TODO*

TODO

Shivank Goel

April 10, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/shivankgoel003/DataBreach_Ransomware_Stats

1 Introduction

In today's digital world, every click, every online transaction, and every shared piece of data poses an entry point for cyber threats. Cybercrime is not just growing quickly, rather it is turning out to be more of a serious concern with both the frequency and severity of attacks rising every year. The purpose of these attacks is to harm companies and organizations financially, however, in some cases these attacks can have military or political purposes. "According to a report published by the Identity Theft Resource Center (ITRC), a record number of 1862 data breaches occurred in 2021 in the US. Sectors like healthcare, finance, business, and retail are the most commonly attacked, impacting millions of Americans every year" (UpGuard 2024)

Despite such severe threats and impacts of cyberattacks, still certain companies tend to oversee this concern. As per PWC 2024 Global Digital Trust Insights report, "about one-third of organisations have no risk management plan to address cloud service provider challenges. Half are 'very satisfied' with their technology capabilities in key cybersecurity areas. More than 30% of companies don't consistently follow what should be standard practices of cyber defence." (PwC 2023)

Therefore, in response to such attacks, there is a need for a plan, that not just keep the intruder or hackers out but also quickly alert if an attack does happen. Our study looks at cyber resilience, which is " the ability to anticipate, withstand, recover from, and adapt to adverse conditions, stresses, attacks, or compromises on systems that use or are enabled by cyber resources." as defined by National Institute of Standards and Technology (National Institute of Standards and Technology (NIST) 2023). We break our study into three major research questions, which forms the thesis of our study:

RQ1: How do organizational factors such as size and sector, influence the severity of cyber breaches experienced by companies?

RQ2: Which cybersecurity strategies, including frameworks, policies, and preventive measures, are the most effective one, in reducing the damage caused by cyber attacks?

RQ3: In what ways do the industry type and digital dependence of a business affect the overall impact of a cyber attack, in terms of preserving confidentiality, integrity and availability of data?

The estimand of our study is the measurable effect of specific characteristics of an organization including size, sector and digital intensity on their cyber resilience. As a key finding, our regression models reveal factors such as organizational size, sector, and digital intensity significantly influence an organization's cyber resilience. For example, larger companies often have stronger defenses against cyber attacks, and on the other hand, companies that use a lot of digital technology in their work have different levels of protection.

We aim to study and answer these questions by performing an analysis on a dataset of data breaches and ransomware attacks over 14 years from 2004, published by the University of Queensland.

The remainder of this paper is structured as follows: We provide the background and overview of the study in Section 2. Section 3 provides an overview of our methodology, including the data collection process and the analytical techniques used to explore the dataset of cyber attacks. Section 4 presents the regression models, discussing how we applied these models to understand the impact of various factors like organizational size, sector, and digital intensity on cyber resilience, Section 5 displays the interpretations of the model alongside other findings from analyzing the data, and Section 6 provides a discussion on the implications of the findings as well as the weaknesses of this paper and its next steps for further study on this subject.

2 Background

As discussed earlier, cyber resilience is about an organization's ability to keep its operations running smoothly in the face of cyber threats. It is not just about preventing cyber attacks, but also being prepared to deal with them effectively when they do happen. It is about recovery and adaptation, and extends beyond traditional cyber security measures. Cyber resilience surrounds various elements:

1. Governance: This is the structure and processes that define the organization's approach to cyber threats. It is about leadership, accountability, and ensuring that the policies are in place and followed as desired.An effective governance is characterized by use of well defined frameworks, and presence of dedicated cybersecurity roles.

- The use of well-defined frameworks that guide the organization's cybersecurity protocols.
- The presence of dedicated cybersecurity roles such as a Chief Information Security Officer can prevent damage to IT systems and network.

2. Prevention, Detection, and Recovery: These are the specific controls and strategies used to prevent attacks, detect them promptly, and recover from any damage caused. This approach involves:

- Setting up appropriate remote access controls to secure unauthorized access.
- Implementing proper network segmentation to control traffic flow and prevent the spread of threats within networks.
- Adding an encryption to protect confidential data.
- Utilizing detection systems to identify potential threats.
- Developing restructuring plans as part of recovery measures to restore systems after the attack.

3. Learning and Adapting: An organization needs to continuously learn from past incidents and attacks. It must adapt its strategies accordingly. This could involve updating its policies, training employees, and revising its approach to security.

4. External Factors: Factors like the industry the organization is in, its size, and its digital intensity (how much it relies on digital technology) can also impact its cyber resilience.

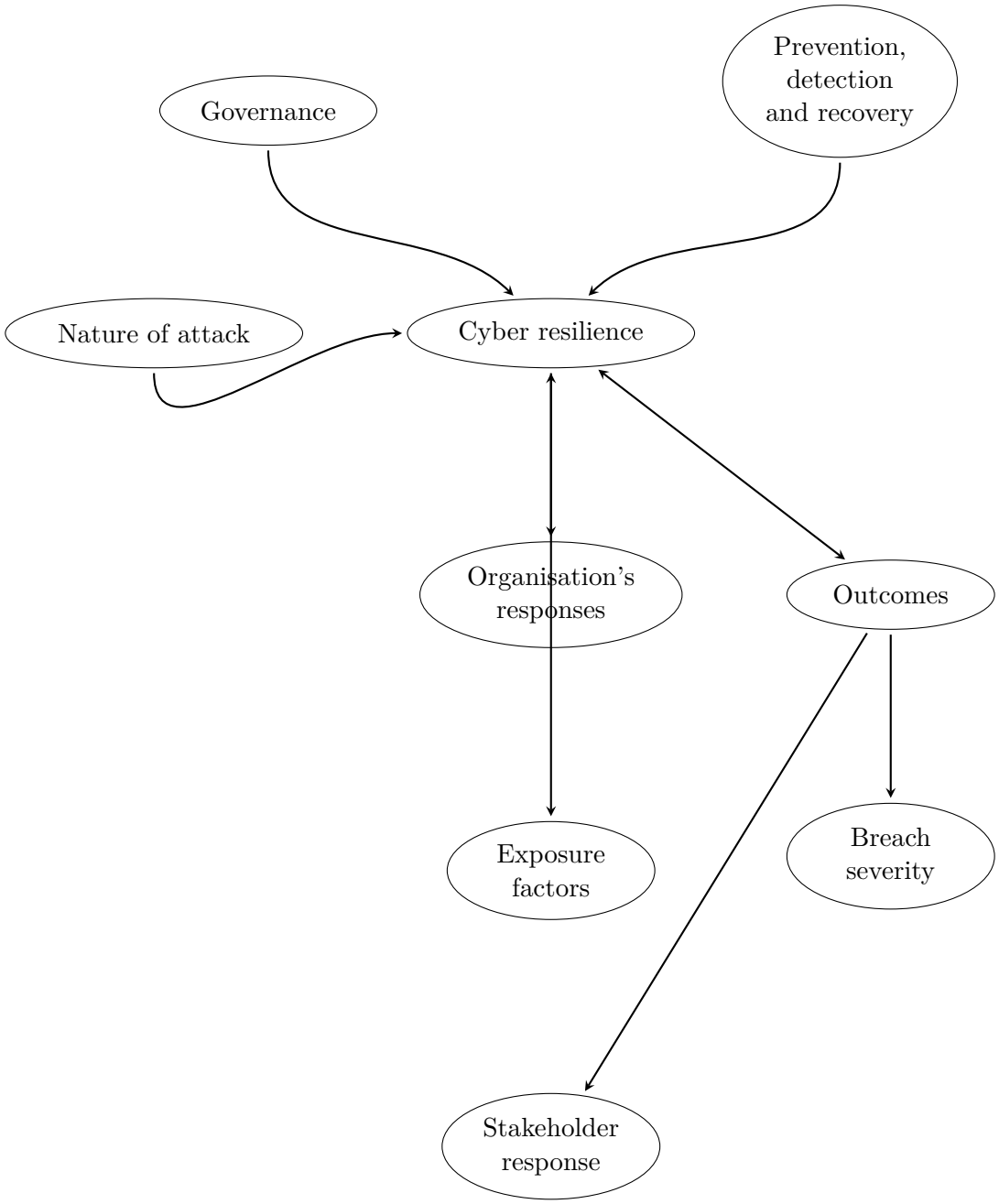

Figure 1: Conceptual model of organizational cyber resilience

3 Data

3.1 Data Source and Collection:

Our analysis is based on sampling of 514 data breaches and ransomware attacks spanning over 14 years from 2004 to 2019. The dataset was obtained from the website of University of Queensland, and was prepared and compiled by researchers Tsen, Elinor, Ko, Ryan, and Slapnicar, Sergeja (Tsen, Ko, and Slapnicar 2020) at the University of Queensland. The data is thorough and encompasses a wide range of cyber attack incidents. It offers insights into various aspects of these incidents, including the types of breaches, affected organizations, and the extent of impact.

The dataset represents a detailed aggregation of data breaches and ransomware attacks, and as per authors, it was originally sourced from publicly disclosed media reports. The data integrates information from multiple public databases such as Privacy Rights Clearinghouse, Information is Beautiful, the Repository of Industrial Security Incidents, and Carnegie Mellon's list of Banking Cyber Incidents. These sources were chosen for their public accessibility and frequent citation in academic and industry literature. The approach to data collection was guided by the PRISMA methodology which ensured a systematic and thorough compilation process.

3.2 Data Cleaning

We used R (R Core Team 2023) and Wickham et al. (2019a) for data cleaning and processing, utilizing packages like tidyverse (Wickham et al. 2019b) for data manipulation and janitor (Firke 2023) for cleaning column names. Other packages used includes `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2024), `tibble` (Müller and Wickham 2023), `janitor` (Firke 2023),`reshape2` (Wickham 2007), `knitr` (Xie 2023), `ggbeeswarm` (Clarke, Sherrill-Mix, and Dawson 2023), `ggrepel` (Slowikowski 2024), `kableExtra`(Zhu 2024), `readxl`(Wickham and Bryan 2023), `MASS`(Venables and Ripley 2002), `rstanarm`(Goodrich et al. 2022), `modelsummary`(Arel-Bundock 2022) and `here` (Müller 2020).

The cyber breach data was preprocessed to remove inconsistencies and irrelevant information. Firstly, variable names were simplified and standardized for consistency and ease of analysis. A key challenge faced was the significant number of missing values in the 'number of users affected' column. This variable was central to our study as we aimed to study trends related to the scale of impact using linear regression analysis. To address this issue, a choice was made to exclude records with missing or uncertain values in this column. While this decision resulted in some data loss, it was a necessary measure to maintain the integrity and accuracy of our trend analysis. Also, in columns like 'attack_type' and 'organisation_size', missing values were replaced with "Unknown" to maintain data integrity.

3.3 Measurement and Exploratory Data Analysis

As part of the measurement, we converted real-world cyber incidents into quantifiable data within our dataset. The dataset variables were defined and measured based on the nature of the cyber incidents they represent. Each entry in the dataset corresponds to a distinct cyber incident, with variables relating to the incident. Here is how we defined and measured key variables:

- `organisation_size`: This categorical variable categorizes the size of the affected organization into 'Small', 'Medium', 'Large', or 'Unknown', based on the number of employees or annual revenue as per commonly accepted business standards.
- `sector`: The sector to which the affected organization belongs is classified according to standard industry classifications. This ensures each entry aligns with the appropriate economic sector.
- `cyber_security_role`: This binary variable indicates the presence (`Yes`) or absence (`No`) of a dedicated cybersecurity role within the organization, at the time of the incident.
- `number_of_users_affected`: Represented as a numerical variable, this measures the estimated number of individuals whose data was compromised during the breach
- `undertook_investigation`: It captures whether an investigation was initiated following the cyber incident (1 for Yes, 0 for No).
- `breach_severity`: To highlight the complexity and impact of cyber incidents, we introduced a custom variable, `breach_severity`. This variable was constructed to study the nature of cyber breaches, combining several key aspects of an incident:
- `impact_on_data`: This reflects the nature of data compromise during the breach (categorized as 'High', 'Medium', or 'Low').
- `subsequent_fraudulent_use_of_data`: Considers if the breached data was later used for fraudulent activities.

The `breach_severity` variable was formulated through a custom function in our data processing script, which combined these elements to classify each incident into 'High', 'Medium', or 'Low' severity categories. This classification was based on the overall impact, the nature of data compromised, and the extent of misuse of data. This measure provides an understanding of the impact of each breach, beyond the simple binary or categorical measures commonly used.

To achieve a clear understanding of the data, we include a variety of graphs and tables that represent the characteristics of each variable within our dataset. These visualizations illustrate the distribution and relationships among key variables, offering a broad picture of the patterns and trends among our data. Table 1 shows the summary statistics for organization

size and sector. For each category within these variables, we present the count and the relative frequency, expressed as a percentage of the total sample. The frequency distribution of variables such as organisation_size and sector indicates the diversity of the dataset showing various sizes of organizations and a range of sectors. Table 2 shows the decripitive statistics for certain variables.

CS Role Yes (27.27%): This shows that approximately 27% of the organizations in the dataset have a designated Cyber Security (CS) role.

CS Role No (72.73%): Conversely, nearly 73% of organizations do not have a designated CS role.

Framework Yes (36.36%): About 36% of the organizations adhere to a cyber security framework. Such frameworks provide structured guidelines and best practices for managing cyber security risks.

Framework No (63.64%): The majority, approximately 64%, do not follow a specific cyber security framework.

Prevention Low (45%): 45% of the organizations were categorized as having Low prevention measures, indicating basic or minimal preventive security measures.

Prevention Medium (36.36%): 36.36% fell into the Medium prevention category, suggesting more substantial but not so strong security measures.

Prevention High (18.18%): Only 18.18% were classified under High prevention, reflecting strong preventive strategies against cyber threats.

Table 2: Descriptive Statistics for Cyber Security Variables

Variable	**Frequency (%)**
a. Governance (N = 514)	
CS Role Yes	27.27
CS Role No	72.73
b. Cyber Security Frameworks (N = 514)	
Framework Yes	36.36
Framework No	63.64
c. Prevention, Detection and Recovery	
Prevention Low	45.45
Prevention Medium	36.36
Prevention High	18.18

In order to observe the trend of number of cyberattacks over a span of years, from 2004 to 2019, we plotted a line graph Figure 2. It is evident from the plot that the frequency of attacks has

Table 1

	Summary Statistics		
Variable	Category	Count	Frequency....
Organization Size			
a. Organization Size	Large	329	64.13
a. Organization Size	Unknown	83	16.18
a. Organization Size	Medium	66	12.87
a. Organization Size	Small	35	6.82
Sector			
b. Sector	Human health activities	191	37.23
b. Sector	Education	65	12.67
b. Sector	Finance and insurance	55	10.72
b. Sector	Arts, entertainment and recreation	37	7.21
b. Sector	Public administration and defence	33	6.43
b. Sector	IT and other information services	24	4.68
b. Sector	Wholesale, retail trade and repair	21	4.09
b. Sector	Advertising and other business services	13	2.53
b. Sector	Accommodation and food service activities	10	1.95
b. Sector	Residential care and social work activities	7	1.36
b. Sector	Telecommunications	7	1.36
b. Sector	Computer, electronic and optical products	6	1.17
b. Sector	Machine equipment	6	1.17
b. Sector	Textiles, wearing apparel and leather	6	1.17
b. Sector	Publishing, audiovisual and broadcasting	5	0.97
b. Sector	Transportation storage	5	0.97
b. Sector	Food products, beverages and tobacco	4	0.78
b. Sector	Scientific research and development	4	0.78
b. Sector	Legal and accounting activities	3	0.58
b. Sector	Administrative and support service	2	0.39
b. Sector	Chemicals and chemical products	2	0.39
b. Sector	Construction	2	0.39
b. Sector	Electricity, gas, steam and air conditioning	2	0.39
b. Sector	Pharmaceutical products	2	0.39
b. Sector	Electrical equipment	1	0.19

Summary Statistics for Organization Size and Sector

fluctuated over the years, with a peak in 2017. However, the decline following this peak may indicate the impact of improved cybersecurity measures, or a possible transition to different types of cyber threats not captured in this dataset. This visualization provides an overview of the nature of cyber threats and the ongoing battle between cybersecurity efforts and threat actors.

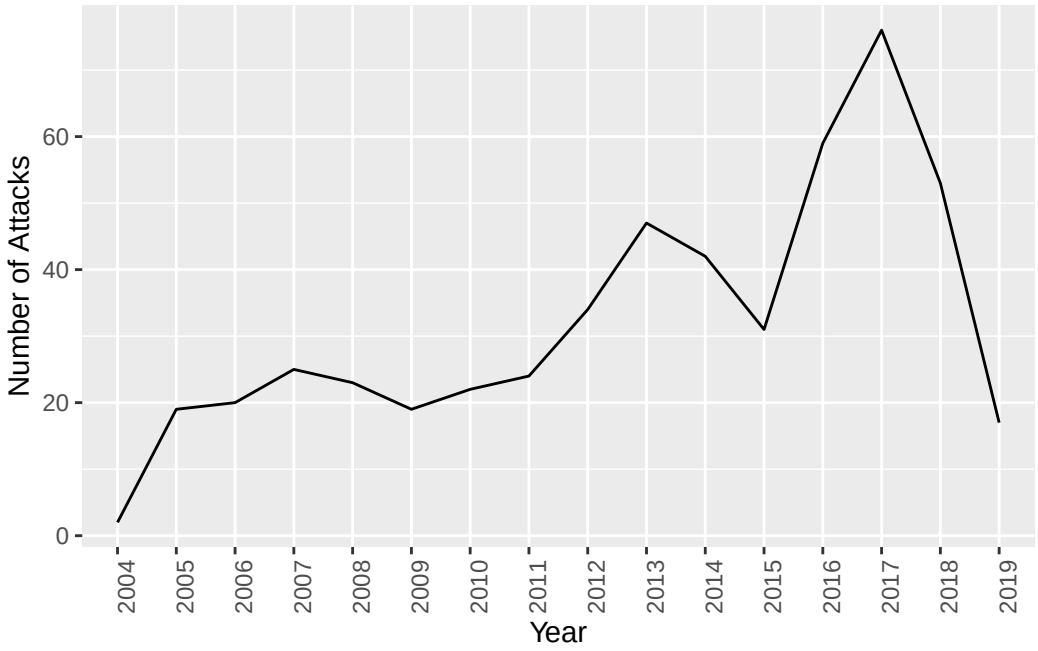

Figure 2: Cyberattacks Over Time

We also plotted a bar graph Figure 3 to count the number of incidents across various sectors. The bar chart clearly indicates that the 'Human Health Activities' sector has the highest count of incidents, standing out significantly from the other sectors. This might suggest that health sector is a more frequent target for cyber incidents or probably it is more diligent in reporting such events. The other sectors show a range of incident counts, with most appearing to have far fewer incidents in comparison. This could point to different levels of risk exposure, varying security measures, or reporting practices across these sectors.

Figure 4 is a creative visualization that effectively depicts the distribution and comparison of cyber attacks across various countries, with a specific emphasis on the United States. It combines a stacked bar chart for multiple countries and a line plot for the USA allowing for a dual-axis comparison due to the disproportionate number of attacks in the USA compared to other countries.

The bar segments represent the frequency of attacks in countries such as Australia, Canada, Japan, the UK, and others, with each color corresponding to a different country. The stacked

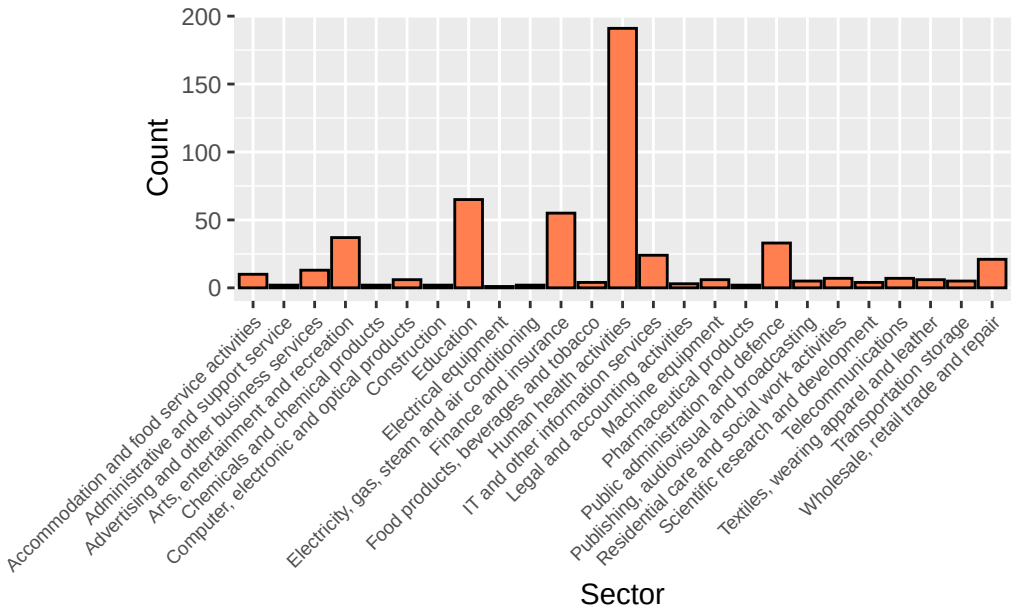

Figure 3: Bar Plot of Count of Cyber Attacks by Sector

nature of the bars shows how the total number of attacks is divided among these countries within each year.

The line plot, on the other hand, tracks the frequency of cyber attacks in the USA across the same timeframe, adjusted by a scale factor for direct comparison on a secondary y-axis. This representation highlights the stark contrast in the volume of attacks between the USA and other countries while providing a clear year-by-year trend analysis.

The choice to categorize all countries with fewer attacks under a consolidated "Other" category is a practical approach to maintain clarity in the visualization, avoiding overcrowding the chart with too many individual country representations.

Figure 5 represents a stacked area chart, with each colored layer representing a different type of attack, allowing for an easy comparison of their occurrences over time. It is clear that some attack types, like installed malware, show peaks and troughs, possibly depicting the nature of cyber threats and security measures. These trends can be helpful for understanding the changing landscape of cyber risks and preparing for future security strategies.

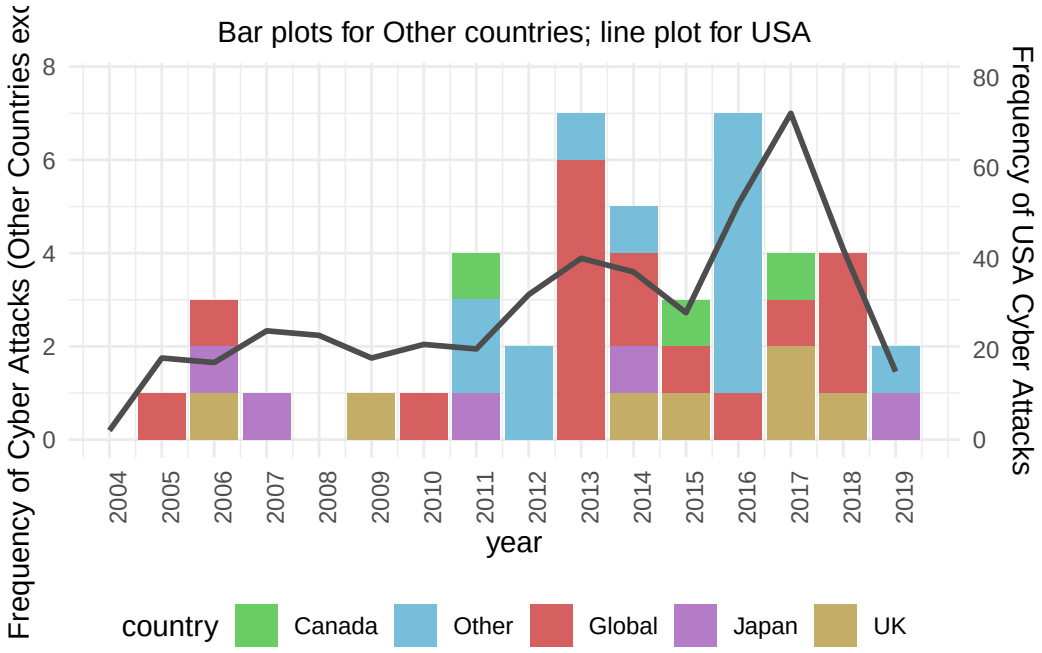

Figure 4: Overview of Cyber Attacks by Year and Country

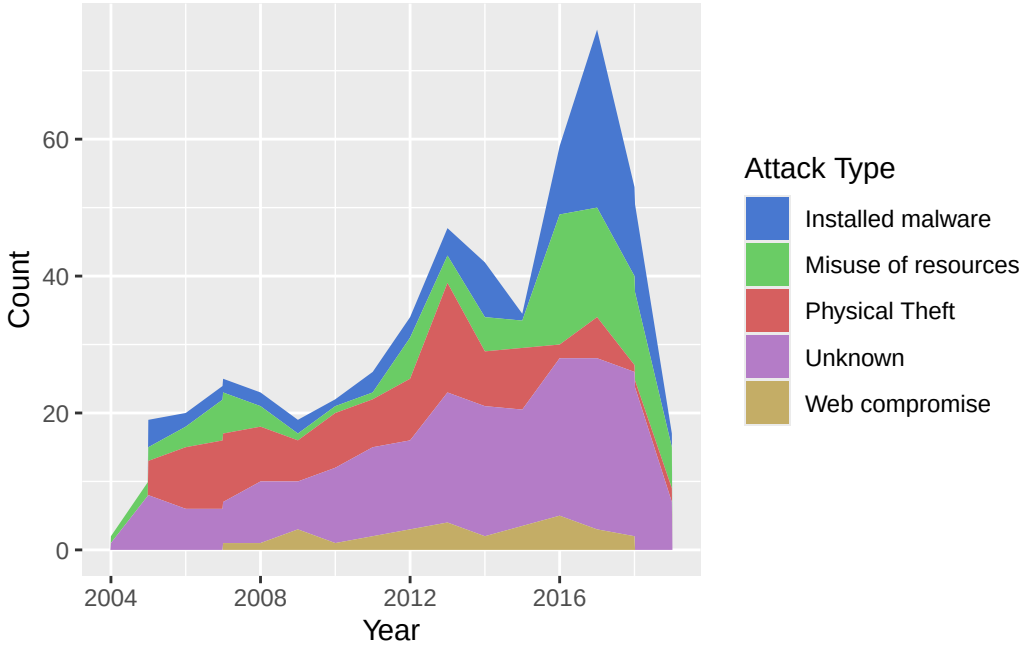

Figure 5: Attack Types Over Years

4 Model

4.1 Model Set-up

After performing exploratory data analysis, we got a broad overview of the general trend between different variables of our dataset. To strengthen answers to our research question claims made earlier, we implement regression models and take a close look at what makes an organization good at dealing with cyber threats. We use these models to understand how certain factors, like the size of a company or the industry, affect its ability to handle a cyberattack.

Simple Models for Numbers (Linear Regression):

We were interested in knowing how often cyberattacks happen and how bad they are based on the company's size or the kind of work it does. We use linear regression for this. The basic form of a linear regression equation is

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$

Where: - Y is the outcome we're interested in.

- β_0 is the intercept, the starting point of our equation when all our predictors (Xs) are zero.
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients. These tell us how much change we expect in Y for a one-unit change in each predictor X, assuming all other predictors are held constant.
- $X_1, X_2, ..., X_n$ are the predictor variables.
- ε represents the error term. It accounts for the difference between our predicted value of Y and its actual value.

We expected to study if bigger companies, for example, have more cyberattacks or not. In form of equation we can represent it as follows:

Impact = (Base Impact) + (Effect of Company Size) + (Effect of Industry) + ...

Complex Models for Yes/No Outcomes (Bayesian Logistic Regression):

Now, what if we want to know if a company will take action after an attack, like starting an investigation? This is a yes/no type of question. We use a different method called logistic regression, which is good for when the outcome is not a number but a choice between two things.

4.2 Specific models for Research Questions

RQ1: Organizational attributes and cyberattack handling

To address the first research question about how an organization's size and business type affect its ability to handle cyber attacks, we developed a specific linear regression model. This model examines the relationships between various organizational attributes and the handling of cyberattacks. It predicts the severity of cyber breaches based on:

1. Organizational Size (Medium, Small, Unknown)
2. Digital Intensity (Low, Low-Medium, Medium-High)
3. Sector (e.g., Public Administration and Defence, Education, Health)
4. Country

- Dependent Variable: Severity of cyber breaches.
- Independent Variables: Organizational size, sector, digital intensity, and geographical location.

Table 3 summaries the results from model.

1. Organisation Size: The coefficients for different organization sizes (Medium, Small, Unknown) indicate how the severity of breaches changes with these categories compared to the baseline category.As compared to a baseline (Large), Medium (0.015) and Small (0.103) sizes have positive coefficients, suggesting potentially higher breach severity. Unknown (-0.072) indicates lower severity.
2. Level of Digital Intensity & Sector: These coefficients compare each level and sector against baselines, "Medium-High" for digital intensity and "Finance and Insurance" for the sector. The public administration and defence sector (-0.581) shows a lower severity.
3. Country: Coefficients for countries are relative to the USA (reference level). They indicate variations in breach severity across geographical locations.

R-squared (0.137): It indicates the proportion of variance in the dependent variable that's predictable from the independent variables. A value of 0.137 suggests that approximately 13.71% of the variation in breach severity is explained by the model.

Adjusted R-squared (0.099): This value is slightly lower than the R-squared, and accounts for the number of predictors. It suggests a good fit of the model considering the complexity introduced by multiple variables.

F-statistic of 3.619 and a p-value of 1.12e-06, suggests that the model demonstrates the statistical significance overall, indicating that at least some of the predictors are likely to be genuinely associated with variations in breach severity.

Table 3: Linear regression analysis with breach severity numeric as dependent variable

Predictor	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.08715	0.09139	22.837	< 2e-16 ***
Organisation Size - Medium	0.01467	0.08055	0.182	0.855
Organisation Size - Small	0.10336	0.10093	1.024	0.306
Organisation Size - Unknown	-0.07151	0.07166	-0.998	0.319
Level of Digital Intensity - Low	-0.29052	0.22750	-1.277	0.202
Level of Digital Intensity - Low-Medium	0.11865	0.16894	0.702	0.483
Level of Digital Intensity - Medium-High	0.45764	0.15186	3.013	0.0027 **
Sector - Public Administration and Defence	-0.58112	0.20077	-2.894	0.0040 **
Sector - Other	-0.44159	0.12026	-3.672	0.0003 ***
Sector - Arts, Entertainment and Recreation	-0.80423	0.20009	-4.019	6.94e-05 ***
Sector - Education	-0.27055	0.20130	-1.344	0.180
Sector - Human Health Activities	-0.20620	0.19467	-1.059	0.290
Country - Canada	-0.55895	0.30172	-1.853	0.0646 .
Country - Global	-0.41952	0.16428	-2.554	0.0110 *
Country - Japan	-0.21109	0.26805	-0.788	0.431
Country - UK	0.07899	0.30459	0.259	0.795
Country - Other	0.13767	0.15593	0.883	0.378
Country - Singapore	-0.68348	0.30035	-2.276	0.0234 *
Country - South Korea	-0.74056	0.52226	-1.418	0.157
Residual standard error	0.5131 on 410 DF			
Multiple R-squared	0.1371			
Adjusted R-squared	0.09921			
F-statistic	3.619 on 18 and 410 DF			
P-value	1.12e-06			

RQ2: Effectiveness of cybersecurity methods

To answer our second research question regarding the most effective strategies and methodsused by organizations to reduce the damage from cyberattacks, we constructed a linear regression model. This model shows the number of users impacted by cyber incidents, considering a relationship between cybersecurity strategies and organizational attributes :

1. Implementation of Cybersecurity Frameworks
2. Adoption of Education and Awareness Policies
3. Execution of Prevention, Detection, and Recovery Measures
4. Organizational Size (Medium, Large, Unknown)
5. Sector Specificity
6. Level of Digital Intensity

- Dependent Variable: Number of users affected by cyberattacks.
- Independent Variables: Cybersecurity measures, organizational size, sector, and digital intensity level.

Table 4 depicts the results derived from the model.

1. Cybersecurity Frameworks(-13,994,198) - The negative coefficient implies that cybersecurity frameworks might reduce the number of users affected.
2. Education and Awareness Policies (38,460,000): A positive coefficient suggests a correlation between the implementation of these policies and an increase in the number of users affected. However, the insignificance of this result (p = 0.843) implies caution in interpreting this finding.
3. Organizational size - Medium size (-7,347,889, p = 0.804), Large size (8,946,092, p = 0.728), and Unknown size (-14,379,354, p = 0.615): The model suggests no significant effect of organization size on the number of users affected, as all the p-values greater than 0.05.
4. Sector Specifics: The model reveals significant positive relationships for 'Other' sectors (73,186,958, p = 0.004) and 'Arts, Entertainment, and Recreation' (93,277,310, p = 0.024), indicating these areas are likely to witness a higher number of users affected by cyberattacks.
5. Digital Intensity Level (-28,344,216): A negative coefficient with p = 0.003 shows that increased digital intensity is associated with a greater number of users affected.

The model's R-squared (0.093) signifies that about 9.3% of the variance in the number of users affected is explained by the model, while the adjusted R-squared (0.037) factors in the number of predictors.

An F-statistic of 1.648 along with a p-value of 0.018 shows the model's overall significance, indicating that specific factors, particularly sector and digital intensity level, are helpful in understanding the scale of cyberattack impacts.

Table 4: Linear regression analysis with number of users affected as dependent variable

Model	Coefficients	Standard errors	T value	Sig
Intercept	36,687,660	32,842,060	1.117	0.264
Cyber Security Role	NA	NA	NA	NA
Cyber Security Frameworks	-13,994,198	139,505,210	-0.100	0.920
Education and Awareness Policy	38,460,000	194,290,149	0.198	0.843
Prevention, Detection, and Recovery	NA	NA	NA	NA
Organization Size - Medium	-7,347,889	29,571,997	-0.248	0.804

Continued on next page

Table 4 – *Continued from previous page*

Model	Coefficients	Standard errors	T value	Sig
Organization Size - Large	8,946,092	25,705,975	0.348	0.728
Organization Size - Unknown	-14,379,354	28,602,186	-0.503	0.615
Sector - Public Administration and Defence	81,232,193	41,730,689	1.947	0.052
Sector - Other	73,186,958	25,457,748	2.875	0.004
Sector - Arts, Entertainment and Recreation	93,277,310	41,274,076	2.260	0.024
Sector - Education	40,468,643	31,559,658	1.282	0.200
Sector - Human Health Activities	47,359,281	28,602,617	1.656	0.098
Level of Digital Intensity	-28,344,216	9,431,846	-3.005	0.003
R Squared		0.093		
Adjusted R Squared		0.037		
F Statistic		1.648		
Significance of F		0.018		

Note: NA indicates that the variable was not applicable or not included in the model.

Also, the trend of our dependent variable `number_of_users_affcted` in RQ2 model, can be seen in Figure 6

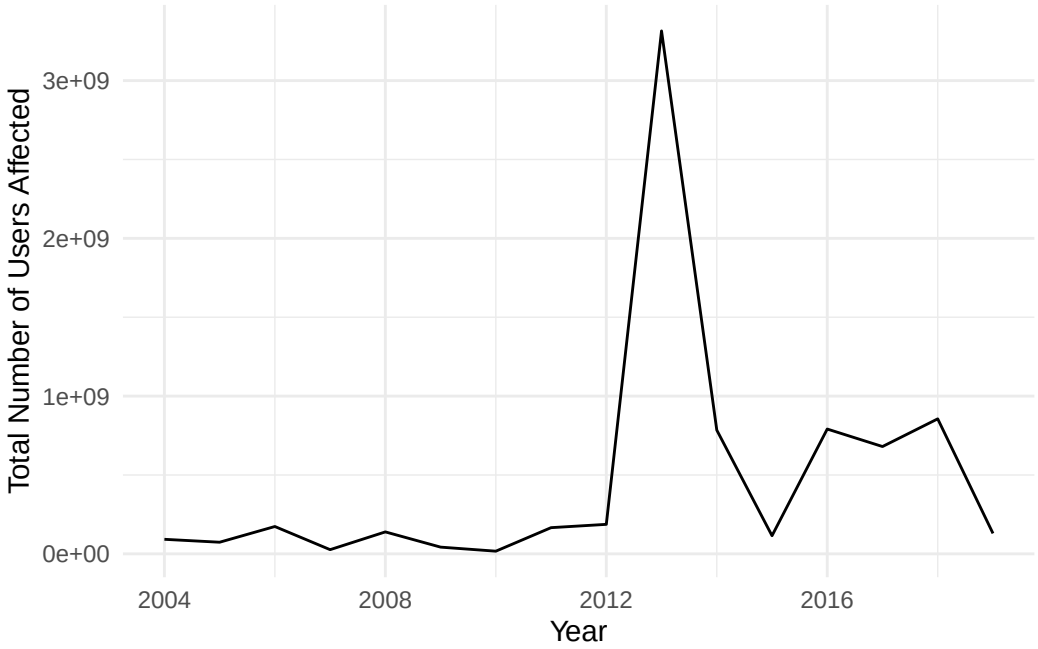

Figure 6: Number of Users Affected Over Years16

RQ3: Impact of Organizational Attributes on Cyberattack Consequences

To analyze how specific attributes of a business, like its industry or dependency on digital tools, impact the consequences of cyberattacks, we used a multinomial logistic regression model. This model helps to determine the relationship between organizational characteristics and the level of impact on data, classified as 'Low', 'Medium', or 'High'.

1. Organizational Size (Medium, Small, Unknown) – with 'Large' as the reference category.
2. Digital Intensity (Low, Low-Medium, Medium-High) – with 'High' as the reference category.
3. Sector – 'Public Administration and Defence', 'Other', 'Arts, Entertainment and Recreation', etc., with 'Finance and Insurance' as the reference category.
4. Country – including Canada, Global, Japan, etc., with 'USA' as the reference category.
5. Dependent Variable: The level of impact on data (Low, Medium, High).

- Independent Variables: Organizational size, digital intensity, sector, country.

Table 5 displays the model outcomes.

1. Organizational Size: Medium (-0.793, p = Medium) and Small (0.308, p = Small) sizes show different degrees of impact compared to the baseline 'Large' size, suggesting the size of the organization plays a role in the impact level of cyberattacks.

2. Level of Digital Intensity: A higher digital intensity has a correlation with a greater likelihood of 'High' impact breaches, as indicated by the positive coefficients for 'Low-Medium' (2.037) and 'Medium-High' (3.073) intensity levels.

3. Sector Specifics: Significant variations are observed across sectors. For example, 'Public Administration and Defence' (-1.907, p = Public Administration) and 'Arts, Entertainment and Recreation' (-4.116, p = Arts) sectors show a specific pattern in their vulnerability to cyberattacks.

4. Geographical Location: The impact of cyberattacks also differs based on geographical location. Canada (-14.948, p = Canada) and Singapore (-15.594, p = Singapore) shows a significant differences in the 'High' impact category, depiciting regional variations in cyberattack severity.

The coefficients represent log odds ratios, where a positive value implies a higher likelihood of the respective impact level compared to the baseline ('High' impact).

Table 5: Multinomial logistic regression analysis with impact on data as dependent variable

Predictor	Coefficients (Medium)	Coefficients (High)	Std. Errors (Medium)	Std. Errors (High)
Intercept	1.305595	1.442978	0.4671905	0.4506995
Organisation Size - Medium	-0.7934138	0.3656129	0.3944130	0.3923776
Organisation Size - Small	0.308093	2.674162	0.8033055	0.7916091

Continued on next page

Table 5 – *Continued from previous page*

Predictor	Coefficients (Medium)	Coefficients (High)	Std. Errors (Medium)	Std. Errors (High)
Organisation Size - Unknown	-0.8242923	0.1584841	0.3454811	0.3649247
Level of Digital Intensity - Low	0.6370277	1.6435254	0.7712827	0.6704278
Level of Digital Intensity - Low-Medium	2.037751	1.534643	0.8804760	0.8738375
Level of Digital Intensity - Medium-High	2.085403	3.073733	0.8548181	0.8186515
Sector - Public Administration and Defence	-1.907019	-5.743329	1.076650	1.186423
Sector - Other	-1.499581	-1.889906	0.5814518	0.5479577
Sector - Arts, Entertainment and Recreation	-4.116223	-5.118463	1.085765	1.012404
Sector - Education	-1.795158	-4.288260	1.047509	1.174420
Sector - Human Health Activities	-1.850178	-3.236125	1.007041	1.004086
Country - Canada	0.03118258	-14.94830280	1.335552e+00	1.259128e-05
Country - Global	-0.7303029	-0.9285532	0.7006000	0.6770341
Country - Japan	0.8169932	-0.8046899	1.265776	1.554310
Country - UK	-0.6114617	0.2831945	1.500414	1.192295
Country - Other	-0.9633399	0.2562872	0.7914678	0.6999920
Country - Singapore	-0.7333256	-15.5943691	1.251740e+00	3.353692e-06
Country - South Korea	-12.880809	-1.007852	488.290745	1.410714

5 Results

5.1 Statistical Results

In this section, we discuss the outcomes from our statistical analysis about cyberattack impacts on organizations and model results. The findings are discussed across three research questions (RQs), where each question aimed at understanding different aspects of cyberattack consequences and responses.

Summary Statistics : Most of the organizations we looked at are large ones as they make up about 64% of our data. This could mean that big companies are more often the targets of cyberattacks, or maybe the big institutes just report these incidents more because they are more noticeable and rely a lot on computers and the internet. We also see that hospitals and schools get attacked a lot. Since they have a lot of sensitive information, hackers target them more.

Our result matches with earlier discussed PWC 2024 Global Digital Trust Insights report that more than 30% of companies don't consistently follow what should be standard practices of cyber defence. Based on our data we also found that only about 27% of these organizations

have someone in charge of keeping their computers safe, and only about 36% follow a set of rules to protect themselves online. This makes us think that maybe not all companies are ready to defend themselves against cyberattacks.

When we look at how they try to stop attacks, almost half of them are only doing the very basic things. This tells us that many organizations might need to do more to keep their data safe.

Cyberattacks Over Time : The surge in cyberattacks observed in 2017 Figure 2 might be linked to various factors, including an increase in attack surfaces, the prevalence of more sophisticated attack methods, or a better reporting system that has led to higher incidents. This upward trend could signal the need for enhanced security measures and more proactive threat detection systems.

Sector and Country Analysis: The disparity in the frequency of attacks among sectors Figure 3 suggests different levels of cyber hygiene or different attacker motivations, such as financial gain or data theft, which are particularly high in sectors dealing with personal data. The dominance of attacks in the USA, alongside high frequencies in other countries too Figure 4, might reflect not only the attractive target that the USA presents but also the global disparity of cyber threats.

Attack Types: The attack types distribution Figure 5 indicates a strong nature of cyber threats, with installed malware and web compromises becoming increasingly common. This pattern could be might indicate that attackers have increased the use of more sophisticated tools and thus, there is a need for businesses to focus on securing web interfaces and being cautious against malwares.

5.2 Model Results

RQ1: Organizational Attributes and Cyberattack Handling

The model about how organizational size and sector influence cyberattack handling shows significant sector-specific effects. Public administration and defence (-0.581, $p < 0.005$), Other (-0.441, $p < 0.001$), and Arts, Entertainment and Recreation (-0.804, $p < 0.001$) sectors show a lower breach severity score, which indicates a better handling or lower impact of cyberattacks compared to the reference sector, Finance and Insurance. Conversely, the effect of organizational size is not statistically significant, further telling us that size alone does not reflect an organization's capacity to handle cyberattacks effectively.

RQ2: Cybersecurity Methods Effectiveness

Regarding the effectiveness of cybersecurity methods, the model reveals a complex picture. The negative coefficient for Cyber Security Frameworks (-13,994,198) suggests that such frameworks might reduce the number of users affected. However, the positive coefficient for Education and Awareness Policy (38,460,000), though not statistically significant ($p = 0.843$), may

require a cautious interpretation, as it suggests a possible correlation with an increase in the number of users affected.

RQ3: Specific Organizational Contexts

For RQ3, the multinomial logistic regression model depicts the variable impact levels of cyber-attacks across different organization sizes, sectors, and countries. Precisely, sectors such as Public Administration and Defence (coefficients ranging from -1.907 to -5.743) and countries like Canada and Singapore show varying degrees of vulnerability to high-impact breaches.

6 Discussion

6.1 Findings

Our study finds that larger organizations report cyberattacks more often, which might be due to their larger online operations and valuable information. Health and educational institutions are also common targets, probably because they hold sensitive personal data. The rising concern is that only about one in three organizations has a dedicated cyber security role or follows a cyber security framework. This suggests that many are not as prepared as they could be. Our models reveal that the presence of cybersecurity frameworks might be linked to a lower number of affected users, suggesting their potential protective effect. However, the implementation of education and awareness policies did not show a significant impact on reducing the number of users affected by cyberattacks.

The models also indicate that sectors such as arts, entertainment, and recreation, alongside 'Other' sectors, tend to see a higher number of users affected, showing possible sector-specific vulnerabilities. Increased digital intensity correlates with a greater number of affected users, which implies that as organizations become more digitally dependent, the potential impact of cyber incidents also rises.

6.2 Limitations and Bias

The analysis we performed hold many limitations. A major challenge we encountered was the high number of missing values within the dataset, which can introduce bias and limit the accuracy of our findings. For example, missing information may lead to under-representation of certain attack types or organizational responses, resulting in skewness of our understanding regarding the cyber threats.

The other obstacle was that the dataset contained just one numerical column (number of users affected), which provided a limited quantitative perspective.

Apart from this, the dataset reflected a biasness as it was more centered on cyber attacks from U.S.. Cybersecurity is a global issue, and the dominance of U.S. data could potentially overshadow the experiences of organizations in other regions. This geographical issues highlights the need for a more diverse compilation of data to study cyber risks more effectively.

These limitations suggests that there is the necessity for a more extensive dataset to study trends in cybercrime.

The weaknesses of the dataset are:

- The dataset spans over 15 years and was from 2004 to 2020, however, cyber threats are growing rapidly. Data from earlier years may not accurately reflect current trends and vulnerabilities.
- Not all organizations may report cyber incidents with the same transparency. This may result in a dataset that does not fully capture the extent of cyber threats across different sectors or sizes of organizations.
- There might be some years when more number of cyber incidents may be reported than other, leading to biasness in the study.

6.3 Future Research

Our study provides a foundational understanding of cyberattack patterns and their relationship with organizational factors and characteristics. However, it also open doors for future research in this area. For example, further investigation into the specific types of cybersecurity frameworks and their effectiveness across different sectors could help us to learn better hoe to deal with cyber attacks. It would also be beneficial to examine the role of cybersecurity education in different organizational cultures and its impact on reducing cyber risks.

It would be valuable if future research is conducted considering more geographic data and more multiple years to understand the long-term effect of preventive measures undertook by companies.Research on the global variations in cyberattacks, specifically focusing on countries outside the USA can be more helpful in providing accurate findings.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Clarke, Erik, Scott Sherrill-Mix, and Charlotte Dawson. 2023. *Ggbeeswarm: Categorical Scatter (Violin Point) Plots*. https://github.com/eclarke/ggbeeswarm.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://tibble.tidyverse.org/.

National Institute of Standards and Technology (NIST). 2023. "Cyber Resiliency." https://csrc.nist.gov/glossary/term/cyber_resiliency.

PwC. 2023. "Global Digital Trust Insights." https://www.pwc.com/us/en/services/consulting/cybersecurity-risk-regulatory/library/global-digital-trust-insights.html.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Slowikowski, Kamil. 2024. *Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot2'*. https://ggrepel.slowkow.com/.

Tsen, Elinor, Ryan KL Ko, and Sergeja Slapnicar. 2020. "Dataset of Data Breaches and Ransomware Attacks over 15 Years from 2004." University of Queensland.

UpGuard. 2024. "The Biggest Data Breaches in the u.s." https://www.upguard.com/blog/biggest-data-breaches-us.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019a. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.