TODO[*]

TODO

Shivank Goel

April 7, 2024

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

1 Introduction

In today's digital world, every click, every online transaction, and every shared piece of data is a potential entry point for cyber threats. Cybercrime is not just evolving, rather it is expanding at an alarming rate, 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" (https://www.upguard.com/blog/biggest-data-breaches-us)

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." (https://www.pwc.com/us/en/services/consulting/cybersecurity-risk-regulatory/library/global-digital-trust-insights.html)

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

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

(https://csrc.nist.gov/glossary/term/cyber_resiliency). To thoroughly analyze our study, we break it into three major research questions :

RQ1: How do things like the size of the company and the type of business it does affect its ability to handle cyber attacks?

RQ2: Which methods or strategies used by companies work best to reduce the damage from cyber attacks?

RQ3: How does the business's specific situation, like its industry or how much it relies on digital tools, change the impact of cyber attacks on it?

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 posture.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: 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. We provide the background and overview of the study in Section 2. **?@sec-model** 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, **?@sec-results** displays the interpretations of the model alongside other findings from analyzing the data, and **?@sec-discussion** 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.

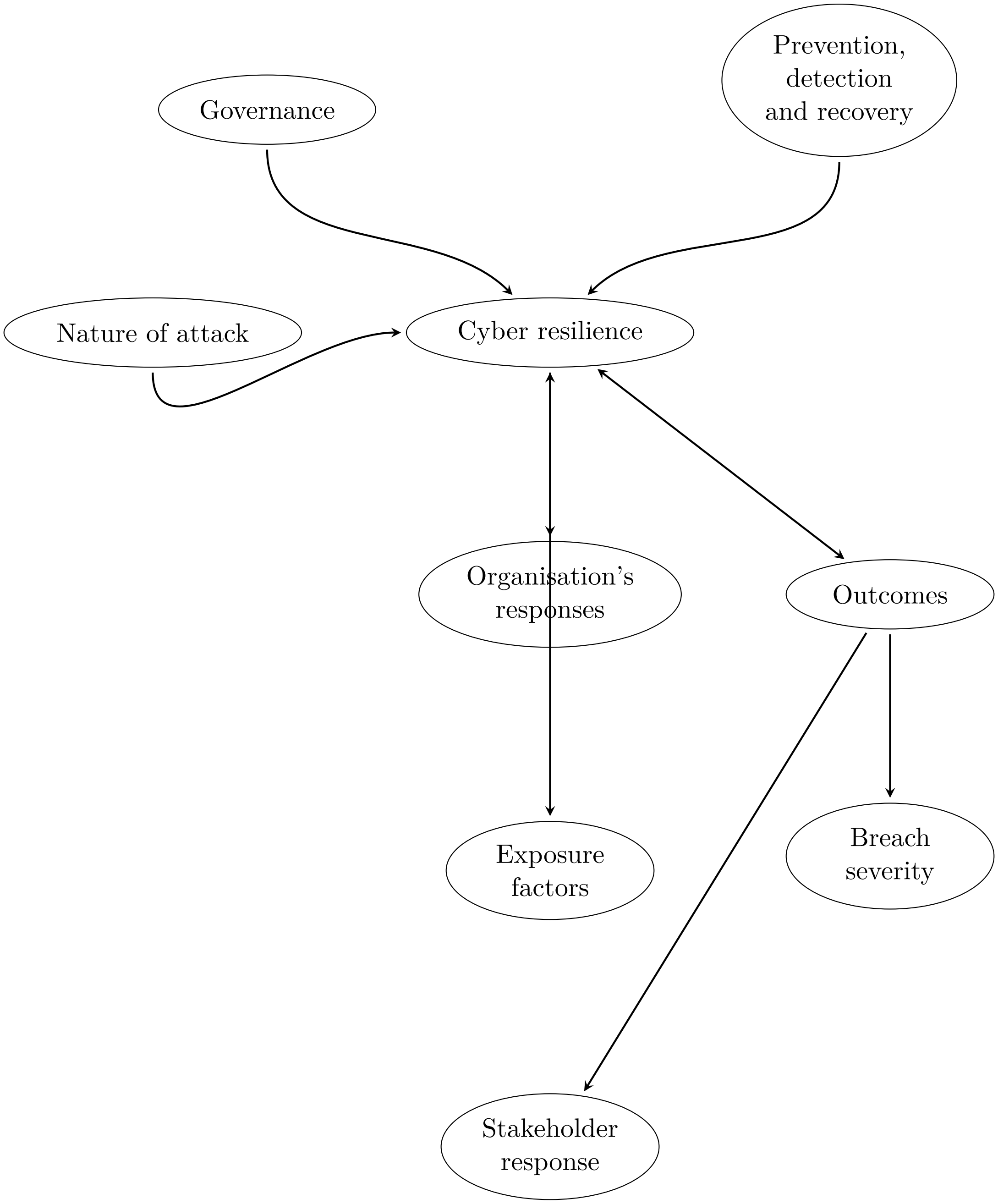


Figure 1: Conceptual model of organizational cyber resilience

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 https://espace.library.uq.edu.au/view/UQ:dfe5027 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.

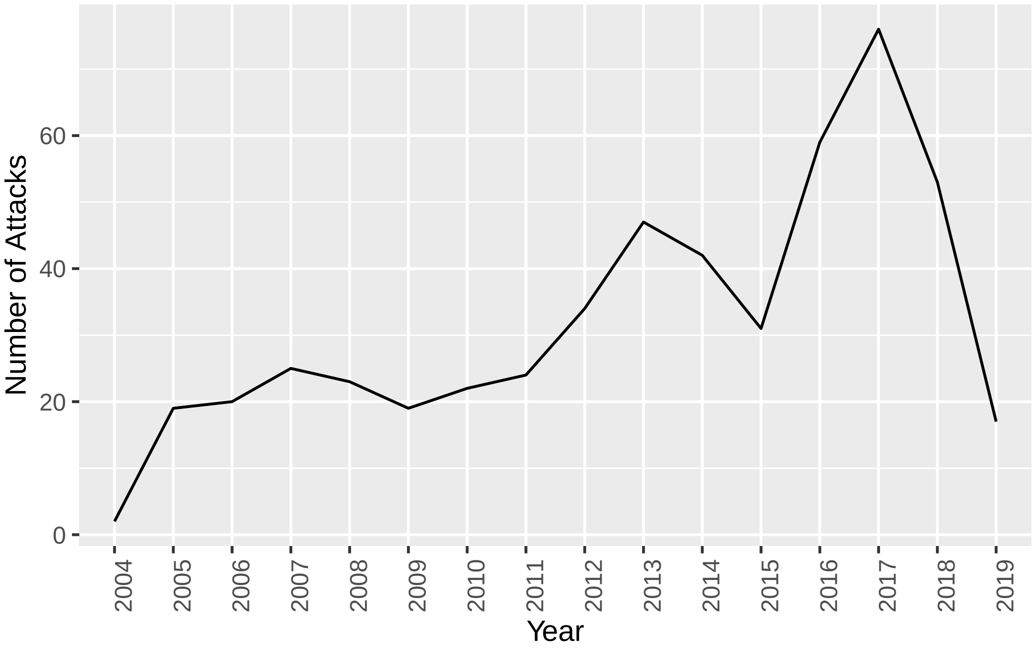


Figure 2: Cyberattacks Over Time

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

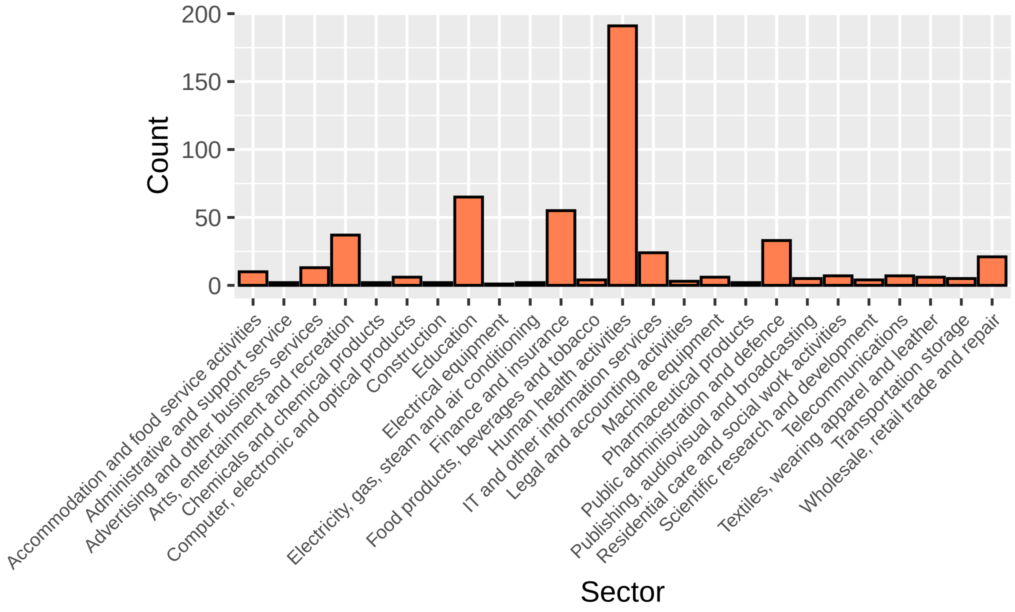


Figure 3: Bar Plot of Sector

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.

```
library(modelsummary)
```

```
Version 2.0.0 of `modelsummary`, to be released soon, will introduce a
  breaking change: The default table-drawing package will be `tinytable`
```

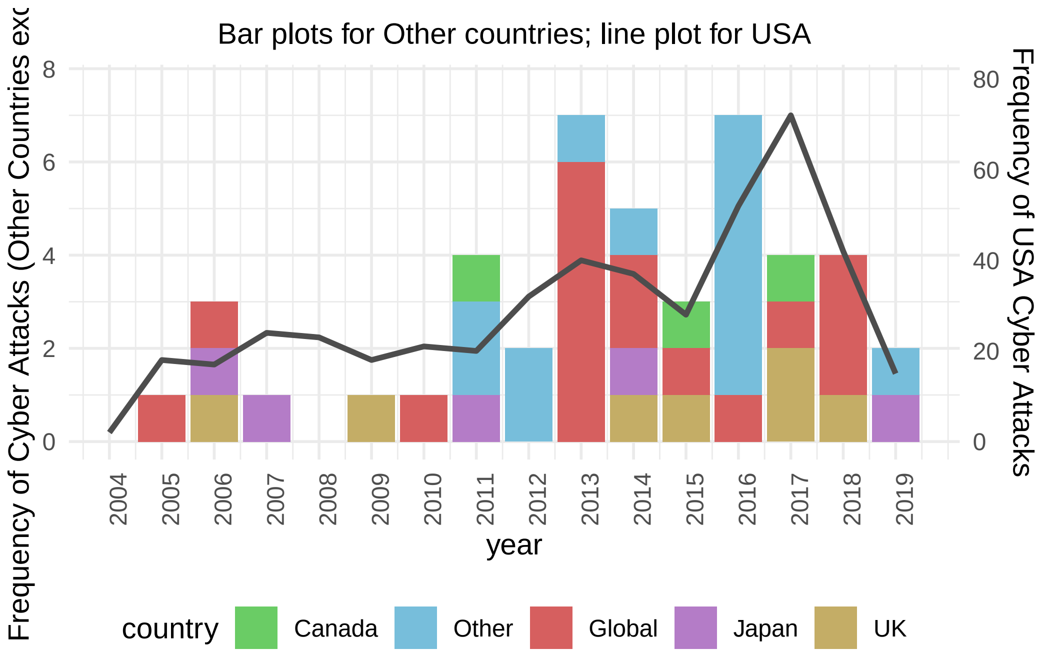


Figure 4: Overview of Cyber Attacks by Year and Country

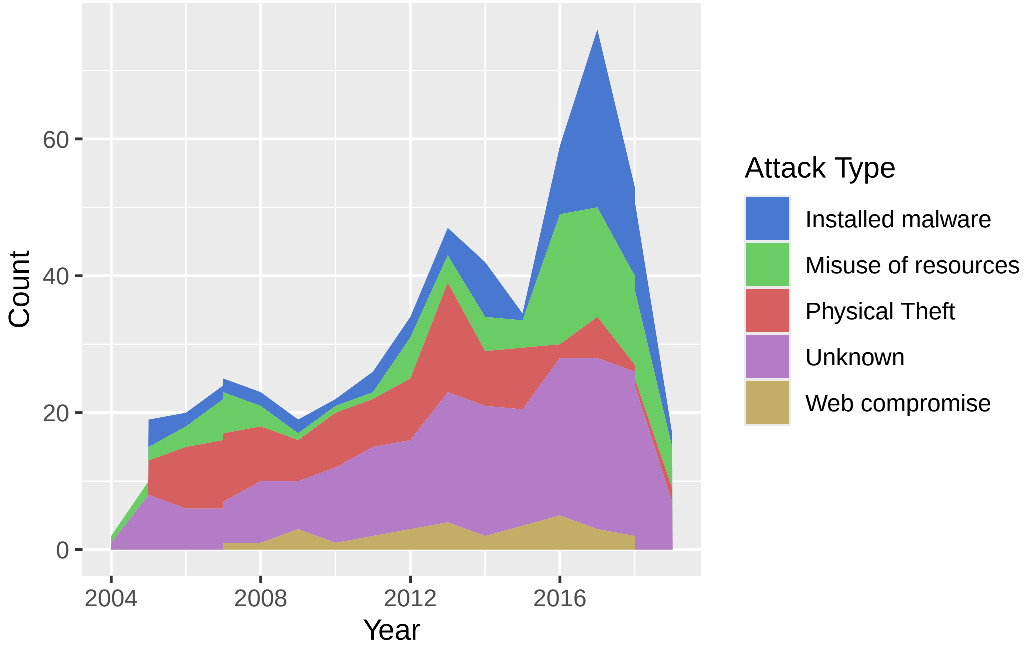


Figure 5: Attack Types Over Years

```
  instead of `kableExtra`. All currently supported table-drawing packages
  will continue to be supported for the foreseeable future, including
  `kableExtra`, `gt`, `huxtable`, `flextable, and `DT`.

  You can always call the `config_modelsummary()` function to change the
  default table-drawing package in persistent fashion. To try `tinytable`
  now:

  config_modelsummary(factory_default = 'tinytable')

  To set the default back to `kableExtra`:

  config_modelsummary(factory_default = 'kableExtra')
```

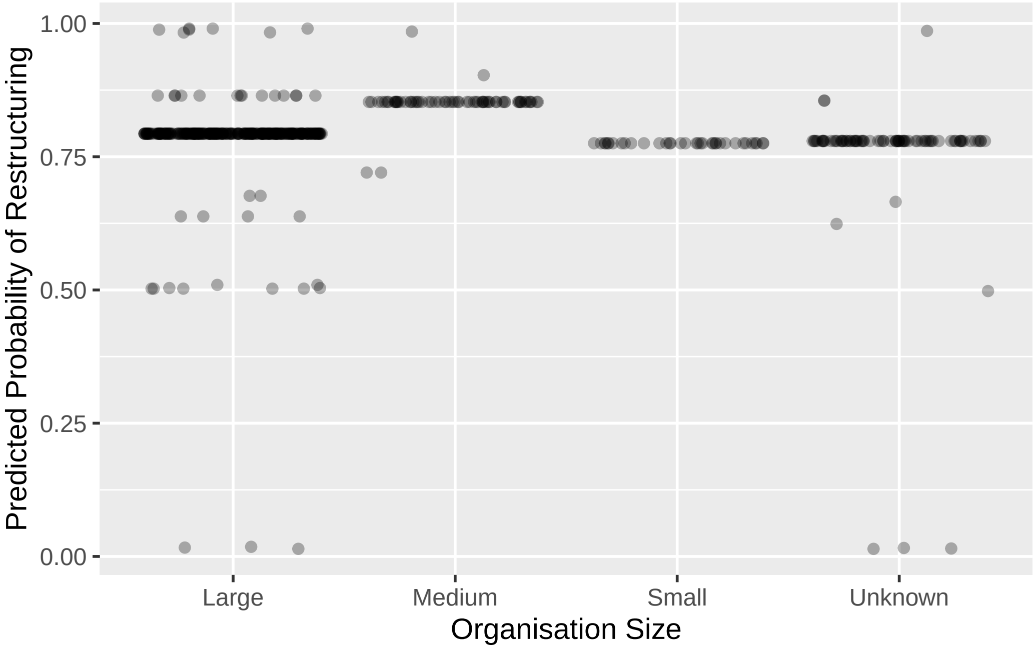
```r
logistic_model <- readRDS(file = here::here("models/restructuring_model.rds"))

modelsummary(list("Logistic Regression" = logistic_model))
```

```
Warning:
`modelsummary` uses the `performance` package to extract goodness-of-fit
statistics from models of this class. You can specify the statistics you wish
to compute by supplying a `metrics` argument to `modelsummary`, which will then
push it forward to `performance`. Acceptable values are: "all", "common",
"none", or a character vector of metrics names. For example: `modelsummary(mod,
metrics = c("RMSE", "R2")` Note that some metrics are computationally
expensive. See `?performance::performance` for details.
 This warning appears once per session.
```

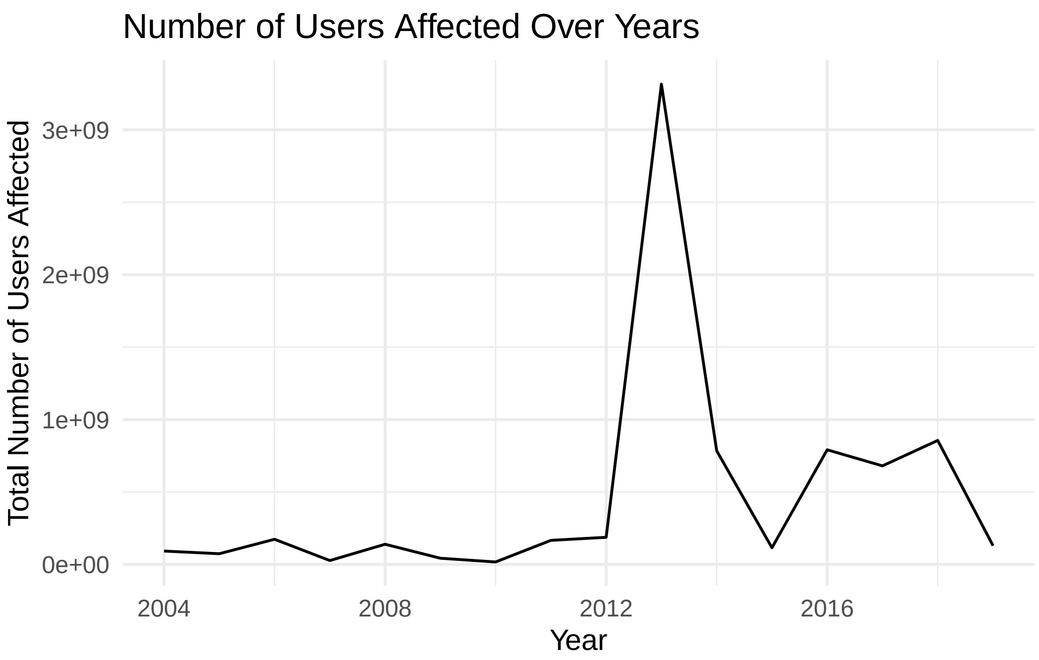
```r
breach_data <- breach_data %>% mutate(row_id = row_number())

# Adjust factors in your data to match the model's training data
breach_data <- breach_data %>%
  mutate(country = factor(country, levels = levels(logistic_model$model$country)))

# Generate predictions
breach_predictions <- predict(logistic_model, newdata = breach_data, type = "response")

# Combine the predictions with the original data
breach_data <- breach_data %>% mutate(predicted_prob = breach_predictions)

# Scatter plot with jitter
```

	Logistic Regression
(Intercept)	0.950
organisation_sizeMedium	0.437
organisation_sizeSmall	−0.076
organisation_sizeUnknown	−0.073
countryChina	33.477
countryFrance	−35.649
countryGermany	22.523
countryGlobal	1.069
countryHong Kong	32.456
countryIndia	−36.446
countryJapan	−0.987
countryNorway	−35.822
countryPhilippines	−35.592
countryQatar	−34.998
countryRussia	−0.999
countrySingapore	19.070
countrySouth Africa	34.073
countrySouth Korea	−0.897
countryTurkey	−36.347
countryUAE	34.251
countryUK	−0.348
countryUSA	0.394
Num.Obs.	417
R2	0.093
Log.Lik.	−203.054
ELPD	−221.7
ELPD s.e.	12.5
LOOIC	443.5
LOOIC s.e.	25.0
WAIC	434.9
RMSE	0.40

```r
ggplot(breach_data, aes(x = organisation_size, y = predicted_prob)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "Organisation Size", y = "Predicted Probability of Restructuring")
```



```r
breach_data %>%
  group_by(year) %>%
  summarize(total_users_affected = sum(number_of_users_affected, na.rm = TRUE)) %>%
  ggplot(aes(x = year, y = total_users_affected)) +
  geom_line() +
  labs(title = "Number of Users Affected Over Years",
       x = "Year",
       y = "Total Number of Users Affected") +
  theme_minimal()
```



```r
breach_data %>%
  filter(year == 2013) %>%
  summarise(
    median_users = median(number_of_users_affected, na.rm = TRUE),
    iqr_users = IQR(number_of_users_affected, na.rm = TRUE),
    upper_bound = median_users + 1.5 * iqr_users
  )
```

```
  median_users iqr_users upper_bound
1        56000    765688     1204532
```

```r
breach_data %>% filter(year == 2019)
```

```
   year                                         organisation critical_industry
1  2019              Blue Cross Blue Shield of Massachusetts               Yes
2  2019                                          Capital One               Yes
3  2019            Centerstone Insurance  Financial Services                No
4  2019     Critical Care, Pulmonary  Sleep Associates, PLLP               Yes
5  2019           Dr. DeLuca Dr. Marciano & Associates, P.C.               Yes
6  2019                                    EyeSouth Partners               Yes
7  2019                Integrated Regional Laboratories, LLC                No
8  2019 Las Colinas Orthopedic Surgery & Sports Medicine, PA               Yes
```

```
9  2019                                        Maffi Clinics               Yes
10 2019                        Memorial Hospital at Gulfport               Yes
11 2019                                  Mitsubishi Electric               Yes
12 2019          Pasquotank-Camden Emergency Medical Service               Yes
13 2019                               Providence Health Plan               Yes
14 2019                                    Quest Diagnostics               Yes
15 2019                   Singapore Ministry of Health - HIV               Yes
16 2019                   Union Labor Life Insurance Company                No
17 2019             Verity Health System of California, Inc.               Yes
   organisation_size level_of_digital_intensity
1              Large                 Low-Medium
2              Large                       High
3             Medium                       High
4             Medium                 Low-Medium
5              Small                 Low-Medium
6             Medium                 Low-Medium
7              Large                       High
8              Small                 Low-Medium
9              Small                 Low-Medium
10             Large                 Low-Medium
11             Large                Medium-High
12            Medium                 Low-Medium
13             Large                 Low-Medium
14             Large                 Low-Medium
15             Large                Medium-High
16             Large                       High
17             Large                 Low-Medium
                                sector   country cyber_security_role
1              Human health activities       USA                 Yes
2                Finance and insurance       USA                 Yes
3                Finance and insurance       USA                  No
4              Human health activities       USA                  No
5              Human health activities       USA                  No
6              Human health activities       USA                 Yes
7  Scientific research and development       USA                  No
8              Human health activities       USA                  No
9              Human health activities       USA                  No
10             Human health activities       USA                  No
11                Electrical equipment     Japan                 Yes
12             Human health activities       USA                  No
13             Human health activities       USA                  No
14             Human health activities       USA                  No
15   Public administration and defence Singapore                  No
```

```
16               Finance and insurance       USA                  No
17             Human health activities       USA                  No
   cyber_security_frameworks education_and_awareness_policy policy
1                         No                             No    Yes
2                         No                             No    Yes
3                         No                             No    Yes
4                         No                             No     No
5                         No                             No    Yes
6                         No                             No    Yes
7                         No                             No    Yes
8                         No                             No    Yes
9                         No                             No    Yes
10                        No                             No    Yes
11                        No                             No    Yes
12                        No                             No     No
13                        No                             No    Yes
14                        No                             No    Yes
15                        No                             No    Yes
16                        No                             No    Yes
17                        No                             No    Yes
   prevention_detection_and_recovery improper_network_segmentation
1                              Medium                          <NA>
2                                 Low                           Yes
3                                 Low                           Yes
4                              Medium                          <NA>
5                                High                           Yes
6                              Medium                            No
7                              Medium                            No
8                                 Low                            No
9                                 Low                           Yes
10                                Low                           Yes
11                                Low                           Yes
12                                Low                           Yes
13                                Low                            No
14                                Low                           Yes
15                                Low                            No
16                             Medium                          <NA>
17                             Medium                            No
   absence_of_encryption       detector restructuring_after_attack
1                   <NA>   Organisation                        Yes
2                   <NA> Federal Agency                        Yes
3                    Yes   Organisation                        Yes
4                   <NA>   Organisation                        Yes
```

```
5                    Yes   Organisation                         Yes
6                     No   Organisation                         Yes
7                    Yes   Organisation                         Yes
8                    Yes   Organisation                        <NA>
9                    Yes   Organisation                         Yes
10                   Yes   Organisation                          No
11                   Yes   Organisation                          No
12                   Yes   Organisation                         Yes
13                   Yes   Organisation                        <NA>
14                   Yes   Organisation                         Yes
15                   Yes           <NA>                         Yes
16                  <NA>   Organisation                         Yes
17                   Yes   Organisation                         Yes
   bribe_ransom_paid free_identity_or_credit_theft_monitoring
1                 No                                      Yes
2                 No                                      Yes
3                 No                                      Yes
4                 No                                       No
5                 No                                      Yes
6                 No                                     <NA>
7                 No                                       No
8                 No                                     <NA>
9                 No                                       No
10                No                                      Yes
11                No                                     <NA>
12                No                                      Yes
13                No                                     <NA>
14                No                                      Yes
15                No                                     <NA>
16                No                                      Yes
17                No                                      Yes
   additional_disclosure_of_information number_of_users_affected
1                                   Yes                 11000000
2                                   Yes                106000000
3                                   Yes                   111589
4                                    No                    23300
5                                   Yes                    23578
6                                  <NA>                    24113
7                                  <NA>                    29644
8                                  <NA>                    76000
9                                   Yes                    10465
10                                 <NA>                    30000
11                                 <NA>                     8000
```

```
12                                 <NA>                    40000
13                                 <NA>                   122000
14                                  Yes                 12000000
15                                  Yes                    14200
16                                   No                    87400
17                                  Yes                    14894
   overall_nature_of_attack         attack_type attacker
1                      <NA>             Unknown External
2                      <NA>             Unknown External
3                    Type 2 Misuse of resources External
4                    Type 2 Misuse of resources External
5                    Type 1   Installed malware External
6                    Type 2 Misuse of resources External
7                    Type 2 Misuse of resources External
8                    Type 3      Physical Theft External
9                    Type 1   Installed malware External
10                   Type 2 Misuse of resources External
11                     <NA>             Unknown External
12                     <NA>             Unknown External
13                   Type 2             Unknown External
14                   Type 2 Misuse of resources External
15                   Type 3      Physical Theft Internal
16                   Type 2             Unknown External
17                   Type 2             Unknown External
                    attack_vector impact_on_data
1          Unknown network attack         Medium
2          Unknown network attack           High
3              Social engineering           High
4                            <NA>           High
5                            <NA>           High
6              Social engineering         Medium
7            Vendor vulnerability         Medium
8                 Physical device         Medium
9                            <NA>           High
10             Social engineering         Medium
11                           <NA>         Medium
12                           <NA>         Medium
13           Vendor vulnerability         Medium
14           Vendor vulnerability            Low
15 Inappropriate use of privilege         Medium
16             Social engineering         Medium
17             Social engineering         Medium
   aspect_of_confidentiality_integrity_availability_triad_affected
```

```
1                                                  Confidentiality
2                                                  Confidentiality
3                                                  Confidentiality
4                                                  Confidentiality
5                                                     Availability
6                                                  Confidentiality
7                                                  Confidentiality
8                                                  Confidentiality
9                                                     Availability
10                                                 Confidentiality
11                                                 Confidentiality
12                                                 Confidentiality
13                                                 Confidentiality
14                                                 Confidentiality
15                                                 Confidentiality
16                                                 Confidentiality
17                                                 Confidentiality
   individual_s_name_s_leaked_exposed address_es_leaked_exposed
1                                 Yes                       Yes
2                                 Yes                       Yes
3                                 Yes                       Yes
4                                 Yes                       Yes
5                                 Yes                       Yes
6                                 Yes                       Yes
7                                 Yes                       Yes
8                                 Yes                       Yes
9                                 Yes                       Yes
10                                Yes                       Yes
11                                Yes                      <NA>
12                                Yes                       Yes
13                                Yes                       Yes
14                                Yes                        No
15                                Yes                       Yes
16                                Yes                       Yes
17                                Yes                       Yes
   other_personally_identifiable_information_pii_leaked_exposed
1                                                           Yes
2                                                           Yes
3                                                           Yes
4                                                           Yes
5                                                           Yes
6                                                           Yes
7                                                           Yes
```

```
8                                                          Yes
9                                                          Yes
10                                                         Yes
11                                                         Yes
12                                                         Yes
13                                                         Yes
14                                                         Yes
15                                                         Yes
16                                                         Yes
17                                                         Yes
   track_1_credit_card_details_leaked_exposed
1                                          No
2                                         Yes
3                                         Yes
4                                          No
5                                          No
6                                          No
7                                          No
8                                          No
9                                          No
10                                         No
11                                         No
12                                         No
13                                         No
14                                         No
15                                         No
16                                         No
17                                         No
   track_2_credit_card_details_leaked_exposed
1                                          No
2                                        <NA>
3                                        <NA>
4                                          No
5                                          No
6                                          No
7                                          No
8                                          No
9                                          No
10                                         No
11                                         No
12                                         No
13                                         No
14                                         No
```

```
15                                        No
16                                        No
17                                        No
   social_security_number_tax_number_leaked_exposed
1                                               Yes
2                                               Yes
3                                               Yes
4                                               Yes
5                                               Yes
6                                               Yes
7                                               Yes
8                                               Yes
9                                               Yes
10                                              Yes
11                                              Yes
12                                              Yes
13                                              Yes
14                                               No
15                                               No
16                                              Yes
17                                              Yes
   subsequent_fraudulent_use_of_data investigation undertook_investigation
1                                 No            No                      No
2                                 No           Yes                     Yes
3                                 No           Yes                     Yes
4                                Yes           Yes                     Yes
5                                 No           Yes                     Yes
6                                 No            No                      No
7                                 No            No                      No
8                                 No            No                      No
9                                 No           Yes                     Yes
10                                No            No                      No
11                              <NA>            No                      No
12                                No            No                      No
13                                No           Yes                     Yes
14                                No           Yes                      No
15                                No           Yes                     Yes
16                                No            No                      No
17                                No            No                      No
   litigation_by_public penalties_settlement_paid_or_actions_imposed
1                    No                                           No
2                    No                                           No
3                    No                                           No
```

```
4                     No                                           No
5                     No                                           No
6                     No                                           No
7                     No                                           No
8                     No                                           No
9                     No                                           No
10                    No                                           No
11                    No                                           No
12                    No                                           No
13                    No                                           No
14                   Yes                                          Yes
15                   Yes                                          Yes
16                    No                                           No
17                    No                                           No
   imposed_penalties_or_actions_on_organisation
1                                            No
2                                            No
3                                            No
4                                            No
5                                            No
6                                            No
7                                            No
8                                            No
9                                            No
10                                           No
11                                           No
12                                           No
13                                           No
14                                           No
15                                          Yes
16                                           No
17                                           No
   fines_issued_by_government_or_relevant_body settlement_paid row_id
1                                           No              No     46
2                                           No              No     61
3                                           No              No     72
4                                           No              No    106
5                                           No              No    124
6                                           No              No    146
7                                           No              No    201
8                                           No              No    219
9                                           No              No    234
10                                          No              No    246
```

```
11                                          No              No    256
12                                          No              No    312
13                                          No              No    327
14                                          No             Yes    335
15                                          No              No    377
16                                          No              No    441
17                                          No              No    482
   predicted_prob
1       0.7932777
2       0.7932777
3       0.8527533
4       0.8527533
5       0.7752741
6       0.8527533
7       0.7932777
8       0.7752741
9       0.7752741
10      0.7932777
11      0.5024668
12      0.8527533
13      0.7932777
14      0.7932777
15      0.9904011
16      0.7932777
17      0.7932777
```

```
breach_data %>%
  filter(year == 2019)%>%
  summarise(
    median_users = median(number_of_users_affected, na.rm = TRUE),
    iqr_users = IQR(number_of_users_affected, na.rm = TRUE),
    upper_bound = median_users + 1.5 * iqr_users
  )
```

```
  median_users iqr_users upper_bound
1        30000     88289    162433.5
```

```
# Summary statistics for numerical variables
summary(breach_data$number_of_users_affected)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
6.900e+02 2.500e+04 9.300e+04 1.482e+07 8.470e+05 3.000e+09         1 
```

```r
summary(breach_data$year)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2004    2010    2014    2013    2017    2019 
```

```r
# Frequency tables for categorical variables
table(breach_data$sector)
```

```

   Accommodation and food service activities 
                                          10 
          Administrative and support service 
                                           2 
     Advertising and other business services 
                                          13 
          Arts, entertainment and recreation 
                                          37 
             Chemicals and chemical products 
                                           2 
   Computer, electronic and optical products 
                                           6 
                                Construction 
                                           2 
                                   Education 
                                          65 
                        Electrical equipment 
                                           1 
Electricity, gas, steam and air conditioning 
                                           2 
                       Finance and insurance 
                                          55 
        Food products, beverages and tobacco 
                                           4 
                     Human health activities 
                                         191 
           IT and other information services 
                                          24 
             Legal and accounting activities 
                                           3 
                           Machine equipment 
                                           6 
                     Pharmaceutical products
                                           2
           Public administration and defence
                                          33
    Publishing, audiovisual and broadcasting
                                           5
 Residential care and social work activities
                                           7
         Scientific research and development
                                           4
                          Telecommunications
                                           7
       Textiles, wearing apparel and leather
                                           6
                      Transportation storage
                                           5
          Wholesale, retail trade and repair
                                          21
```

```
table(breach_data$organisation_size)
```

```
  Large  Medium   Small Unknown
    329      66      35      83
```

```
table(breach_data$critical_industry)
```

```
 No Yes
182 331
```

```
table(breach_data$level_of_digital_intensity)
```

```
       High         Low  Low-Medium Medium-High
        109          22         273         109
```

```
table(breach_data$country)
```

```
      Canada        China       France      Germany       Global    Hong Kong 
           3            1            1            2           17            1 
       India        Japan       Norway  Philippines        Qatar       Russia 
           1            5            1            1            1            2 
   Singapore South Africa  South Korea       Turkey          UAE           UK 
           3            1            3            1            1            7 
         USA 
         461 
```

```r
table(breach_data$cyber_security_role)
```

```
 No Yes 
452  61 
```

```r
table(breach_data$cyber_security_frameworks)
```

```
 No Yes 
511   2 
```

```r
table(breach_data$education_and_awareness_policy)
```

```
 No Yes 
512   1 
```

```r
table(breach_data$policy)
```

```
 No Yes 
  3 499 
```

```r
table(breach_data$prevention_detection_and_recovery)
```

```
  High    Low Medium 
     4    286    223 
```

```r
table(breach_data$detector)
```

```

Credit card/bank   Federal Agency     Organisation           Public
              12               15              457               16
```

```r
table(breach_data$restructuring_after_attack)
```

```

 No Yes
 90 327
```

```r
table(breach_data$bribe_ransom_paid)
```

```

 No Yes
512   1
```

```r
table(breach_data$free_identity_or_credit_theft_monitoring)
```

```

 No Yes
236 195
```

```r
table(breach_data$additional_disclosure_of_information)
```

```

 No Yes
200 205
```

```r
table(breach_data$overall_nature_of_attack)
```

```

Type 1 Type 2 Type 3 Type 4 Type 5
    83    107    106     25     15
```

```r
table(breach_data$attack_type)
```

```

  Installed malware Misuse of resources      Physical Theft             Unknown
                 83                  91                 106                 208
     Web compromise
                 25
```

```r
table(breach_data$attacker)
```

```

External Internal
     495       18
```

```r
table(breach_data$attack_vector)
```

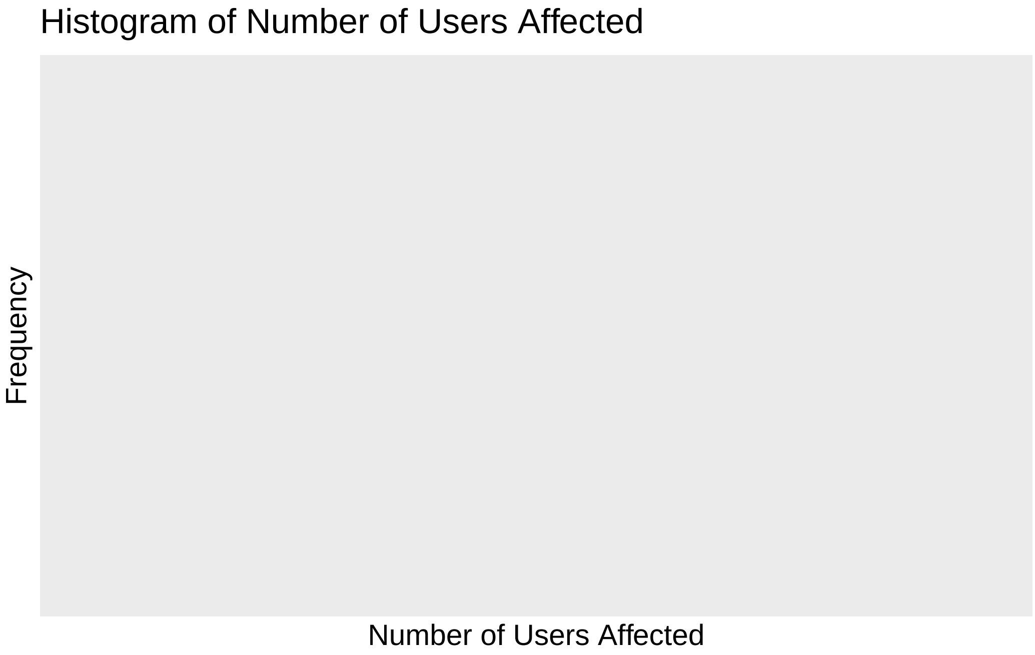
```

        Inappropriate use of privilege Insufficient authentication validation
                                    13                                      8
         Insufficient input validation                        Physical device
                                    23                                    100
                    Social engineering                  Unknown device attack
                                    46                                      7
                Unknown network attack Unknown website/web application attack
                                    82                                     13
                  Vendor vulnerability
                                    39
```

```r
table(breach_data$impact_on_data)
```

```

  High    Low Medium
   155    111    247
```

```r
# ... continue for other categorical variables as needed

# Histogram for a continuous variable (e.g., number_of_users_affected)
ggplot(breach_data, aes(x = number_of_users_affected)) +
  geom_histogram(binwidth = 1000, fill = "blue", color = "black") +
  labs(title = "Histogram of Number of Users Affected", x = "Number of Users Affected", y = "
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_bin()`).
```

```
Warning: Computation failed in `stat_bin()`.
Caused by error in `bin_breaks_width()`:
! The number of histogram bins must be less than 1,000,000.
i Did you make `binwidth` too small?
```

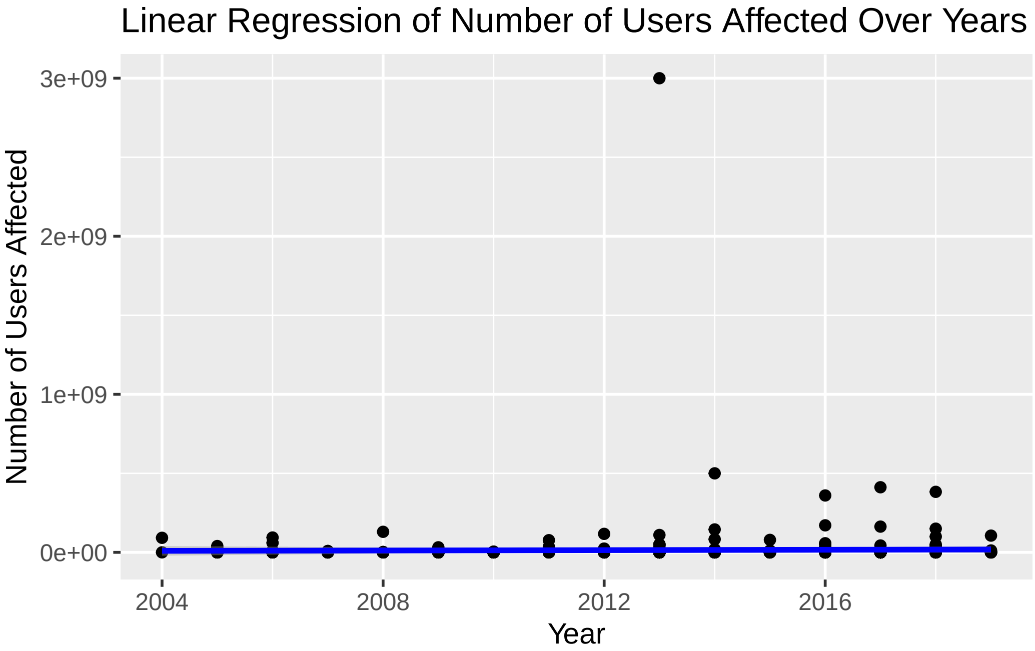


```r
linear_model_RQ2 <- readRDS(file = here::here("models/linear_model_RQ2.rds"))
# Plotting diagnostics for the linear regression model (Example: linear_model_RQ2)
ggplot(breach_data, aes(x = year, y = number_of_users_affected)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Linear Regression of Number of Users Affected Over Years",
       x = "Year",
       y = "Number of Users Affected")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
```

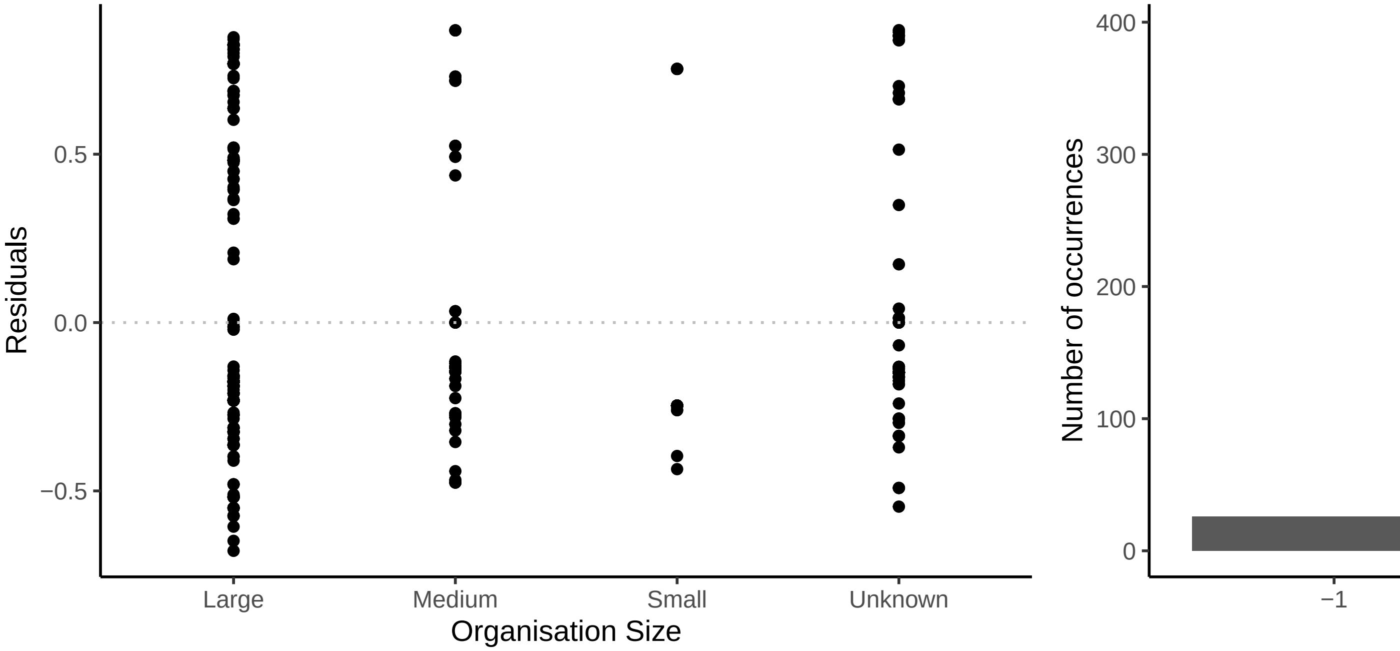
```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



```
# Assuming breach_data is your dataset and you have already created models named linear_model

# Partial regression plot for breach severity with a specific predictor (e.g., 'organisation_
# install.packages("car") # Uncomment if the car package is not installed
```

```
Warning: package 'broom' was built under R version 4.3.3
```

```
Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```


Residuals
0.5
0.0
−0.5
Large
Medium
Small
Unknown
Organisation Size
Number of occurrences
400
300
200
100
0
−1

$

$

Table 3: Linear regression analysis with dependent variable

	Dependent variable:
	undertook_investigation
critical_industry	−0.03 (0.12)
organisation_sizeMedium	−0.04 (0.06)
organisation_sizeSmall	0.07 (0.08)
organisation_sizeUnknown	−0.03 (0.06)
level_of_digital_intensityLow	−0.03 (0.62)
level_of_digital_intensityLow-Medium	−0.29 (0.67)
level_of_digital_intensityMedium-High	−0.31 (0.66)
sectorAdministrative and support service	−0.62 (0.72)
sectorAdvertising and other business services	−0.48 (0.66)
sectorArts, entertainment and recreation	−0.14 (0.12)
sectorChemicals and chemical products	0.20 (0.37)
sectorComputer, electronic and optical products	0.19 (0.23)
sectorConstruction	−0.61 (0.45)
sectorEducation	−0.11 (0.18)
sectorElectrical equipment	−0.23 (0.51)
sectorElectricity, gas, steam and air conditioning	−0.56 (0.34)
sectorFinance and insurance	−0.09 (0.63)
sectorFood products, beverages and tobacco	−0.06 (0.26)
sectorHuman health activities	−0.14 (0.22)
sectorIT and other information services	−0.29 (0.63)
sectorLegal and accounting activities	−0.63 (0.69)
sectorMachine equipment	−0.08 (0.24)
sectorPharmaceutical products	−0.27 (0.38)
sectorPublishing, audiovisual and broadcasting	−0.12 (0.22)
sectorResidential care and social work activities	−0.18 (0.24)
sectorScientific research and development	−0.37 (0.68)
sectorTelecommunications	−0.40 (0.65)
sectorTextiles, wearing apparel and leather	
sectorTransportation storage	0.06 (0.31)
sectorWholesale, retail trade and repair	
countryChina	0.88 (0.51)
countryFrance	−0.17 (0.51)
countryGermany	−0.01 (0.40)
countryGlobal	0.28 (0.29)
countryHong Kong	0.56 (0.57)
countryIndia	−0.33 (0.51)
countryJapan	0.13 (0.35)
countryNorway	0.04 (0.50)
countryPhilippines	−0.15 (0.51)
countryQatar	−0.33 (0.51)
countryRussia	−0.12 (0.41)
countrySingapore	0.83* (0.36)
countrySouth Africa	0.85 (0.51)
countrySouth Korea	0.53 (0.36)
countryTurkey	0.85 (0.51)
countryUAE	0.60 (0.57)
countryUK	0.35 (0.31)
countryUSA	0.19 (0.26)

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/.

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/.

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.