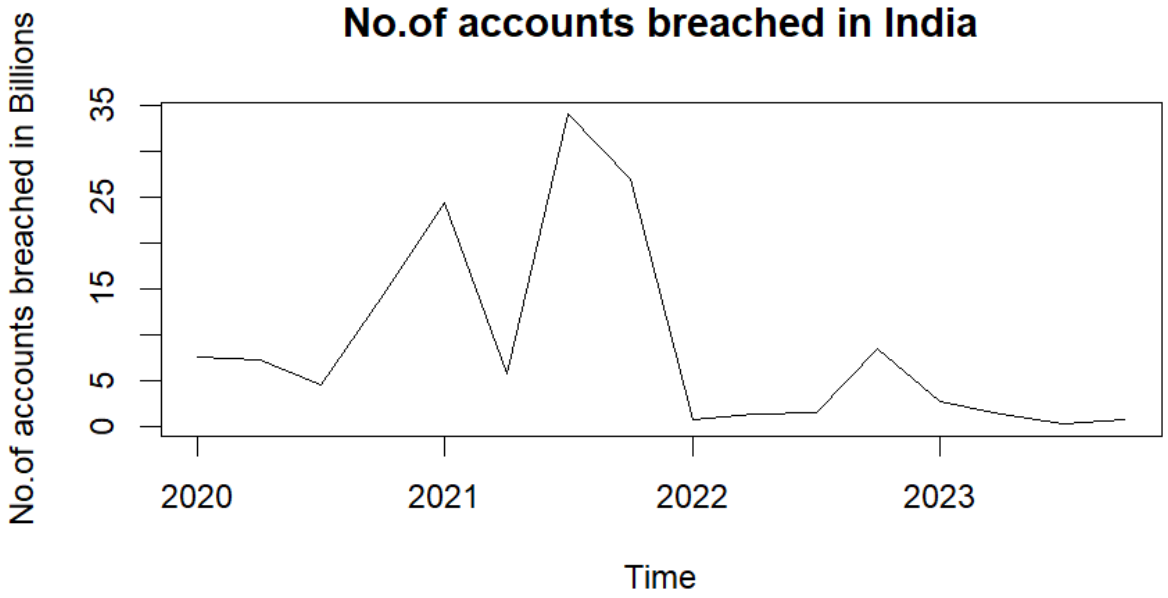
As the SSE for single exponential smoothing is less we use it to forecast the future no. of accounts that can be breached in the year 2024-2025

Forecast:

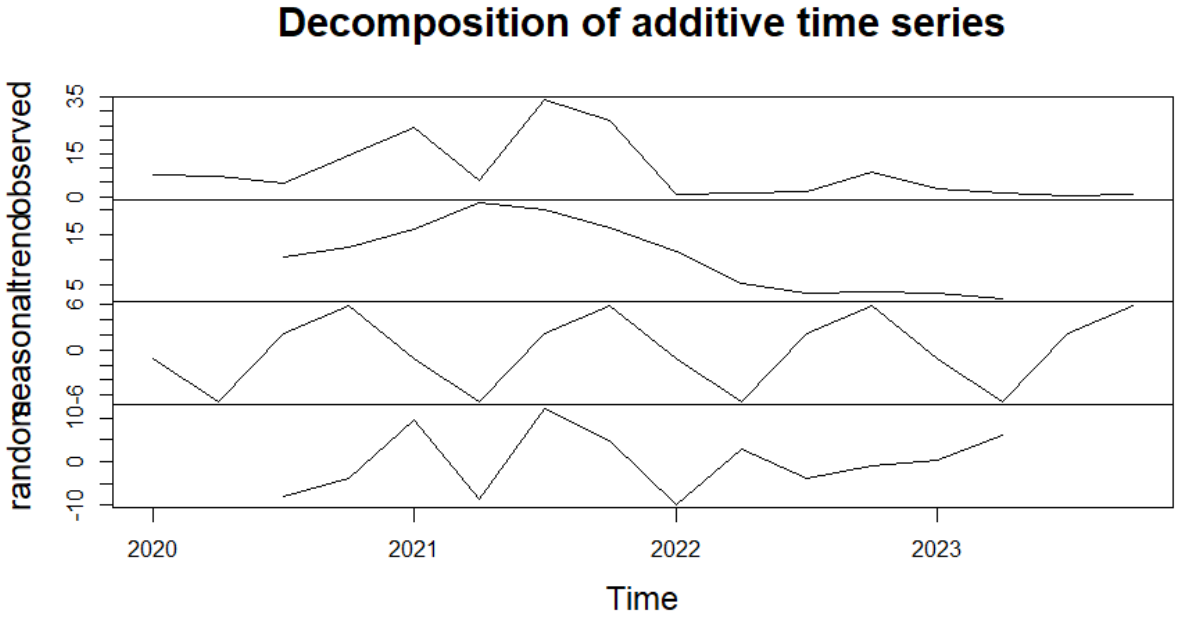


The predicted value of the number of accounts that can be breached in the year 2024-2025 is approximately 0.08370329 billion

3C) Time Series Analysis on number of accounts breached in India in Billions



Decomposing the Time Series model we get an additive type model



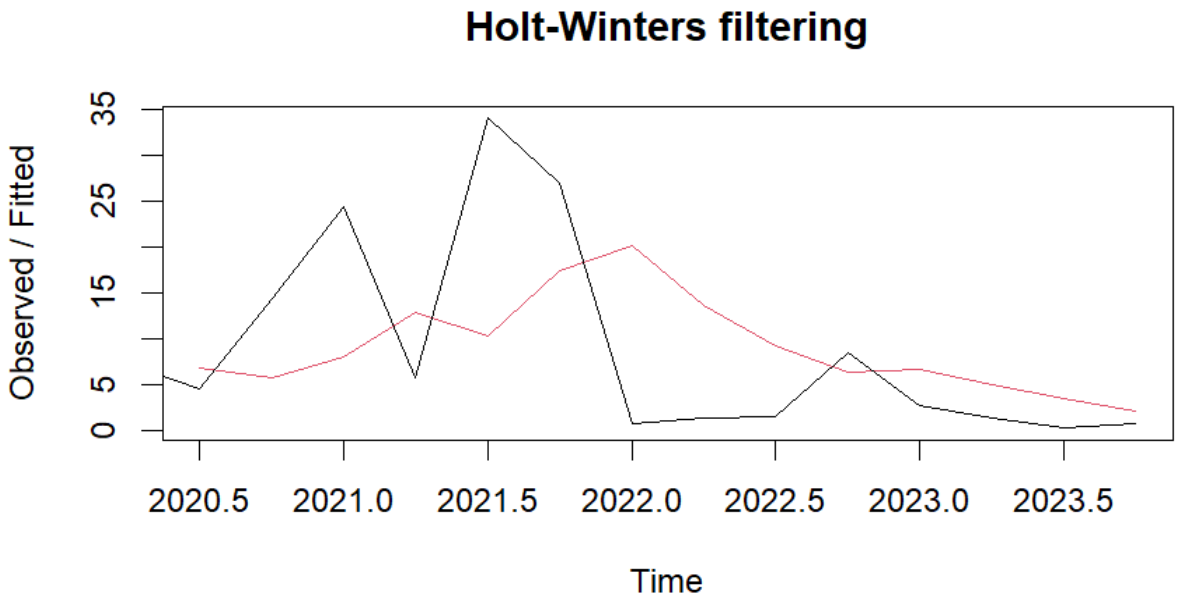
Time Series Smoothing

1) Single exponential smoothing

Smoothing parameters α=0.3755119 SSE=1704.176

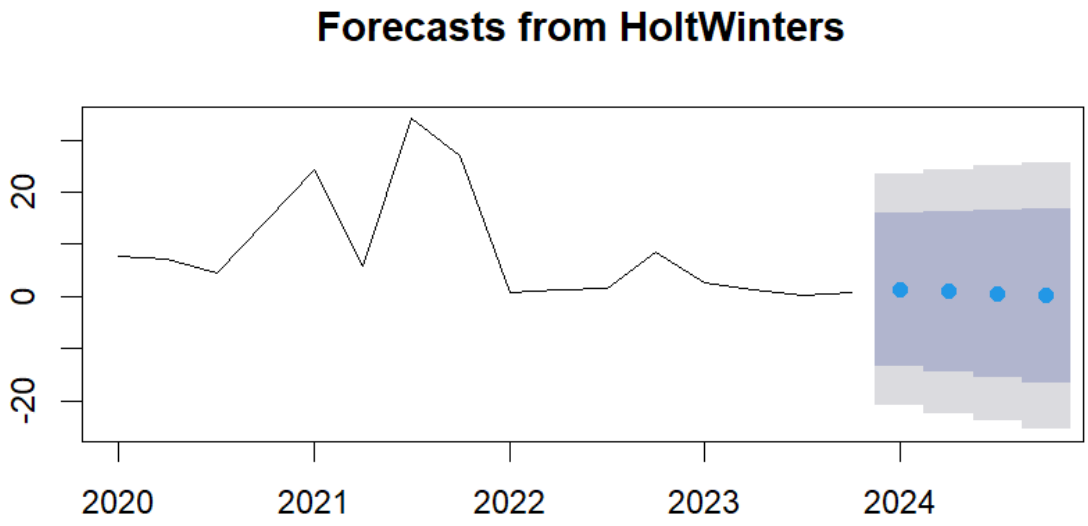
2) Double Exponential Smoothing

Smoothing parameters α=0.3199311 β=0 SSE=1677.229



Since the SSE for double exponential smoothing is less we will use it to predict the future number of accounts that can be breached in India

Forecast:



The predicted values of the number of accounts that can be breached in the year 2024 are

2024 Q1	1.3172467 billion
2024 Q2	0.9333167 billion
2024 Q3	0.5493867 billion
2024 Q4	0.1654567 billion

3D) Kruskal Wallis Test on median value of an average cost of a data breach in USD millions for different sectors

The average total cost of a data breach represents the total cost incurred by an organization as a result of a data breach . This cost provides a comprehensive estimate of the overall impact of a security incident on an organization's finances .

Sectors:

X1=Healthcare , X2=Financial , X3=Pharmaceutical , X4= Energy , X5=Industry

X6= Technology

Hypothesis:

Ho: The median value of an average cost of a data breach in USD Millions is same for different sectors

H1: The median value of an average cost of a data breach in USD Millions is different for different sectors

Result:

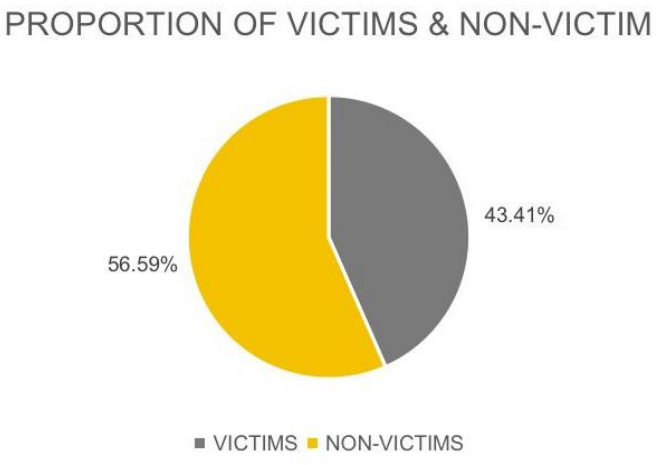
```
Kruskal-Wallis rank sum test

data:  Y by X
Kruskal-Wallis chi-squared = 19.857, df = 5, p-value = 0.00133
```

Decision: Since the p-value is less than the (α=0.05) level of significance, we reject the null hypothesis (Ho)

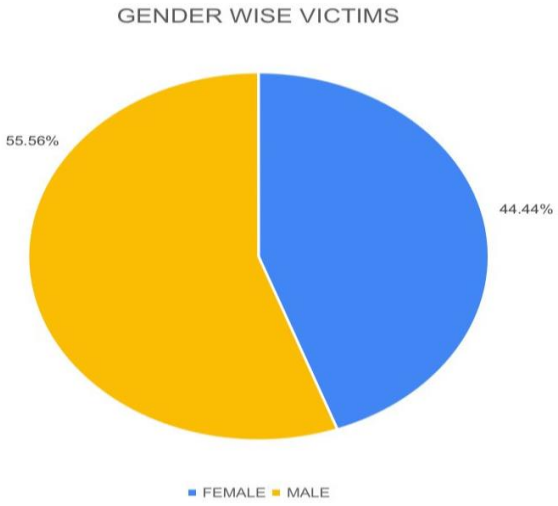
Conclusion: The median value of an average cost of a data breach in USD Millions is different for different sector

PRIMARY DATA ANALYSIS

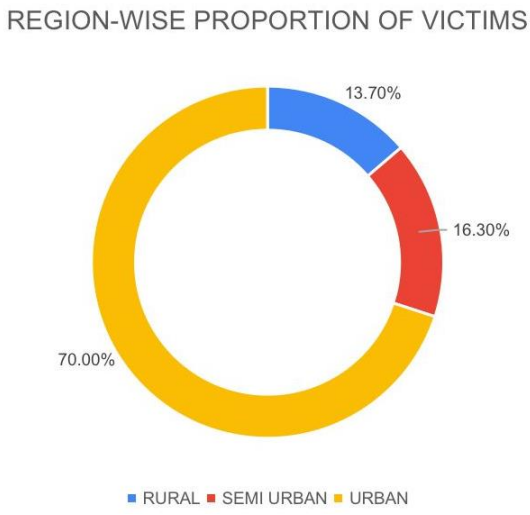
Out Of 622 responses collected from our Primary survey on Cyber Security, Crime & Awareness there are 270 people who have been the victim of Cyber Crime .



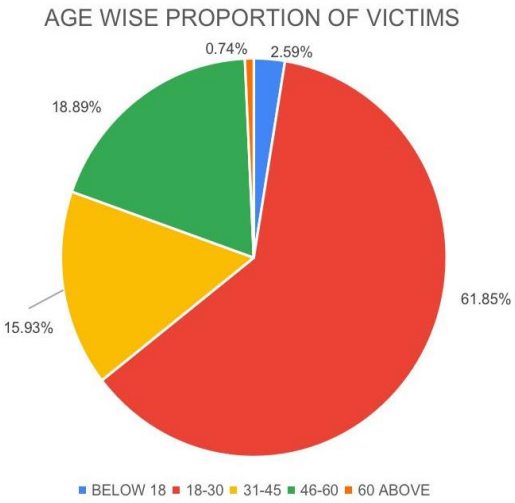
This diagram indicates that among 622 people there are 270 people who has been victim of cyber crime.



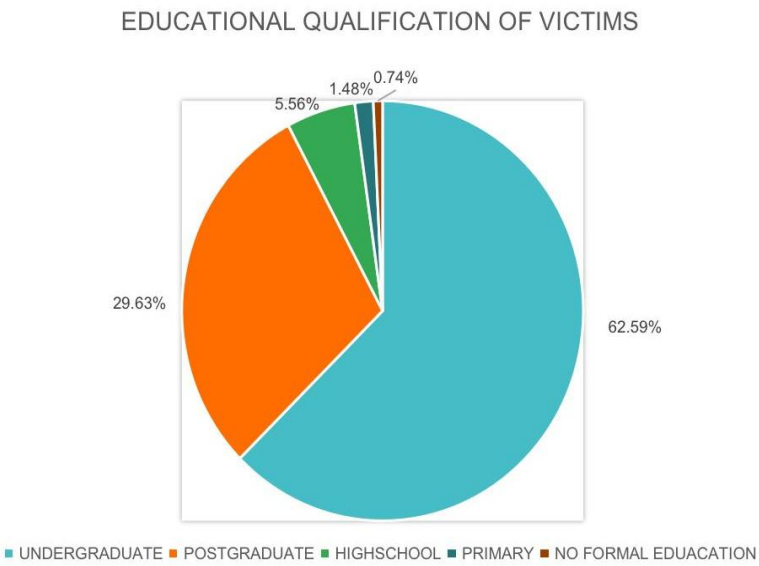
This diagram indicates the proportion of males & females



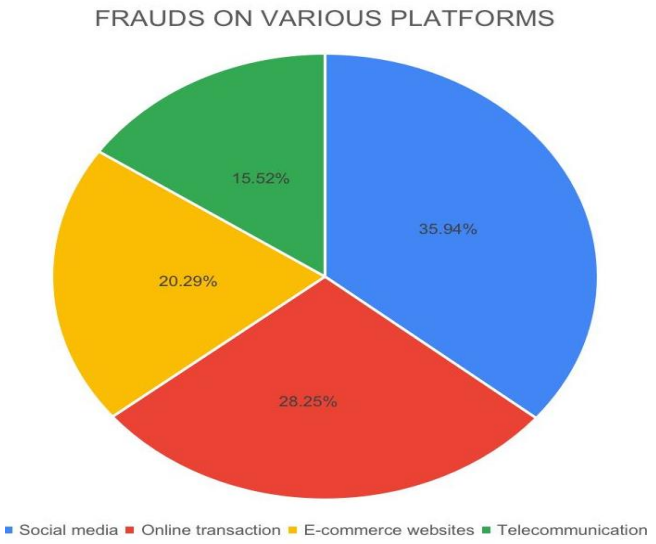
This diagram represents the proportion of people belonging to Urban, Semi-Urban & Rural areas



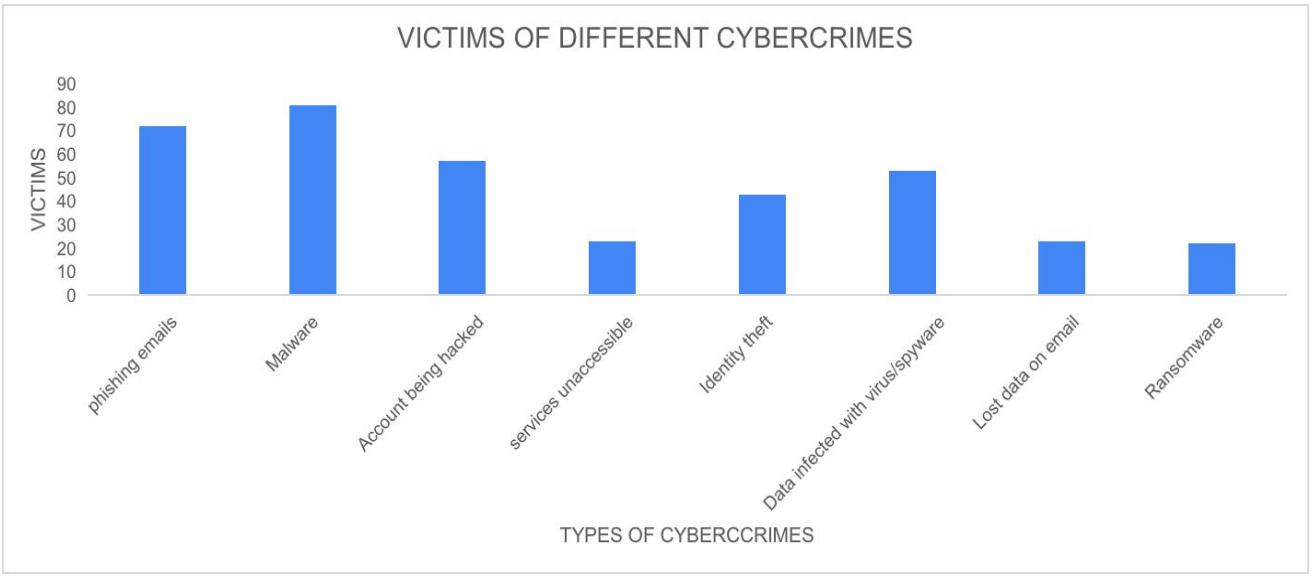
This diagram indicates the proportion of people belonging to different age groups, 170 out of 270 people belong to the age group 18-30



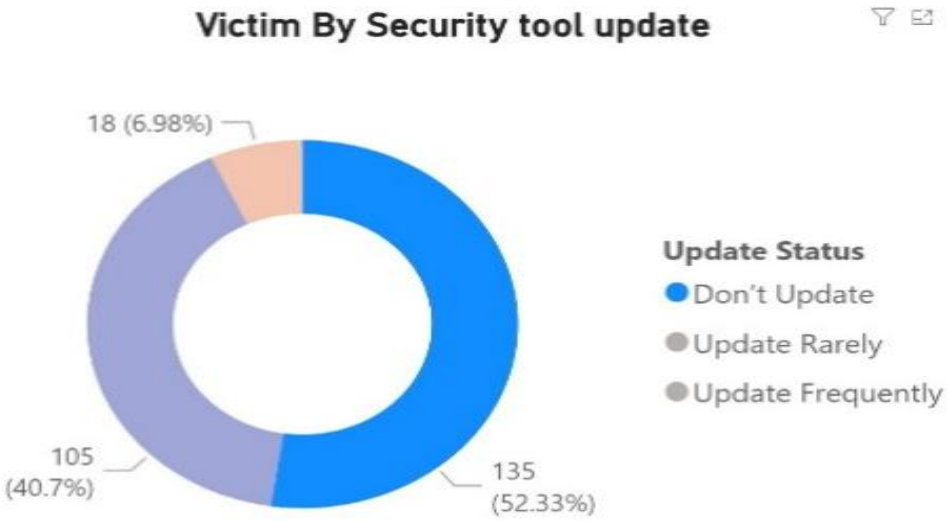
This pie chart indicates the educational qualification of victims from which 169 people are undergraduate .



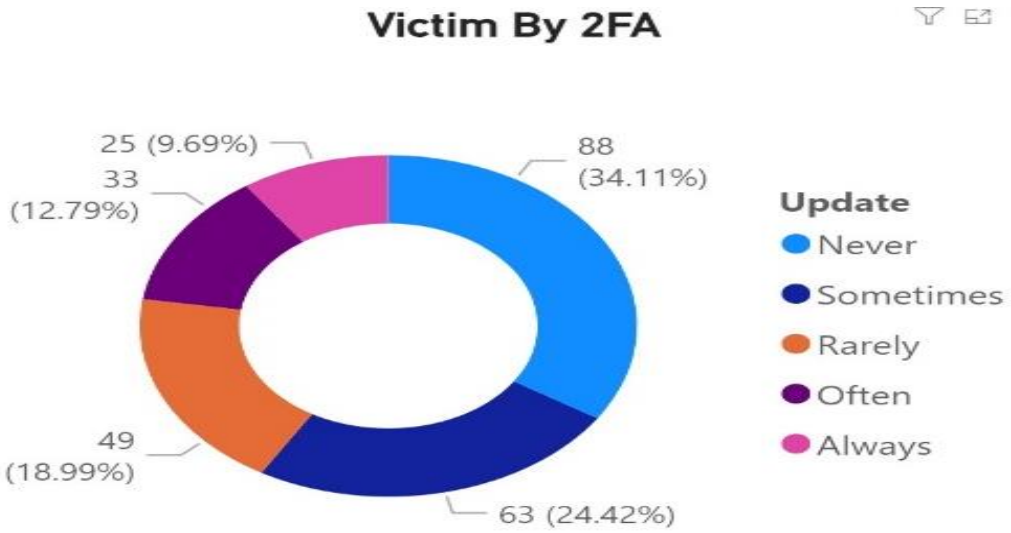
Most the people face cyber crime on social media platform followed by online transaction



From this bar graph we can see the highest number of people have been a victim of malware attacks followed by phishing attacks & hacking of accounts.



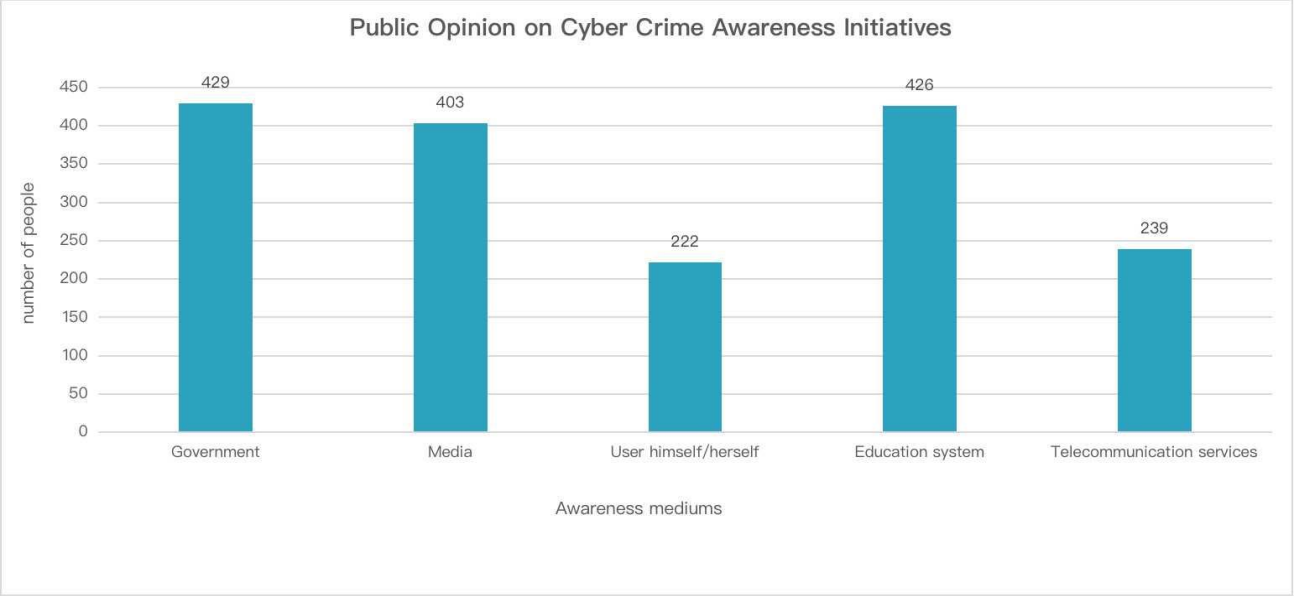
As we can see from these pie-charts most of the victims of cybercrime don't update their security tools . This signifies the importance of updating a device regularly .



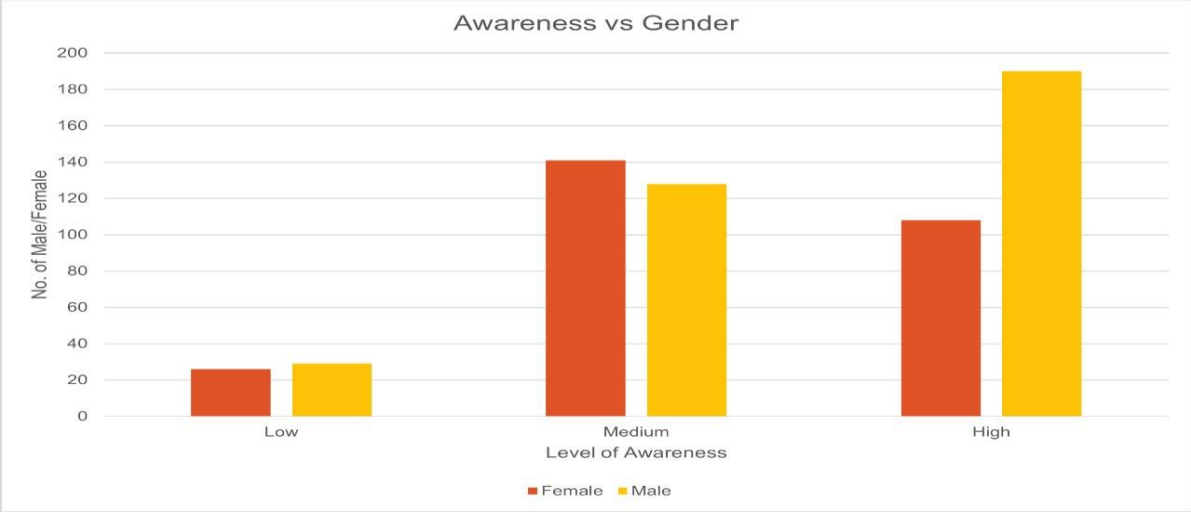
Here we can see that there is high percentage of people who never practice 2 factor authentication (2FA) and have been a victim of cyber crime. On the other hand there are very few people who always practices 2 factor authentication (2FA) and have been a victim of cyber crime . This tells us that practicing 2FA always reduces the chance of cyber attack.

BAR DIAGRAMS:

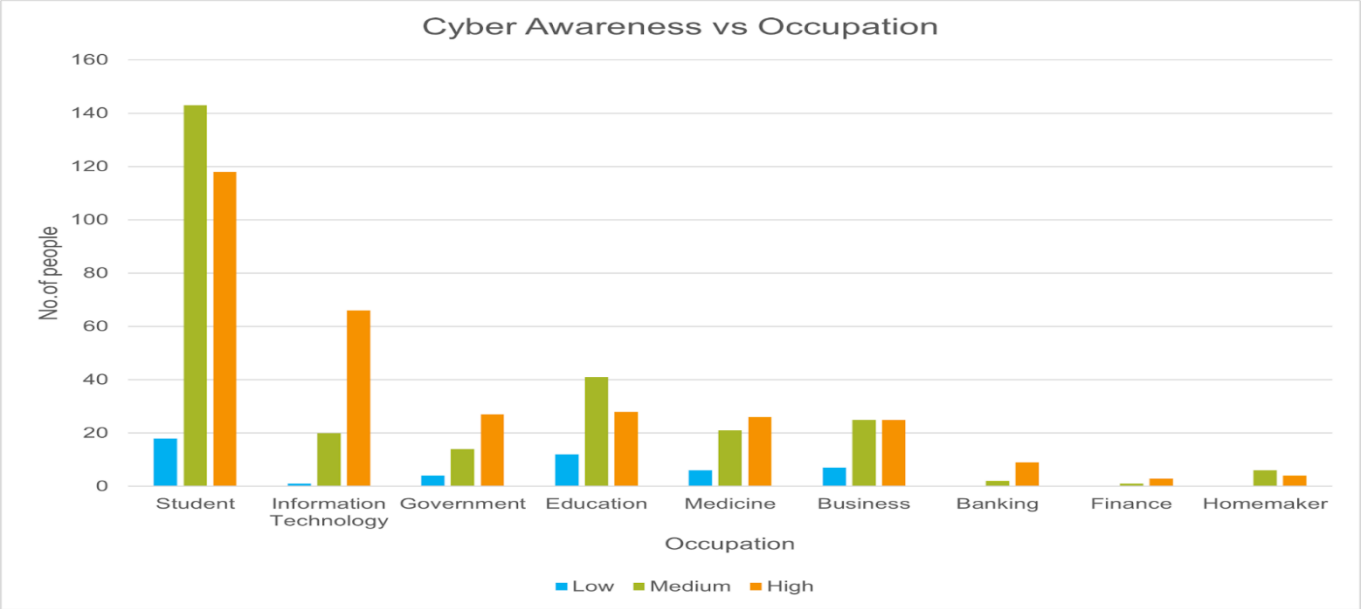
When people were asked who according to them should spread awareness about Cyber Crime, following were the responses



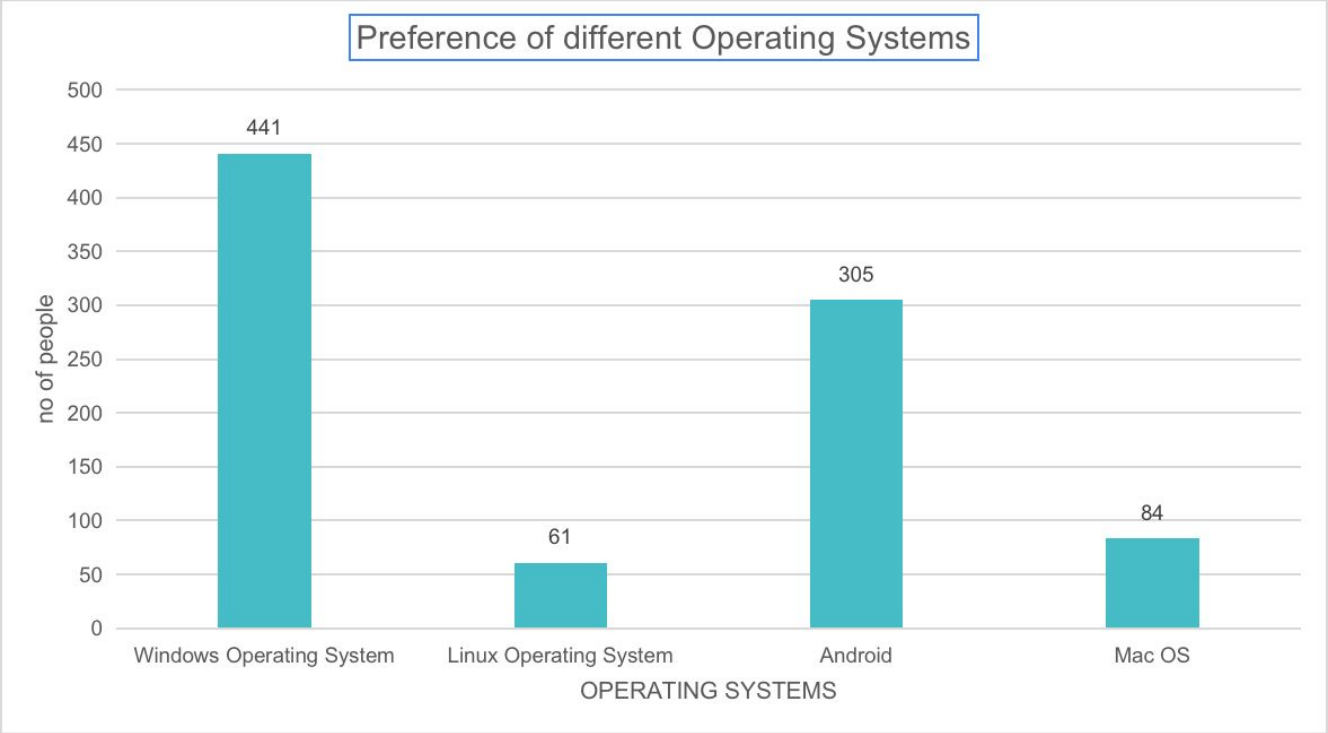
Majority of the people want the Government & the Education System to spread awareness about Cyber Crime



Interpretation: This graph tells us that there are approximately the same no. of males & females in the low & medium Cyber Awareness Score range. Whereas in the high Cyber Awareness score range there are more no. of males than females

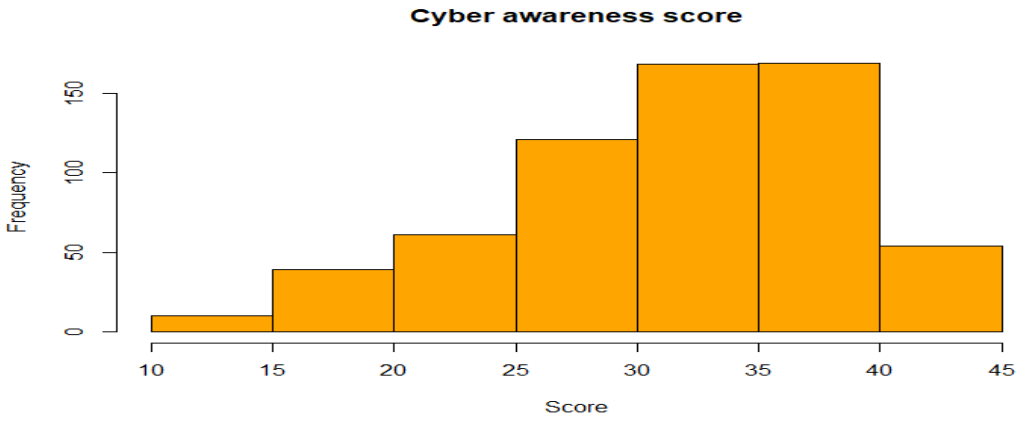


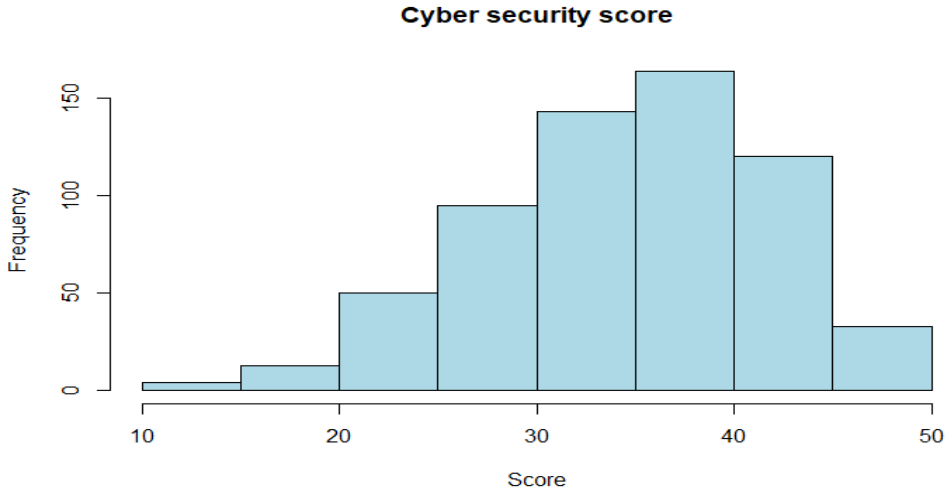
Interpretation: Among students there are only few who have a less cyber awareness score and most of them have a high score. Then the next set of people having a high cyber awareness score belong to the IT sector. Most of the other people have a medium cyber awareness score and work in different sectors such as government, business, medicine etc. This tells us that different occupations have an effect on the cyber awareness score of people.



We can see from this bar graph that most of the people prefer Windows Operating system as their OS, followed by Android

In order to proceed with further tests we need to check if the variables under consideration are normally distributed or not





Interpretation: It is quite clear from the above histograms that Cyber awareness and Cyber security scores are not normally distributed. We can interpret that more no. of people are aware and secure since most of our data is collected from the age group 18-30

I) To check the normality of variables

SHAPIRO-WILK NORMALITY TEST

Hypothesis:

Ho: Given variable is normally distributed

H1: Given variable is not normally distributed

Result:

For Cyber awareness score

```
        Shapiro-Wilk normality test

data:  CAS
W = 0.965, p-value = 5.111e-11
```

For Cyber security score

```
        Shapiro-Wilk normality test

data:  CSS
W = 0.98193, p-value = 5.909e-07
```

Decision: As p-value is less that level of significance (α = 0.05), we reject null hypothesis for both the variables.

Conclusion: 1) Cyber awareness score is not normally distributed

2) Cyber security score is not normally distributed

As the variables are not normally distributed, we perform non-parametric tests for further analysis.

CHI – SQUARE GOODNESS OF FIT

Hypothesis :

Ho: Fitting of uniform distribution is good (proper)

H1: Fitting of uniform distribution is not proper

```
> observed=c(75,80,55,25,45,50,25,25)
> expected=c(rep(1/8,8))
> chisq.test(x=observed,p=expected)

        Chi-squared test for given probabilities

data:  observed
X-squared = 71.579, df = 7, p-value = 7.082e-13
```

Decision : As p-value is less than level of significance (α= 0.05), we reject the null hypothesis

Conclusion : Fitting of uniform distribution is not good for the given data .

Interpretation : This signifies that all the crimes have different frequencies of happening .

CHI-SQUARE TEST

A) To check independence of Cyber Security Score (CSS) and Cyber Awareness Score (CAS)

Range for CAS:

Low	**Medium**	**High**
9-20	21-33	34-45

Range for CSS:

Low	**Medium**	**High**
10-22	23-37	38-50

Hypothesis:

Ho: Cyber awareness scores (CAS) and Cyber security scores (CSS) are independent

H1: Cyber awareness scores and Cyber security scores are associated

Contingency table:

CAS / CSS	**Low**	**Medium**	**High**
Low	20	33	2
Medium	19	194	56
High	1	104	193

Result:

```
    Pearson's Chi-squared test

data:  data
X-squared = 219.16, df = 4, p-value < 2.2e-16
```

Decision: As p-value is less than level of significance (α= 0.05), we reject the null hypothesis

Conclusion: Cyber awareness scores and Cyber security scores are associated

Interpretation : People with high cyber awareness scores tend to have high cyber security scores .

Hence we can see a positive association between the attributes .

B) To check independence of Age and Cyber Awareness score

Hypothesis:

Ho: Age and Cyber awareness scores are independent

H1: Age and Cyber awareness scores are associated

Contingency table:

CAS / Age	18-30	31-45	46-60	Above 60
Low	28	12	10	5
Medium	169	42	43	5
High	180	47	57	4

Result:

```
    Pearson's Chi-squared test

data:  data
X-squared = 14.77, df = 6, p-value = 0.02212
```

Decision: As p-value is less than level of significance (α= 0.05), we reject the null hypothesis

Conclusion: Age and Cyber awareness scores are associated

C) To check independence of Age and Cyber Security score

Hypothesis:

Ho: Age and Cyber security scores are independent

H1: Age and Cyber security scores are associated

Contingency table:

CSS / Age	18-30	31-45	46-60	Above 60
Low	18	4	9	3
Medium	200	60	58	7
High	159	34	45	4

Result:

```
    Pearson's Chi-squared test

data:  data
X-squared = 10.948, df = 6, p-value = 0.08998
```

Decision: As p-value is more than level of significance (α= 0.05), we fail to reject the null hypothesis

Conclusion: Age and Cyber security scores may be independent

MANN WHITNEY TEST

A) To check if Cyber Awareness score is same in males and females

Hypothesis:

Ho: There is no significant difference between the median cyber awareness score of male and female

H1: There is significant difference between the median cyber awareness score of male and females

Result:

Wilcoxon rank sum test with continuity correction

data: x1 and x2

W = 38724, p-value = 5.288e-05

alternative hypothesis: true location shift is not equal to 0

Decision: As p-value is less than level of significance (α = 0.05), we reject null hypothesis

Conclusion: Median Cyber awareness score of male and female differ significantly

B) To check if Cyber Security score is same in males and females

Hypothesis:

Ho: There is no difference between the median cybersecurity scores of male and female participants.

H1: There is a difference between the median cybersecurity scores of male and female participants.

Result:

Wilcoxon rank sum test with continuity correction

data: y1 and y2
W = 38858, p-value = 6.835e-05
alternative hypothesis: true location shift is not equal to 0

Decision: As p-value is less than level of significance ($\alpha = 0.05$), we reject null hypothesis

Conclusion: Cyber security scores of male and females differ significantly

IV) KRUSKAL-WALLIS TESTS

The Kruskal-Wallis test is a rank-based nonparametric test that can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable.

IVA) Checking median level of Cyber Awareness Score for different levels of Expertise

Levels of expertise: Beginner, Intermediate, Expert

Beginner- You can go to specific web pages, utilize social media and a few applications such as MS Word

Intermediate- You can load applications, manage settings of devices and has knowledge about hardware as well as software

Expert- You are a Computer specialist, network engineer etc

Hypothesis:

Ho: The median Level of Cyber Awareness score is same across different levels of expertise

H1: The median Level of Cyber Awareness score is different across different levels of expertise

Result:

Kruskal-Wallis rank sum test

data: a by b
Kruskal-Wallis chi-squared = 103.94, df = 2, p-value < 2.2e-16

Decision: As the p-value is less than level of significance (α=0.05), we reject Ho

Conclusion: The median Level of Cyber Awareness score is different across different levels of expertise

B) Checking median level of Cyber Security Score for different levels of Expertise

Levels of expertise: Beginner, Intermediate, Expert

Hypothesis:

Ho: The median Level of Cyber Security score is same across different levels of expertise

H1: The median Level of Cyber Security score is different across different levels of expertise

Result:

Kruskal-Wallis rank sum test

data: y by x
Kruskal-Wallis chi-squared = 70.729, df = 2, p-value = 4.378e-16

Decision: As the p-value is less than level of significance (α=0.05), we reject Ho

Conclusion: The median Level of Cyber Security score is different across different levels of expertise

A) Checking median level of Cyber Awareness Score for different levels of Education

Levels of education: High School, Graduate, Post graduate

Hypothesis:

Ho: The median Level of Cyber Awareness score is same across different levels of education

H1: The median Level of Cyber Awareness score is different across different levels of education

Result:

Kruskal-Wallis rank sum test

data: s by t
Kruskal-Wallis chi-squared = 2.6379, df = 2, p-value = 0.2674

Decision: As the p value is greater than the level of significance (α=0.05), we accept Ho

Conclusion: The median Level of Cyber Awareness score is same across different levels of education

D) Checking median level of Cyber Security Score for different levels of Education

Levels of education: High School, Graduate, Post graduate

Hypothesis:

Ho: The median Level of Cyber Security score is same across different levels of education

H1: The median Level of Cyber Security score is different across different levels of education

Result:

Kruskal-Wallis rank sum test

data: u by v
Kruskal-Wallis chi-squared = 1.4345, df = 2, p-value = 0.4881

Decision: As the p value is greater than the level of significance (α=0.05), we accept Ho

Conclusion: The median Level of Cyber Security Score is same across different levels of Education

V) ODDS RATIO

Odds ratio of people falling victim to cybercrime and gender

```
$data
        Victim Not victim Total
Males      151         198   349
Females    121         156   277
Total      272         354   626

$measure
                        NA
odds ratio with 95% C.I.  estimate     lower    upper
                  Males   1.0000000        NA       NA
                  Females 0.9831927 0.7150175 1.352542
```

Conclusion:

The Odds of falling victim to a cybercrime is 98.31% less for males as compared to females.

Hence, females are more likely to be a victim of cybercrimes.

CONCLUSIONS

1. The **Digital Quality of Life Index** plays a significant role on the **Human Development Index**. Improving the digital quality index is essential for driving economic growth, improving education and skills development. Thus, improving DQL will eventually lead to decrease in the number of cyberattacks.

2. Countries should invest in e- governance and e-infrastructure to improve the Digital Quality of Life index.

3. The median value of the **Global Cyber Security Index (GCI)** varies with **GDP.**

4. There is a significant growth in no. of Cyber Crime incidents in India as the no of incidents in 2023 were 1320106 and the predicted no. of cybercrime incidents for the year 2024 are 1433683.

5. **Financial sector and Health Care sector** are most likely to be a victim of **Phishing, Ransomware and Data Breach** attacks resulting in high monetary losses in these sectors.

6. More Cyber Awareness implies more Cyber Security practices.

7. High Cyber Awareness and high Cyber Security practices are undertaken by the people of the age group 18-30.

8. Cyber Awareness and Cyber Security practices vary for males and females, for people with different levels of Expertise but surprisingly is the same for different levels of Education.

9. Females are more likely to fall victims to Cyber Crimes as compared to males.

GENERAL AWARENESS

Q) How to tackle Cyber Attacks?

There are several approaches one can take to tackle cyberattacks, depending on whether it is an individual or an organization.

For Individuals:

Prevention is Key

- **Strong Passwords**: Create strong and unique passwords for all your accounts. Use a password manager to help you keep track.
- **Software Updates**: Keep your operating system, applications, and firmware updated with the latest security patches.
- **Anti-virus/Anti-malware**: Use a reputable anti-virus and anti-malware software to protect your devices.
- **Wi-Fi Security**: Be cautious when using public Wi-Fi networks. Avoid accessing sensitive information on public Wi-Fi and consider using a VPN.
- **Email and Phishing Awareness**: Don't click on suspicious links or attachments in emails. Be wary of unsolicited emails and messages, even if they appear to be from legitimate sources.
- **Data Backups**: Regularly backup your important data to an external device or cloud storage.

- Defense in Depth:

- **Two-factor Authentication (2FA):** Enable 2FA whenever possible. This adds an extra layer of security by requiring a second verification code in addition to your password.
- **Data Encryption**: Consider encrypting sensitive data on your devices. This can make it more difficult for attackers to access your information even if they breach your defenses.

- Be Prepared to Respond:
- Know what to do: Familiarize yourself with the signs of a cyberattack and have a plan in place for how to respond. This might involve contacting your IT support team, changing your passwords, or notifying the authorities.

For Organizations:

- Comprehensive Security Strategy: Develop a comprehensive cybersecurity strategy that addresses all aspects of your IT infrastructure. This should include:
- Risk Assessment: Regularly assess your security risks and vulnerabilities.
- Security Policies and Procedures: Implement clear security policies and procedures for your employees.

- Employee Training: Provide regular security awareness training to your employees to help them identify and avoid cyber threats.
- Data Security Controls: Implement data security controls such as access controls, data encryption, and activity monitoring.
- Incident Response Plan: Have a plan in place for responding to cyberattacks. This should include procedures for identifying, containing, and recovering from an attack.

- Security Technologies:
- Invest in security technologies such as firewalls, intrusion detection systems (IDS), and data loss prevention (DLP) solutions.
- Stay Informed: Keep up-to-date on the latest cyber threats and vulnerabilities

Portals to report to in case of a Cyberattack

- National Cyber Crime Reporting Portal (NCRP)
- Internet Crime Complaint Centre (IC3)
- Indian Cyber Crime Coordination Centre (I4C)
- Cyber Crime Helpline No.-155260/1930

LIMITATIONS & SCOPE

LIMITATIONS

- The project relied on primary data, which may only represent a small section of the entire population due to selection bias or other factors
- The project faced challenges in collecting data for secondary analysis, such as the losses due to Phishing, Ransomware & Data Breach which could have affected the accuracy of the analysis
- Interpretations of the scores may vary since there were limited questions asked on the topic & more factors could have been considered
- Due to unavailability of Cyber Crime data on Indian websites we were able to draw only limited conclusions
- As fewer cybercrime cases are reported, there is unavailability of the data on victims and losses

SCOPE

- The study can be further expanded by collecting data from a heterogeneous population, and be unbiased towards all groups leading to appropriate conclusions
- The Government & the Education system should spread more awareness about Cyber safety and practices which will further reduce cyber crimes .

APPENDIX

I) SOFTWARE USED

a) R Studio

b) MS Excel

II) RELEVANT CODES

A) Regression Analysis:

```
> d=DQL_index$`DQL Index`
> h=DQL_index$`HDI r`
> #fitting simple linear regression
> fit=lm(h~d)
> summary(fit)

Call:
lm(formula = h ~ d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.201499 -0.038611 -0.005559  0.045262  0.117223

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.40327    0.01884   21.41   <2e-16 ***
d            0.80224    0.03909   20.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06321 on 115 degrees of freedom
Multiple R-squared:  0.7855,    Adjusted R-squared:  0.7836
F-statistic: 421.1 on 1 and 115 DF,  p-value: < 2.2e-16

> #ANOVA for signicance of regression
> summary.aov(fit)
             Df Sum Sq Mean Sq F value Pr(>F)
d             1 1.6824   1.682   421.1 <2e-16 ***
Residuals   115 0.4594   0.004
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Residual analysis plot codes
> par(mfrow=c(2,2))
> plot(fit)
> plot(fit$residuals)
```

B) Mann-Whitney U test:

```
#Male and female cyber awareness score
x1=SCORE$`female cyber awareness score`
x1
x2=SCORE$`male cyber awareness score`
x2
result<-wilcox.test(x1,x2)
result
#
#Wilcoxon rank sum test with continuity correction
#data:  x1 and x2
#W = 38724, p-value = 5.288e-05
#alternative hypothesis: true location shift is not equal to 0
```

C) Two way Analysis of Variance (RBD):

```
> a=Copy_of_phishing$`Financial institution`
> b=Copy_of_phishing$webmail
> c=Copy_of_phishing$`social media`
> d=Copy_of_phishing$Logistics
> e=Copy_of_phishing$Payment
> f=Copy_of_phishing$ecommerce
> d1=data.frame(a,b,c,d,e,f)
> s=stack(d1)
> y1=c(rep(c("2020","2021","2022","2023"),6))
> D1=data.frame(s,y1)
> names(D1)=c("Readings","Treatments","Blocks")
> fit=aov(Readings~Treatments+Blocks,data=D1)
> summary(fit)
            Df Sum Sq Mean Sq F value  Pr(>F)
Treatments   5 1.3532 0.27064   6.615 0.00193 **
Blocks       3 0.7315 0.24384   5.960 0.00695 **
Residuals   15 0.6137 0.04091
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> plot(fit)
> #Post-Hoc
> p1=pairwise.t.test(p,q,p.adjust.method="none")
> p1

        Pairwise comparisons using t tests with pooled SD

data:  p and q

  a      b      c      d      e
b 0.4680 -      -      -      -
c 0.2617 0.6813 -      -      -
d 0.0052 0.0255 0.0588 -      -
e 0.0087 0.0408 0.0909 0.8197 -
f 0.0109 0.0505 0.1106 0.7382 0.9151
```

D) Time Series Analysis:

```
> a=internet_statista$`Number of incidents handled by CERT-IN`
> #Time series pot
> t=ts(a,start=c(2004,1),frequency=1)
> t
Time Series:
Start = 2004
End = 2023
Frequency = 1
 [1]      23     254     552    1237    2565    8266   10315   13301   22060   71780
[11]  130338   49455   50362   53117  208456  394499 1158208 1402809 1391457 1320106
> par(mfrow=c(1,1))
> plot(t,ylab="Number of cases reported",xlab="Year",main="Cyber incidents in India fro
m 2004 to 2023")
> #Single Exp smoothing
> h1=HoltWinters(t,alpha=NULL,beta=F,gamma=F)
> h1
Holt-Winters exponential smoothing without trend and without seasonal component.

Call:
HoltWinters(x = t, alpha = NULL, beta = F, gamma = F)

Smoothing parameters:
 alpha: 0.9999225
 beta : FALSE
 gamma: FALSE

Coefficients:
     [,1]
a 1320112
> plot(h1)
> h1$SSE
[1] 719674401679
> #Double Exp smoothing
> h2=HoltWinters(t,alpha=NULL,beta=NULL,gamma=F)
> h2
```

```
Holt-Winters exponential smoothing with trend and without seasonal component.

Call:
HoltWinters(x = t, alpha = NULL, beta = NULL, gamma = F)

Smoothing parameters:
 alpha: 1
 beta : 0.1622423
 gamma: FALSE

Coefficients:
       [,1]
a 1320106.0
b  113577.2
> plot(h2)
> h2$SSE
[1] 660050827911
> #Forecasting
> library(forecast)
> f=forecast(h2,h=3)
> f
     Point Forecast   Lo 80   Hi 80     Lo 95   Hi 95
2024        1433683 1186402 1680965 1055499.0 1811867
2025        1547260 1168120 1926401  967415.4 2127105
2026        1660838 1159822 2161853  894599.8 2427075
> plot(f)
>
```

E) <u>Chi-Square Test of Independence of Attributes:</u>

```
> # CAS and CSS
> data=matrix(c(20,33,2,19,194,56,1,104,193),ncol=3,byrow=T)
> row.names(data)=c("Low", "Medium", "High")
> colnames(data)=c("Low", "Medium", "High")
> data
       Low Medium High
Low     20     33    2
Medium  19    194   56
High     1    104  193
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 219.16, df = 4, p-value < 2.2e-16
```

F) Odds Ratio:

```
> #Cybercrime and gender
> library(epitools)
> x1=matrix(c(151,198,121,156),ncol=2,byrow = T)
> row.names(x1)=c('Males','Females')
> colnames(x1)=c('Victim','Not victim')
> x1
        Victim Not victim
Males      151         198
Females    121         156
> oddsratio(x1)
$data
        Victim Not victim Total
Males      151         198   349
Females    121         156   277
Total      272         354   626

$measure
                        NA
odds ratio with 95% C.I.   estimate     lower    upper
                  Males    1.0000000        NA       NA
                  Females 0.9831927 0.7150175 1.352542

$p.value
         NA
two-sided midp.exact fisher.exact chi.square
  Males            NA           NA         NA
  Females   0.9169376    0.935421  0.9169712

$correction
[1] FALSE

attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"
>
```

G) Kruskal Wallis Test:

```
> View(Scores_1)
> a=Scores_1$CAS
> b=Scores_1$Expertise
> k1=kruskal.test(a~b,data=Scores_1)
> k1

        Kruskal-Wallis rank sum test

data:  a by b
Kruskal-Wallis chi-squared = 103.94, df = 2, p-value < 2.2e-16
```

III) QUESTIONNAIRE*:*

CYBER SECURITY:

1. Do you use strong, unique passwords for all your accounts?
2. Do you check the authenticity of the website before accessing it?
3. Do you accept cookies while browsing websites?
4. Do you update security tools?
5. Do you regularly update your computer's Operating System?
6. Which security tools do you have downloaded?
7. Do you take proactive steps to protect your devices from Malware & Viruses?
8. Do you encrypt sensitive data stored on your devices or transmitted over the internet?
9. Do you regularly back up important files and data to prevent loss in case of a cyberattack?
10. Do you use two-factor authentication (2FA) whenever possible to enhance the security of your accounts?

CYBER AWARENESS:

1. Do you regularly update passwords?
2. Are you cautious while using public Wi-Fi for sensitive tasks?
3. Can you recognize suspicious emails or messages and do you handle them effectively?
4. Are you familiar with Internet Cookies and their usage while browsing websites?
5. Are you aware of the potential risks and threats associated with downloading pirated content (such as software, movies, music), including malware, legal consequences, and supporting criminal activities?
6. Are you aware of common Cyber Security threats such as Phishing, Malware & Ransomware?
7. Are you confident in your ability to identify phishing emails and avoid falling victim to them?
8. Which of the following Cyber Security portals are you aware of?
9. Do you share your personal data, OTP, bank details with anyone?

REFERENCES

- https://surfshark.com/dql2023/insights
- https://www.itu.int/en/ITU-D/Cybersecurity/Pages/global-cybersecurity-index.aspx
- https://ncsi.ega.ee/ncsi-index/
- https://www.ibm.com/reports/data-breach
- https://www.statista.com/statistics/617136/digital-population-worldwide/
- https://apwg.org/trendsreports/
- https://www.stationx.net/
- surfshark.com/research/data-breach-monitoring
- https://www.coveware.com/ransomware-quarterly-reports
- https://data.gov.in/search?title=crime
- https://www.ic3.gov/Home/AnnualReports
- https://www.cyfirma.com/outofband/tracking-ransomware-december-2023/
- https://www.sonicwall.com/threat-report/
- https://cybercrime.gov.in/