TODO*

TODO

Shivank Goel

April 2, 2024

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

1 Introduction

TODO today's world, where everything is connected online, cybersecurity is super important. It's all about keeping our digital stuff safe from bad guys who try to steal or mess with it. Cyber threats, like hackers breaking into computer systems or stealing personal information, are becoming more common and tricky to deal with.

This paper is all about diving into the world of cybersecurity to learn about the problems we face, the trends we're seeing, and how we can protect ourselves better. We'll start by looking at some big cyber incidents that have happened recently. By understanding what happened in these incidents, we can figure out how to stop similar attacks in the future.

Then, we'll talk about the new kinds of cyber threats that are popping up all the time. From sneaky tricks like ransomware to people tricking us into giving away our info, we'll explore the different ways bad guys try to break into our digital stuff.

Finally, we'll talk about what we can do to fight back against cyber threats. This includes using smart technology, making rules and policies to keep things safe, and teaching people how to be careful online. By learning more about cyber risks and working together to stay safe, we can make sure our digital world is a lot safer for everyone.

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

2 Data

2.1 Data Source and Collection:

TODO

The study relies on datasets obtained from the provincial open databases of Alberta, accessible through the official website government (2024). Three key datasets were utilized to extract relevant variables for analysis, aiming to uncover the relationship between air quality and mortality rates in Alberta. The analysis begins with the leading causes of death dataset for Alberta, sourced from the provincial open data portal government (2023b). This dataset provides insights into mortality rates associated with various illnesses, facilitating the examination of trends related to respiratory and heart-related illnesses. To explore potential correlations between air quality and mortality rates, the study incorporates the Air Quality Health Index (AQHI) dataset for Alberta, sourced from the provincial open data portal government (2023a). This dataset offers comprehensive information on the AQHI across different municipalities in Alberta over multiple years. Additionally, the study utilizes PM2.5 air pollutant concentration level data sourced from Alberta's official resources Alberta Government (2023). This dataset provides detailed information on the concentration levels of PM2.5 pollutants over several years, offering valuable insights into air quality trends. The following subsections outline the sources, collection methodologies, and data-cleaning procedures implemented to ensure the accuracy and reliability of the datasets used in the analysis. This meticulous approach ensures that the data is prepared for thorough analysis, facilitating the exploration of correlations between air quality indicators and mortality rates in Alberta.

Leading Causes of Death in Alberta Data: The disease data is found from the government of Alberta's open data portal, and was last updated on September 22, 2023 and continues to be updated annually. This dataset encompasses mortality data related to the top 30 common causes of death. It reports on types of diseases, causes of death, mortality denoted by total death counts, and ranking for 2000-2022. Due to our focus on respiratory illnesses, in the leading cause of death dataset, we grouped diseases by categories. Our category of focus included filtering on illnesses like acute myocardial infarction, malignant neoplasms of the trachea, bronchus, and lung, other chronic obstructive pulmonary disease, and all other forms of chronic ischemic heart disease. Leading causes of death are measured and ranked by the top 30 most common death causes each specific year. The causes of death are classified based on the International Classification of Diseases 10th Edition.

AQHI Data: The second dataset we used is the air quality health index (AQHI) dataset found at the government of Alberta's open data portal. This dataset contains AQHI by municipality for the years 2012-2022 and reports air quality health index, and health risk both quantitatively and qualitatively. To use the AQHI dataset we employed simple data-cleaning practice to maintain descriptive variable names and readability. The data is measured by the percentage of hours for each year at a given air quality level, by municipality. The Air Quality Health Index is calculated based on the relative risks of a combination of common air pollutants that is known to harm human health. These pollutants are ozone (O3) at ground level, particulate matter (PM2.5), and nitrogen dioxide (NO2). Risks are defined as follows: 1-3 High Quality; 4-6 Moderate Quality; 7-9 Low Quality; 10+ Very Low Quality.

PM2.5 Data: We used the PM2.5 data set retrieved from Alberta.ca (Government of Alberta) which was last updated in April 2023. It reports on average PM2.5 concentration levels through the years 2000-2021 using a provincial average, the 10th percentile quantities, and the 90th percentile quantities, with a focus on 8 municipalities Edmonton, Fort McMurray, Grande Prairie, Lethbridge, Medicine Hat, and Red Deer respectively and lastly reports the Canadian Ambient Air Quality Standard (CAAQS) value.

The Alberta Air Zone report Environment and Areas (2021), which is linked to our dataset, provides a detailed explanation of the measurement and processing of the PM2.5 quantity. Alberta Air Zones divides Alberta into six air zones which are aligned with Alberta's Land-use Framework regional boundaries. Ambient air quality in Alberta is monitored at continuous air monitoring stations located within these air zones. PM2.5 quantities are taken throughout these stations across Alberta, and they measure the quantities in µg (micrograms per cubic meter of air).

2.2 Data Cleaning

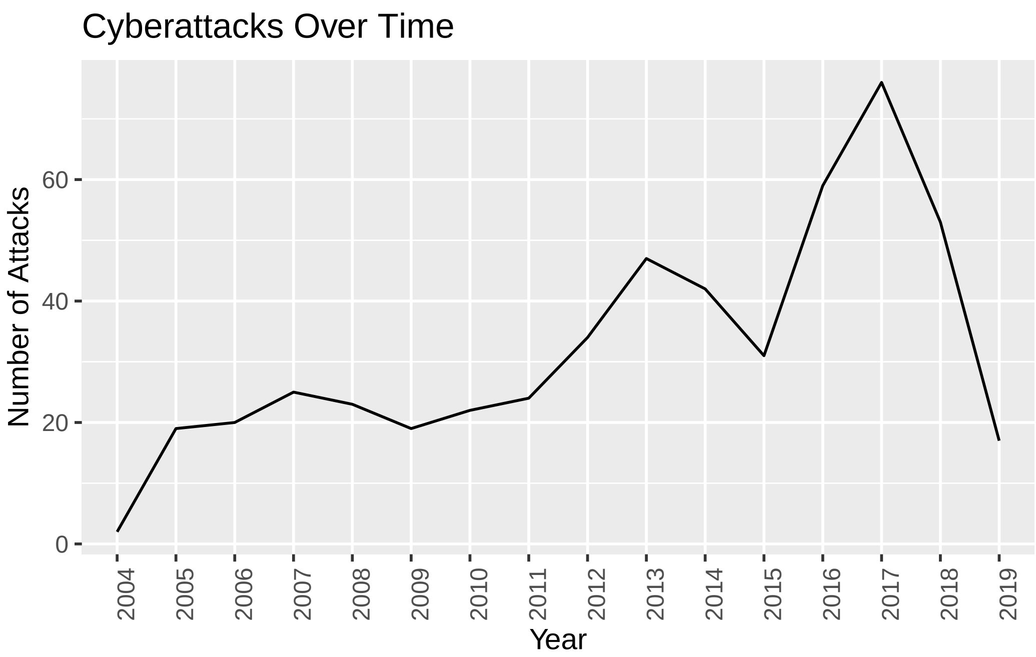
We used R (R Core Team 2023) and Wickham et al. (2019a) for data cleaning and processing, utilizing packages like tidyverse (Wickham et al. 2019b) for data manipulation and janitor (Firke 2023) for cleaning column names. Other packages used includes `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2023), `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 raw air quality data were preprocessed to remove inconsistencies and irrelevant information. Specifically, we filtered the dataset to include observations from the years 2012 to 2021, which are relevant to our analysis. Additionally, we merged this dataset with additional information on peak pollution levels for comprehensive analysis. Similar to the previous datasets, the raw mortality data underwent cleaning procedures to focus on specific causes relevant to our analysis. We filtered the dataset to include observations up to 2021 and merged it with additional information on air quality for correlation analysis. The raw data on AQHI were filtered to include observations from the years 2011 to 2021 for consistency with other datasets. Additionally, the data were aggregated at the municipal level for further analysis.

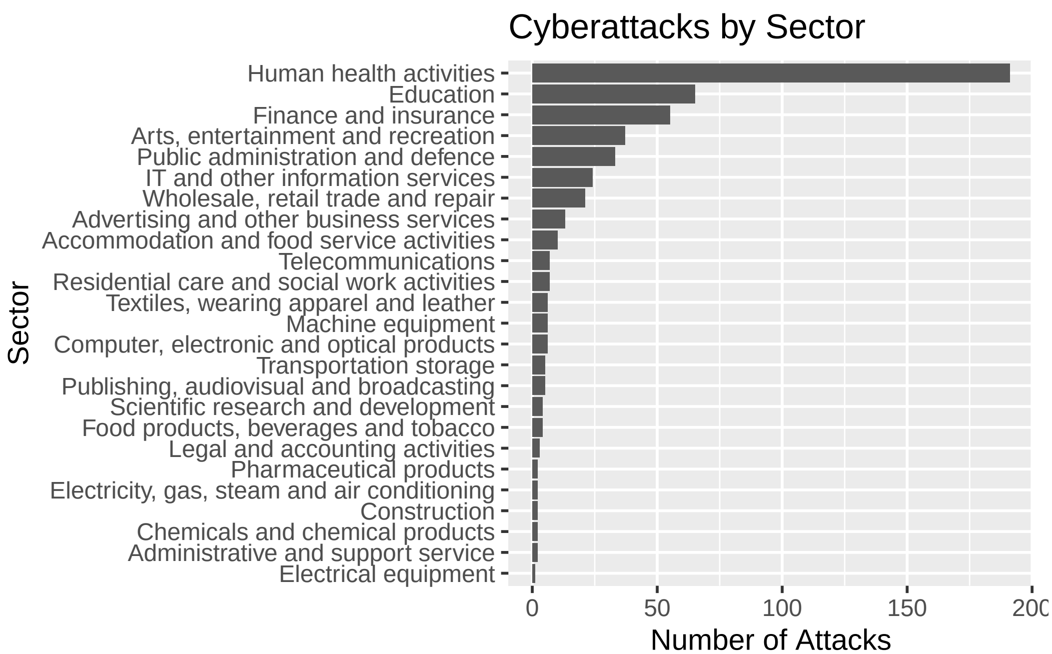
2.3 Data Modifications

In this study, we constructed unique datasets by thoughtfully selecting and merging data from the Government of Alberta's open data portal, Alberta.ca, spanning the years 2012 to 2022. Our process involved merging variables from various datasets to create specific datasets tailored for model building and analysis. One such dataset, 'cleaned_chart_data,' was created by merging variables such as causes of death, total deaths, provincial average PM2.5 levels, and CAAQS. This dataset was designed to facilitate our analysis of any significant correlations between these variables. A snapshot of this data is referenced in Table X. Additionally, we derived two other datasets, 'merged_data' and 'merged_heart_data,' by merging variables related to heart disease numbers, lung disease numbers, and provincial average PM2.5 values. These datasets were instrumental in examining the impact of PM2.5 on each type of illness, as previously discussed. Overall, our methodology ensured the creation of comprehensive datasets that allowed for a detailed investigation into the relationships between PM2.5 levels and various health outcomes in Alberta.

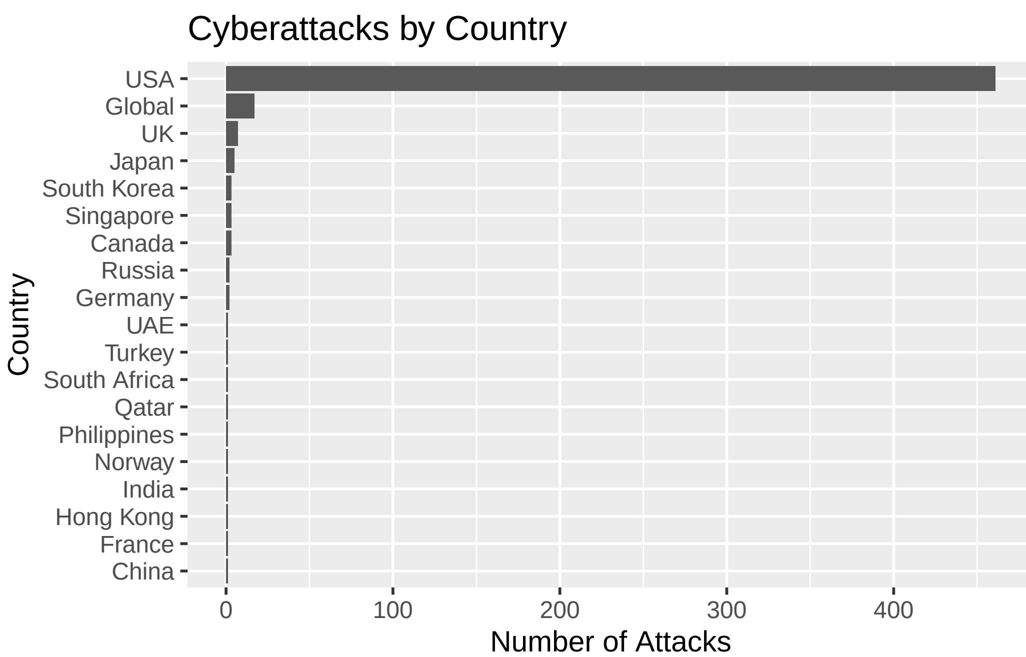
```
breach_data %>%
  count(year) %>%
  ggplot(aes(x = as.factor(year), y = n)) +  # Convert year to factor to treat it as discrete
  geom_line(group=1) +  # Ensure geom_line treats the data as connected points
  scale_x_discrete(breaks = levels(as.factor(breach_data$year))) +  # Specify breaks at every
  labs(title = "Cyberattacks Over Time", x = "Year", y = "Number of Attacks") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x labels to fit them al
```



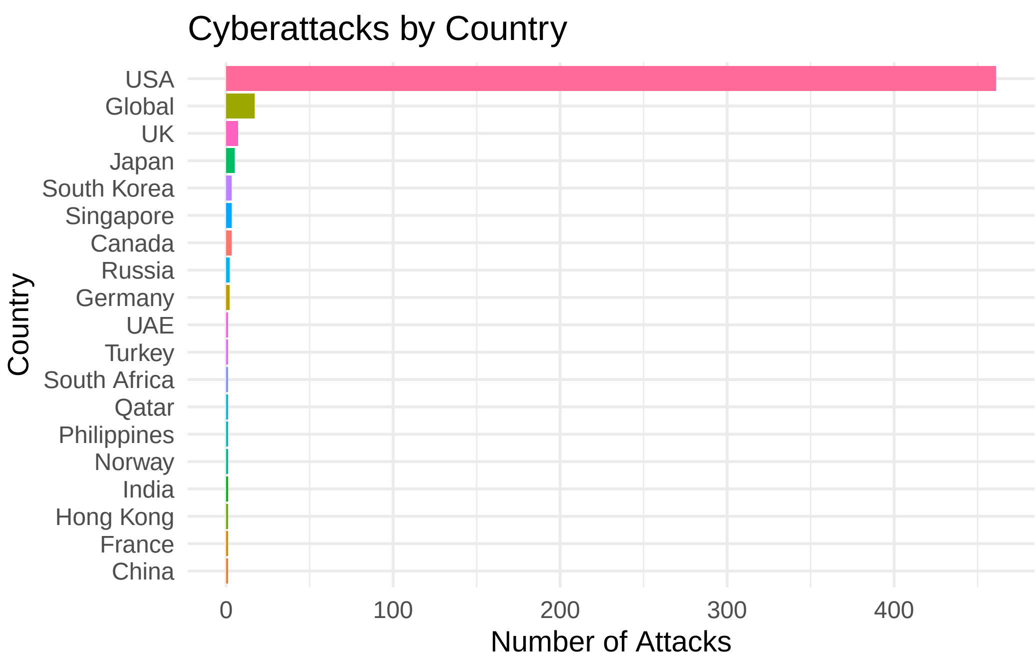
```r
breach_data %>%
  count(sector) %>%
  ggplot(aes(x = reorder(sector, n), y = n)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Cyberattacks by Sector", x = "Sector", y = "Number of Attacks")
```



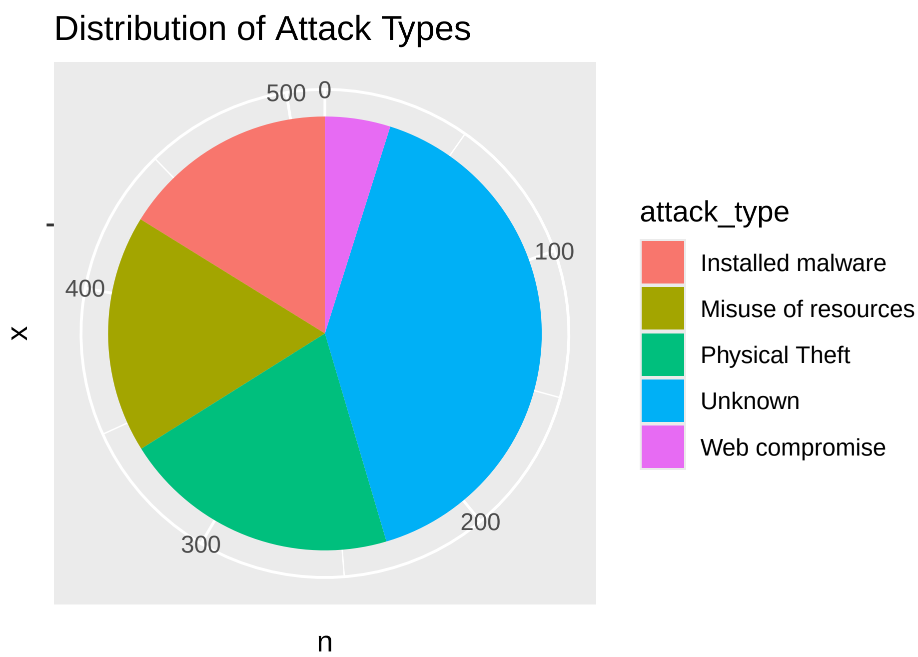

```r
breach_data %>%
  count(country) %>%
  ggplot(aes(x = reorder(country, n), y = n)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Cyberattacks by Country", x = "Country", y = "Number of Attacks")
```




```
breach_data %>%
  count(country) %>%
  ggplot(aes(x = reorder(country, n), y = n, fill = country)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Cyberattacks by Country", x = "Country", y = "Number of Attacks") +
  theme_minimal() +
  theme(legend.position = "none") # Hides the legend
```



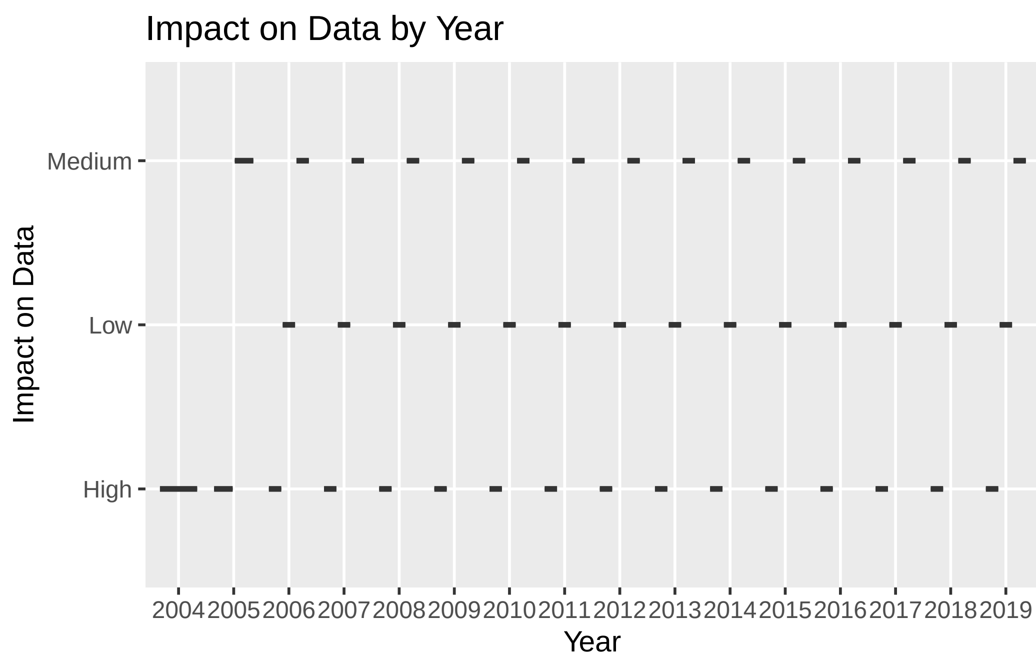

```
breach_data %>%
  count(attack_type) %>%
  ggplot(aes(x = "", y = n, fill = attack_type)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  labs(title = "Distribution of Attack Types")
```



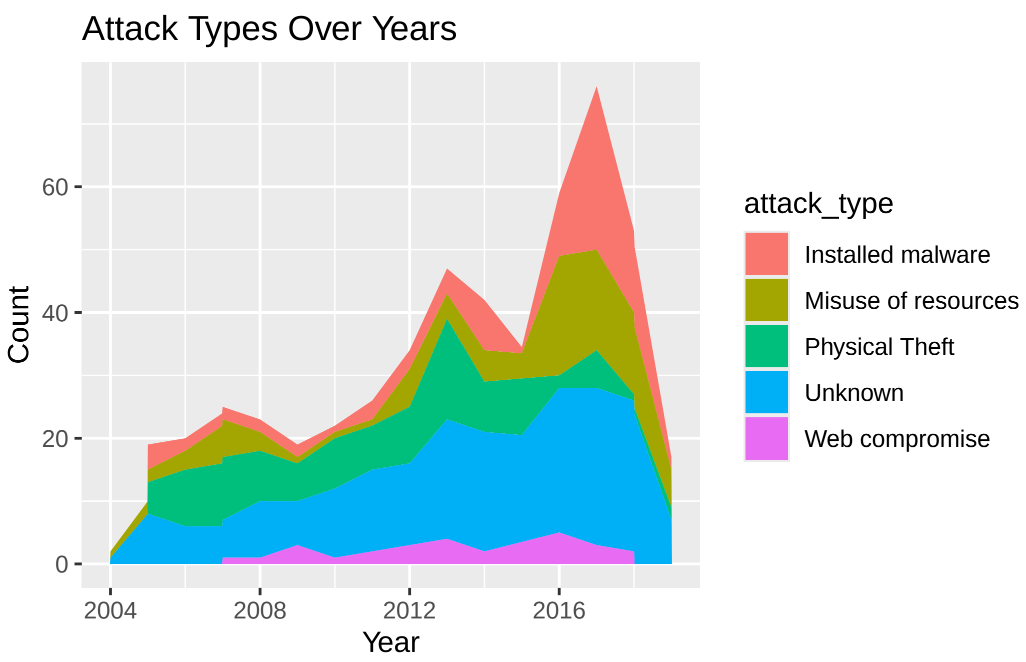
Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

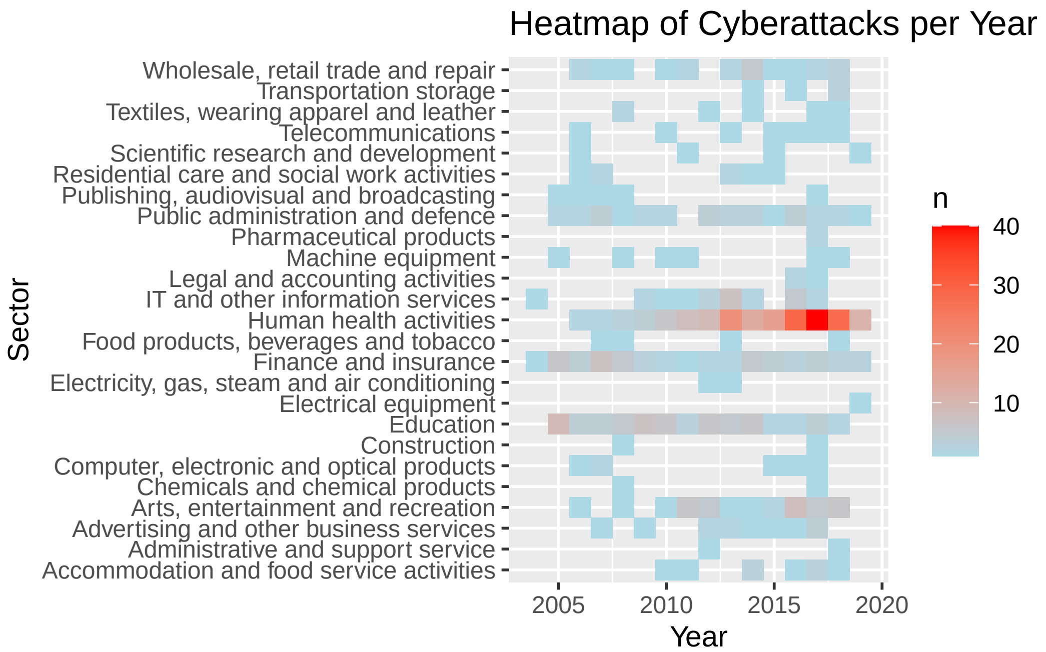
```
breach_data %>%
  ggplot(aes(x = as.factor(year), y = impact_on_data)) +
  geom_boxplot() +
  labs(title = "Impact on Data by Year", x = "Year", y = "Impact on Data")
```



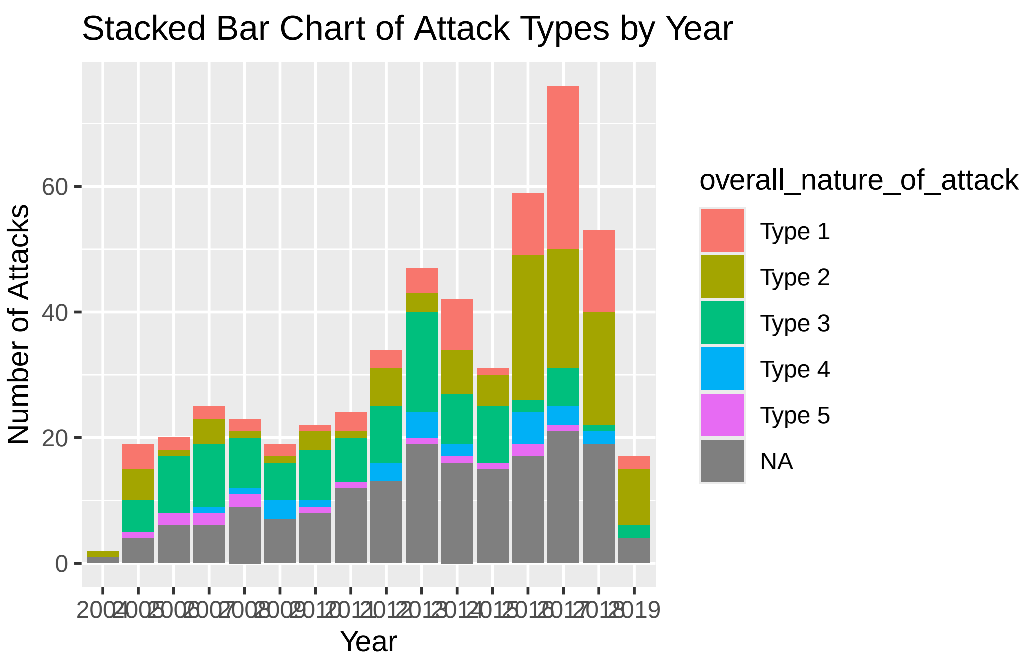

```
breach_data %>%
  count(year, attack_type) %>%
  ggplot(aes(x = year, y = n, fill = attack_type)) +
  geom_area(position = 'stack') +
  labs(title = "Attack Types Over Years", x = "Year", y = "Count")
```




```r
library(ggplot2)
library(dplyr)

# Assuming 'overall_nature_of_attack' is your categorical variable indicating the nature/type
# and 'number_of_users_affected' is a numeric variable indicating the number of users impacte

# First, let's calculate summary statistics for each nature of attack

breach_data %>%
  count(year, sector) %>%
  ggplot(aes(x = year, y = sector, fill = n)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "red") +
  labs(title = "Heatmap of Cyberattacks per Year and Sector",
       x = "Year",
       y = "Sector")
```

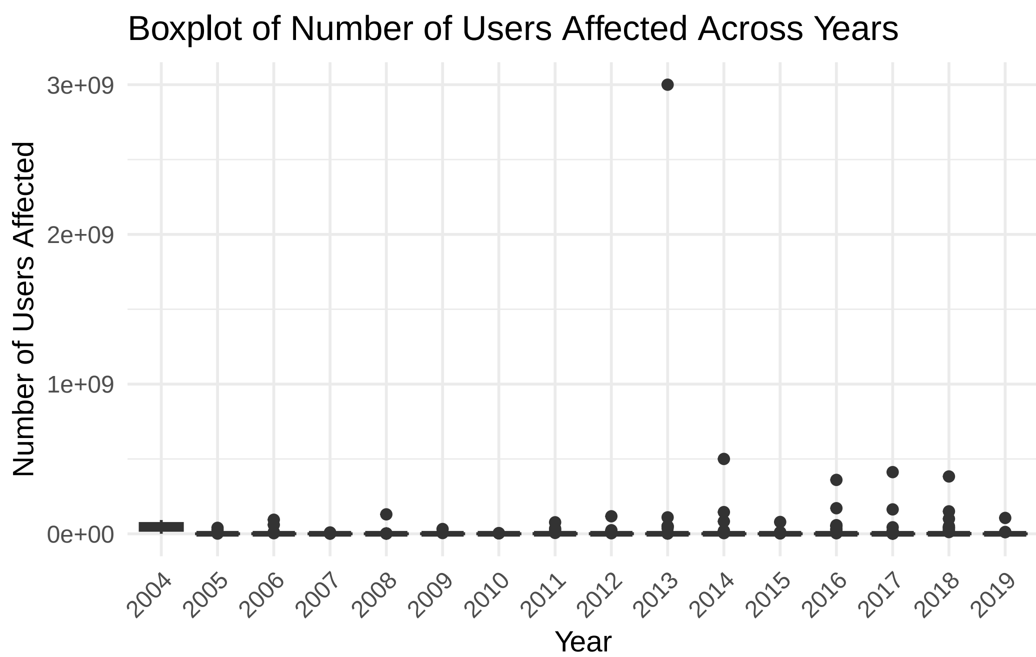



```r
breach_data %>%
  count(year, overall_nature_of_attack) %>%
  ggplot(aes(x = as.factor(year), y = n, fill = overall_nature_of_attack)) +
  geom_bar(stat = "identity") +
  labs(title = "Stacked Bar Chart of Attack Types by Year",
       x = "Year",
       y = "Number of Attacks")
```




```
library(ggplot2)

ggplot(breach_data, aes(x = as.factor(year), y = number_of_users_affected)) +
  geom_boxplot() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Boxplot of Number of Users Affected Across Years",
       x = "Year",
       y = "Number of Users Affected")
```

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




```
breach_data %>%
  filter(year == 2013) %>%
  arrange(desc(number_of_users_affected)) %>%
  head()  # This shows the top entries for 2013
```

```
  year  organisation critical_industry organisation_size
1 2013         Yahoo               Yes             Large
2 2013        Target                No             Large
3 2013      Evernote               Yes            Medium
4 2013 Living Social                No             Large
5 2013        Scribd               Yes            Medium
6 2013         Adobe               Yes             Large
  level_of_digital_intensity                                 sector country
1                       High       IT and other information services     USA
2                Medium-High      Wholesale, retail trade and repair     USA
3                       High       IT and other information services  Global
4                       High Advertising and other business services     USA
5                       High       IT and other information services     USA
6                       High       IT and other information services  Global
  cyber_security_role cyber_security_frameworks education_and_awareness_policy
1                 Yes                        No                             No
2                 Yes                        No                             No
```

```
3                 Yes                        No                              No
4                  No                        No                              No
5                  No                        No                              No
6                  No                        No                              No
  policy prevention_detection_and_recovery improper_network_segmentation
1    Yes                            Medium                          <NA>
2    Yes                              High                           Yes
3    Yes                            Medium                          <NA>
4    Yes                               Low                          <NA>
5    Yes                               Low                          <NA>
6    Yes                               Low                           Yes
  inappropriate_remote_access absence_of_encryption     detector
1                        <NA>                  <NA> Organisation
2                         Yes                   Yes Organisation
3                        <NA>                    No Organisation
4                        <NA>                   Yes Organisation
5                          No                   Yes Organisation
6                        <NA>                   Yes Organisation
  restructuring_after_attack bribe_ransom_paid
1                        Yes                No
2                        Yes                No
3                        Yes                No
4                         No                No
5                         No                No
6                        Yes                No
  free_identity_or_credit_theft_monitoring additional_disclosure_of_information
1                                      Yes                                  Yes
2                                      Yes                                  Yes
3                                       No                                  Yes
4                                       No                                  Yes
5                                       No                                   No
6                                       No                                   No
  number_of_users_affected overall_nature_of_attack       attack_type attacker
1                  3.0e+09                     <NA>           Unknown External
2                  1.1e+08                   Type 1 Installed malware External
3                  5.0e+07                     <NA>           Unknown External
4                  5.0e+07                     <NA>           Unknown External
5                  5.0e+07                     <NA>           Unknown External
6                  3.8e+07                     <NA>           Unknown External
           attack_vector impact_on_data
1                   <NA>            Low
2   Vendor vulnerability           High
3                   <NA>            Low
```

```
4                  <NA>            Low
5                  <NA>            Low
6 Unknown network attack          High
  aspect_of_confidentiality_integrity_availability_triad_affected
1                                                 Confidentiality
2                                                 Confidentiality
3                                                 Confidentiality
4                                                 Confidentiality
5                                                 Confidentiality
6                                                 Confidentiality
  individual_s_name_s_leaked_exposed address_es_leaked_exposed
1                                Yes                        No
2                                Yes                       Yes
3                                Yes                        No
4                                Yes                        No
5                                Yes                        No
6                                Yes                        No
  other_personally_identifiable_information_pii_leaked_exposed
1                                                          Yes
2                                                          Yes
3                                                          Yes
4                                                          Yes
5                                                          Yes
6                                                          Yes
  track_1_credit_card_details_leaked_exposed
1                                         No
2                                        Yes
3                                         No
4                                         No
5                                         No
6                                        Yes
  track_2_credit_card_details_leaked_exposed
1                                         No
2                                       <NA>
3                                         No
4                                         No
5                                         No
6                                       <NA>
  social_security_number_tax_number_leaked_exposed
1                                               No
2                                              Yes
3                                               No
4                                               No
```

```
5                                              No
6                                              No
  subsequent_fraudulent_use_of_data investigation undertook_investigation
1                                No           Yes                     Yes
2                               Yes           Yes                     Yes
3                                No            No                      No
4                                No            No                      No
5                                No           Yes                     Yes
6                                No           Yes                     Yes
  litigation_by_public penalties_settlement_paid_or_actions_imposed
1                  Yes                                          Yes
2                  Yes                                          Yes
3                   No                                           No
4                   No                                           No
5                   No                                           No
6                  Yes                                          Yes
  imposed_penalties_or_actions_on_organisation
1                                           No
2                                          Yes
3                                           No
4                                           No
5                                           No
6                                           No
  fines_issued_by_government_or_relevant_body settlement_paid
1                                          No             Yes
2                                         Yes             Yes
3                                          No              No
4                                          No              No
5                                          No              No
6                                          No             Yes
  effect_on_share_price                       summary
1                  <NA>                       Unknown
2                     ↓ Through vendor access, PoS target
3                  <NA>                       Unknown
4                  <NA>                       Unknown
5                  <NA>                       Unknown
6                     ↓  Found a backup server and raided
```

```
breach_data %>%
  filter(year == 2014) %>%
  arrange(desc(number_of_users_affected)) %>%
  head()  # This shows the top entries for 2013
```

```
  year        organisation critical_industry organisation_size
1 2014               Yahoo               Yes             Large
2 2014                eBay                No             Large
3 2014     JP Morgan Chase               Yes             Large
4 2014 Korea Credit Bureau               Yes           Unknown
5 2014            Experian               Yes             Large
6 2014         P.F. Changs               Yes             Large
  level_of_digital_intensity                                    sector
1                       High         IT and other information services
2                Medium-High        Wholesale, retail trade and repair
3                       High                    Finance and insurance
4                       High                    Finance and insurance
5                       High                    Finance and insurance
6                        Low Accommodation and food service activities
      country cyber_security_role cyber_security_frameworks
1         USA                 Yes                        No
2      Global                  No                        No
3         USA                  No                        No
4 South Korea                  No                        No
5          UK                  No                        No
6         USA                  No                        No
  education_and_awareness_policy policy prevention_detection_and_recovery
1                             No    Yes                               Low
2                             No    Yes                               Low
3                             No    Yes                               Low
4                             No    Yes                               Low
5                             No    Yes                               Low
6                             No    Yes                               Low
  improper_network_segmentation inappropriate_remote_access
1                           Yes                        <NA>
2                           Yes                          No
3                           Yes                          No
4                           Yes                          No
5                          <NA>                        <NA>
6                            No                          No
  absence_of_encryption       detector restructuring_after_attack
1                   Yes   Organisation                        Yes
2                   Yes   Organisation                        Yes
3                   Yes           <NA>                        Yes
4                   Yes         Public                        Yes
5                  <NA>   Organisation                       <NA>
6                   Yes Federal Agency                       <NA>
  bribe_ransom_paid free_identity_or_credit_theft_monitoring
```

```
1              No                                       Yes
2              No                                        No
3              No                                        No
4              No                                       Yes
5              No                                       Yes
6              No                                       Yes
  additional_disclosure_of_information number_of_users_affected
1                                  Yes                 5.00e+08
2                                   No                 1.45e+08
3                                  Yes                 8.30e+07
4                                  Yes                 2.00e+07
5                                 <NA>                 1.50e+07
6                                  Yes                 7.00e+06
  overall_nature_of_attack         attack_type attacker
1                   Type 2 Misuse of resources External
2                     <NA>             Unknown External
3                   Type 2 Misuse of resources External
4                   Type 5 Misuse of resources Internal
5                     <NA>             Unknown External
6                   Type 1   Installed malware External
                           attack_vector impact_on_data
1                     Social engineering            Low
2                                   <NA>           High
3 Insufficient authentication validation         Medium
4         Inappropriate use of privilege           High
5                 Unknown network attack         Medium
6                 Unknown network attack           High
  aspect_of_confidentiality_integrity_availability_triad_affected
1                                                 Confidentiality
2                                                 Confidentiality
3                                                 Confidentiality
4                                                 Confidentiality
5                                                 Confidentiality
6                                                 Confidentiality
  individual_s_name_s_leaked_exposed address_es_leaked_exposed
1                                Yes                        No
2                                Yes                       Yes
3                                Yes                       Yes
4                                Yes                        No
5                                Yes                        No
6                                Yes                       Yes
  other_personally_identifiable_information_pii_leaked_exposed
1                                                          Yes
```

```
2                                                          Yes
3                                                          Yes
4                                                          Yes
5                                                          Yes
6                                                          Yes
  track_1_credit_card_details_leaked_exposed
1                                         No
2                                        Yes
3                                         No
4                                        Yes
5                                         No
6                                        Yes
  track_2_credit_card_details_leaked_exposed
1                                         No
2                                       <NA>
3                                         No
4                                       <NA>
5                                         No
6                                       <NA>
  social_security_number_tax_number_leaked_exposed
1                                               No
2                                               No
3                                               No
4                                               No
5                                              Yes
6                                               No
  subsequent_fraudulent_use_of_data investigation undertook_investigation
1                                No           Yes                    Yes
2                                No           Yes                     No
3                                No           Yes                    Yes
4                                No           Yes                    Yes
5                                No           Yes                    Yes
6                                No           Yes                    Yes
  litigation_by_public penalties_settlement_paid_or_actions_imposed
1                  Yes                                          Yes
2                  Yes                                           No
3                   No                                           No
4                   No                                           No
5                   No                                           No
6                  Yes                                           No
  imposed_penalties_or_actions_on_organisation
1                                           No
2                                           No
```

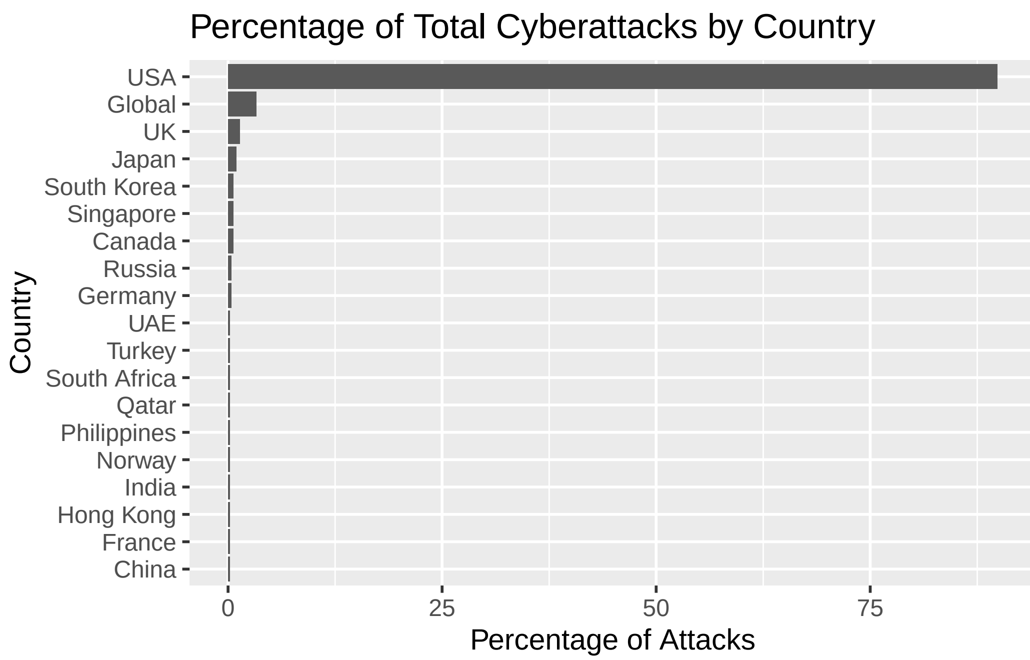
```
3                                          No
4                                          No
5                                          No
6                                          No
  fines_issued_by_government_or_relevant_body settlement_paid
1                                         No             Yes
2                                         No              No
3                                         No              No
4                                         No              No
5                                         No              No
6                                         No              No
  effect_on_share_price                                         summary
1                  <NA>                                  Spear phishing
2             No change                  Unauthorised access but unknown
3                     ↓ Server was hacked due to lack of proper authentication
4                  <NA>                               Contractor took data
5                     ↓                  Unauthorised access but unknown
6                  <NA>                            Malware installed on PoS
```

```r
# Install and load the scales package
if (!requireNamespace("scales", quietly = TRUE)) {
  install.packages("scales")
}
library(scales)
```

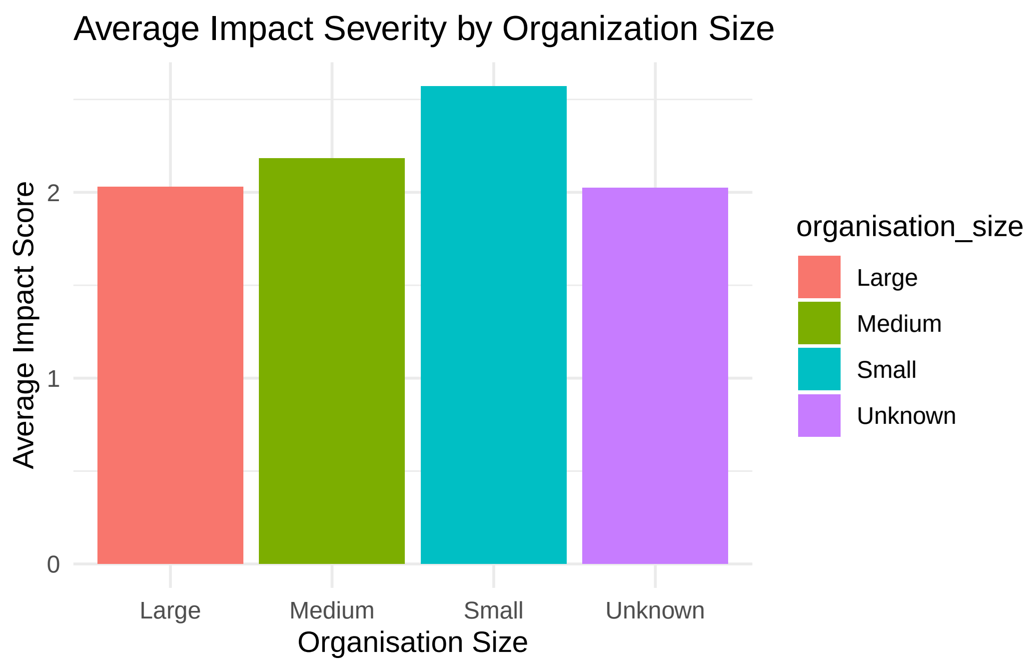
```

Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```r
breach_data %>%
  count(country) %>%
  mutate(percentage = n / sum(n) * 100) %>%
  ggplot(aes(x = reorder(country, percentage), y = percentage)) +
  geom_bar(stat = "identity") +
```

```
  coord_flip() +
  labs(title = "Percentage of Total Cyberattacks by Country",
       x = "Country",
       y = "Percentage of Attacks")
```

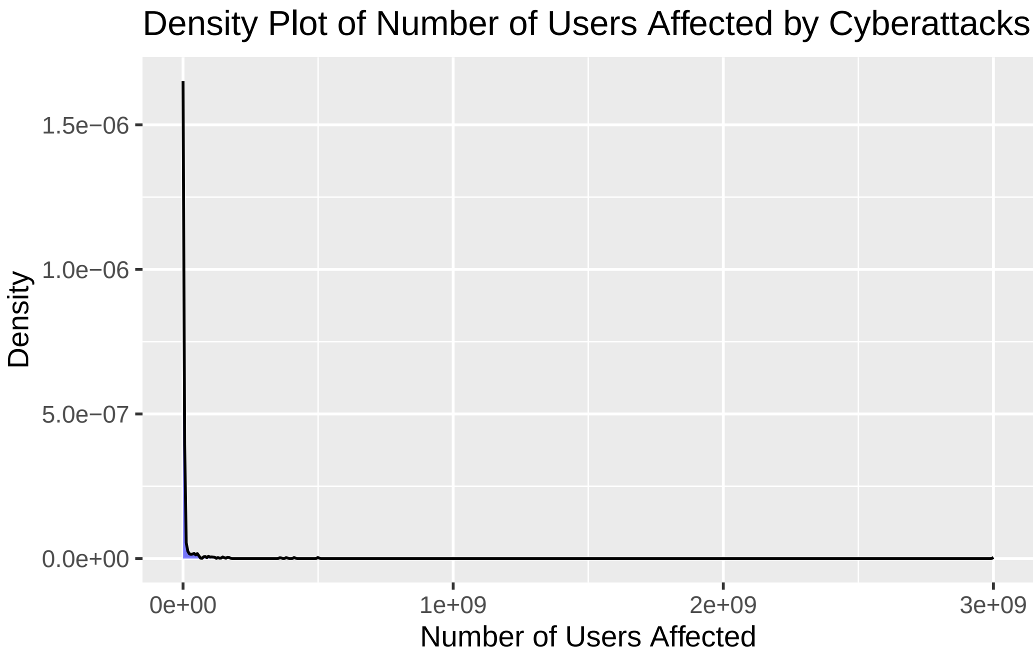



```
breach_data %>%
  mutate(impact_score = case_when(
    impact_on_data == "Low" ~ 1,
    impact_on_data == "Medium" ~ 2,
    impact_on_data == "High" ~ 3,
    TRUE ~ NA_real_
  )) %>%
  group_by(organisation_size) %>%
  summarize(average_impact = mean(impact_score, na.rm = TRUE)) %>%
  ggplot(aes(x = organisation_size, y = average_impact, fill = organisation_size)) +
  geom_col() +
  labs(title = "Average Impact Severity by Organization Size",
       x = "Organisation Size",
       y = "Average Impact Score") +
  theme_minimal()
```

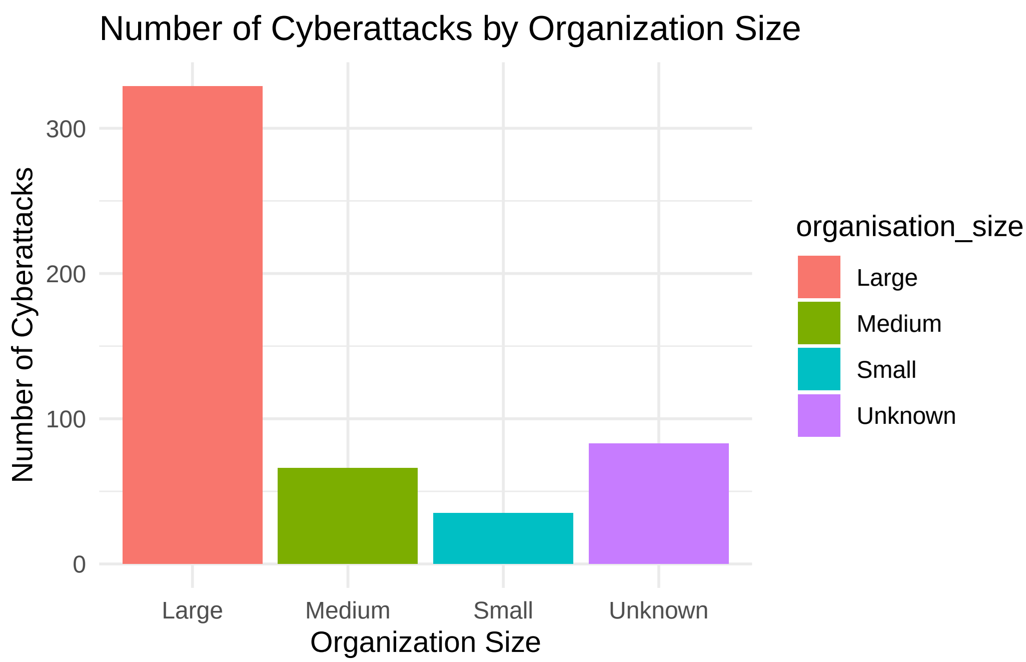



```r
ggplot(breach_data, aes(x = number_of_users_affected)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Density Plot of Number of Users Affected by Cyberattacks",
       x = "Number of Users Affected",
       y = "Density")
```

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




```
ggplot(breach_data, aes(x = organisation_size, fill = organisation_size)) +
  geom_bar() +
  labs(title = "Number of Cyberattacks by Organization Size",
       x = "Organization Size",
       y = "Number of Cyberattacks") +
  theme_minimal()
```




```r
library(modelsummary)
```

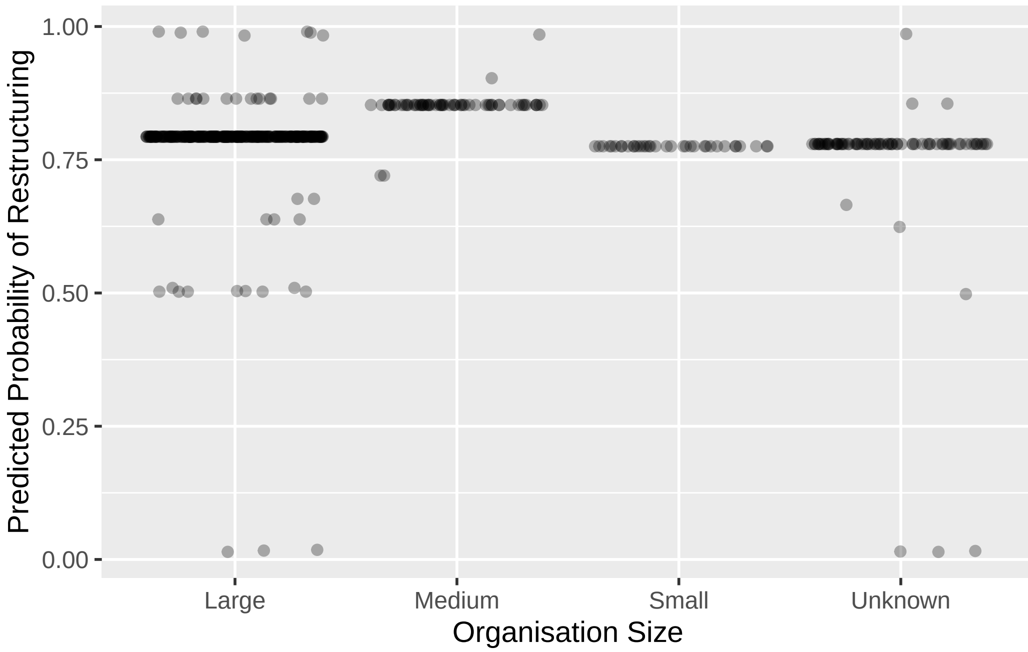
```
Version 2.0.0 of `modelsummary`, to be released soon, will introduce a
  breaking change: The default table-drawing package will be `tinytable`
  instead of `kableExtra`. All currently supported table-drawing packages
  will continue to be supported for the foreseeable future, including
  `kableExtra`, `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.
```

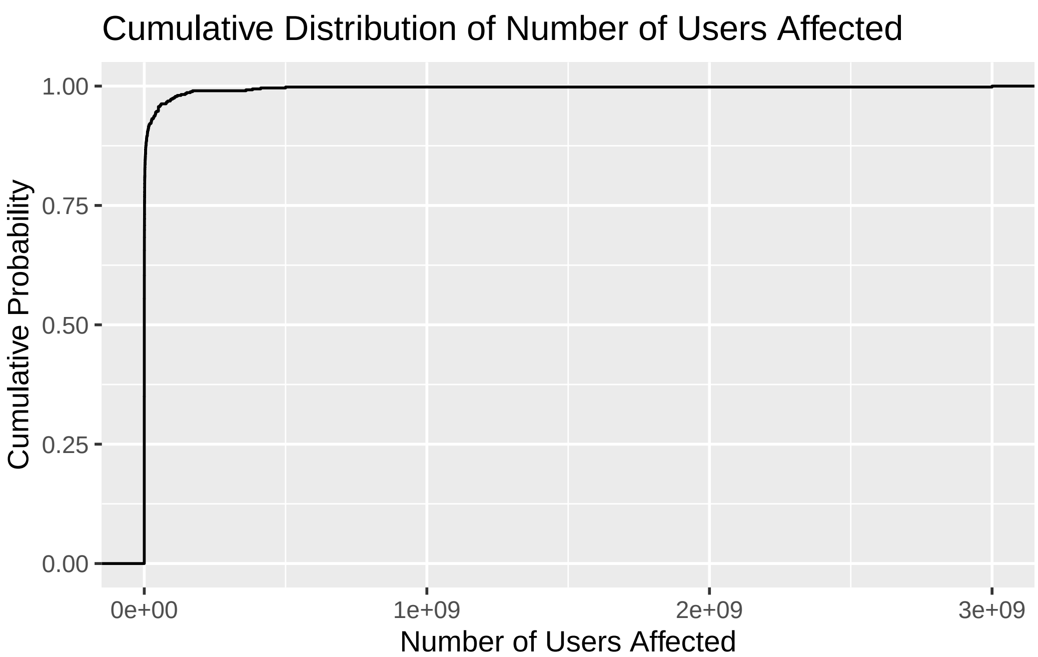
```r
breach_data <- breach_data %>% mutate(row_id = row_number())

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

	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

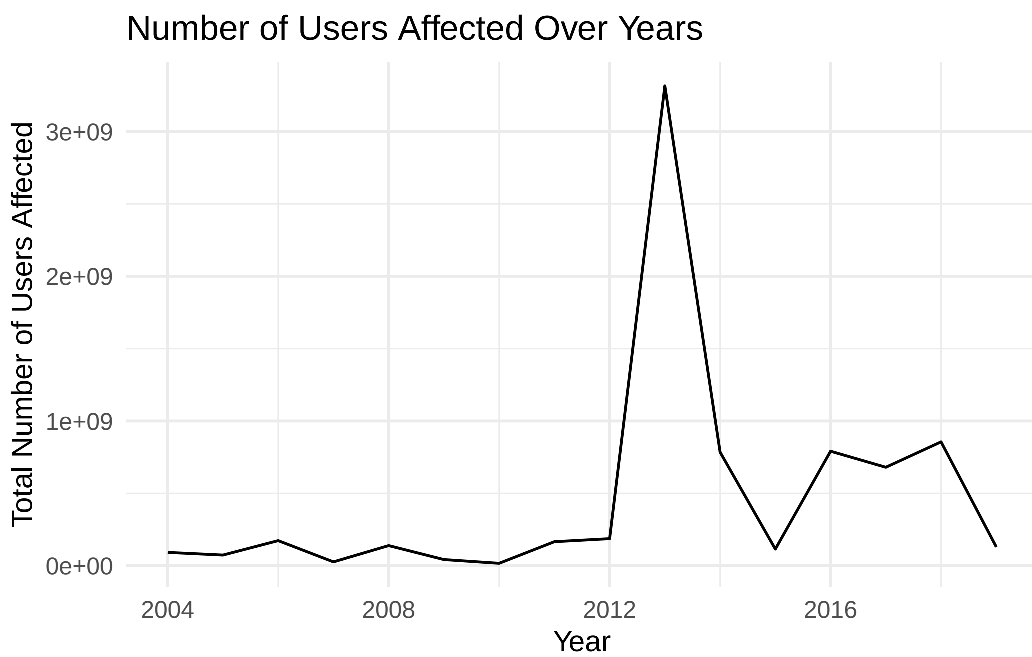



```r
ggplot(breach_data, aes(x = number_of_users_affected)) +
  stat_ecdf(geom = "step") +
  labs(title = "Cumulative Distribution of Number of Users Affected",
       x = "Number of Users Affected",
       y = "Cumulative Probability")
```

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

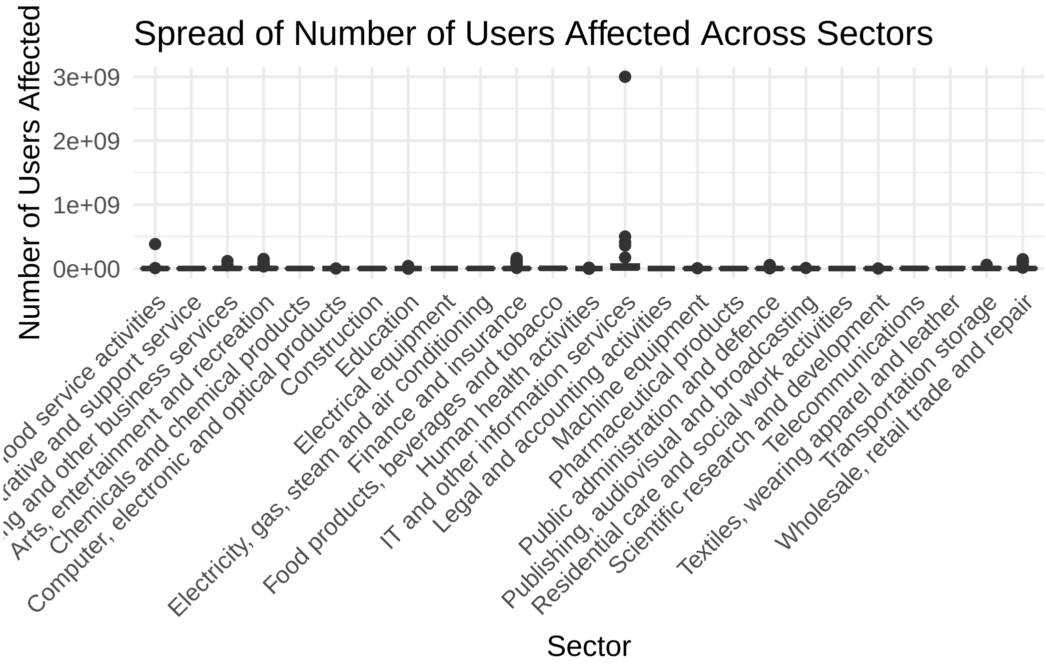



```r
breach_data %>%
  group_by(year) %>%
  summarize(total_users_affected = sum(number_of_users_affected, na.rm = TRUE)) %>%
  ggplot(aes(x = year, y = total_users_affected)) +
  geom_line() +
  labs(title = "Number of Users Affected Over Years",
       x = "Year",
       y = "Total Number of Users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



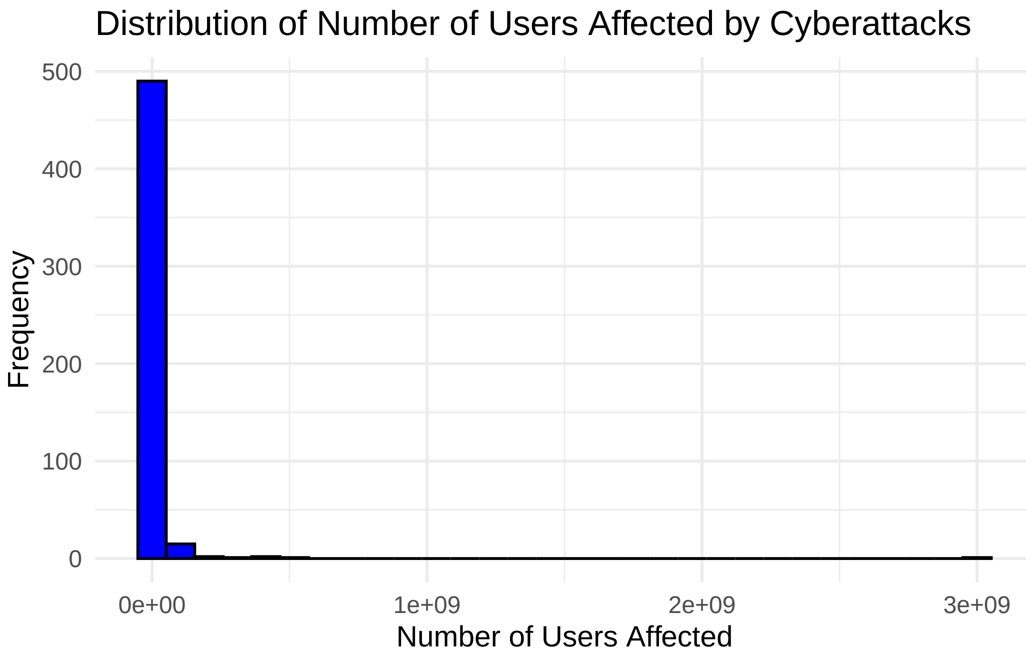

```r
breach_data %>%
  ggplot(aes(x = sector, y = number_of_users_affected)) +
  geom_boxplot() +
  labs(title = "Spread of Number of Users Affected Across Sectors",
       x = "Sector",
       y = "Number of Users Affected") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

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




```r
breach_data %>%
  ggplot(aes(x = number_of_users_affected)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  labs(title = "Distribution of Number of Users Affected by Cyberattacks",
       x = "Number of Users Affected",
       y = "Frequency") +
  theme_minimal()
```

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




```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
   inappropriate_remote_access absence_of_encryption       detector
1                         <NA>                  <NA>   Organisation
2                         <NA>                  <NA> Federal Agency
3                           No                   Yes   Organisation
4                           No                  <NA>   Organisation
```

```
5                          Yes                    Yes   Organisation
6                           No                     No   Organisation
7                           No                    Yes   Organisation
8                           No                    Yes   Organisation
9                           No                    Yes   Organisation
10                          No                    Yes   Organisation
11                        <NA>                    Yes   Organisation
12                          No                    Yes   Organisation
13                          No                    Yes   Organisation
14                          No                    Yes   Organisation
15                          No                    Yes           <NA>
16                        <NA>                   <NA>   Organisation
17                          No                    Yes   Organisation
   restructuring_after_attack bribe_ransom_paid
1                         Yes                No
2                         Yes                No
3                         Yes                No
4                         Yes                No
5                         Yes                No
6                         Yes                No
7                         Yes                No
8                        <NA>                No
9                         Yes                No
10                         No                No
11                         No                No
12                        Yes                No
13                       <NA>                No
14                        Yes                No
15                        Yes                No
16                        Yes                No
17                        Yes                No
   free_identity_or_credit_theft_monitoring
1                                       Yes
2                                       Yes
3                                       Yes
4                                        No
5                                       Yes
6                                      <NA>
7                                        No
8                                      <NA>
9                                        No
10                                      Yes
11                                     <NA>
```

```
12                                    Yes
13                                   <NA>
14                                    Yes
15                                   <NA>
16                                    Yes
17                                    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
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                                          No             Yes
15                                          No              No
16                                          No              No
17                                          No              No
   effect_on_share_price                                         summary row_id
1                   <NA>                                         Unknown     46
2              No change                             Unauthorised access     61
3                   <NA>                            Email phishing scam     72
4                   <NA>                                 Phishing attack    106
5                   <NA>                                      Ransomware    124
6                   <NA>          Unauthorised access to employee email    146
7                   <NA> American Medical Collection Agency data breach    201
8                   <NA>                               Hard drive stolen    219
9                   <NA>                                      Ransomware    234
10                  <NA>                                 Phishing attack    246
11                  <NA>                                         Unknown    256
12                  <NA>                Unauthorised access but unknown    312
13                  <NA>                    Dominon National was hacked    327
14                  <NA>                              AMCA vulnerability    335
15                  <NA>         Individual gained access through doctor    377
16                  <NA>                                 Phishing attack    441
17                  <NA>                                 Phishing attack    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
```

```
breach_data$year <- as.numeric(as.character(breach_data$year))

# Now proceed with your data manipulation
breach_data_summary <- breach_data %>%
  mutate(country = fct_collapse(country,
                                "Other" = setdiff(unique(country),
                                                   c("USA", "Australia", "Canada", "Global",
  group_by(year, country) %>%
  summarize(frequency = n(), .groups = 'drop')

# Separating the data by country for ease of plotting
country_data <- breach_data_summary %>%
```

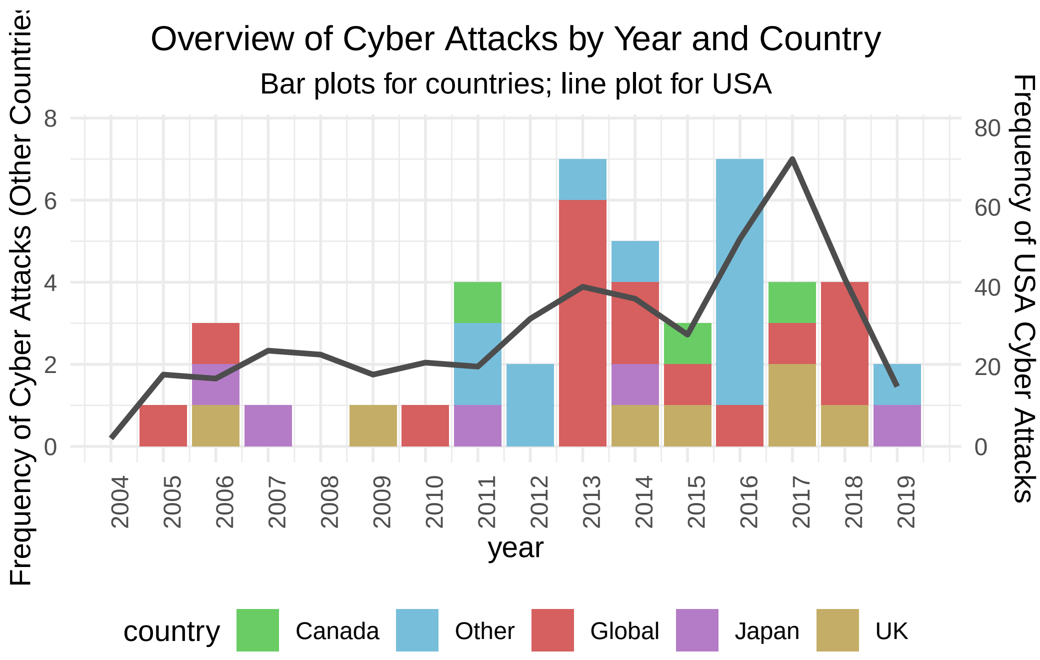
```r
  filter(country != "USA")

usa_data <- breach_data_summary %>%
  filter(country == "USA") %>%
  arrange(year)

# Ensure year is numeric for the line plot
breach_data_summary$year <- as.numeric(as.character(breach_data_summary$year))

# Recalculate max frequencies if necessary
max_usa_freq <- max(usa_data$frequency, na.rm = TRUE)
max_other_freq <- max(country_data$frequency, na.rm = TRUE)

# Determine scale factor for secondary axis
scale_factor <- max_other_freq / max_usa_freq
breach_data_summary$year <- as.numeric(as.character(breach_data_summary$year))
breach_data_summary <- breach_data_summary %>% filter(!is.na(year))

# Plot
# Define a color palette
formal_palette <- c("Australia" = "#4878D0", "Canada" = "#6ACC65", "Global" = "#D65F5F",
                    "Japan" = "#B47CC7", "UK" = "#C4AD66", "Other" = "#77BEDB")

# Plot
gg <- ggplot() +
  # Add bars for all countries except USA
  geom_bar(data = country_data, aes(x = year, y = frequency, fill = country), stat = "identi
  # Add line for USA
  geom_line(data = usa_data, aes(x = year, y = frequency * scale_factor, group = 1), color =
  # Define the primary y-axis with the secondary axis for the USA
  scale_y_continuous(
    name = "Frequency of Cyber Attacks (Other Countries)",
    limits = c(0, max_other_freq * 1.1),  # Set limits for better control, slightly above ma
    sec.axis = sec_axis(~ . / scale_factor, name = "Frequency of USA Cyber Attacks", labels =
  ) +
  # Set breaks for the x-axis to unique years
  scale_x_continuous(breaks = sort(unique(breach_data_summary$year))) +
  # Apply the formal color palette
  scale_fill_manual(values = formal_palette) +
  labs(title = "Overview of Cyber Attacks by Year and Country",
       subtitle = "Bar plots for countries; line plot for USA") +
  theme_minimal() +
```

```
  theme(axis.text.x = element_text(angle = 90, hjust = 1), # Rotate x-axis labels
        legend.position = "bottom", # Adjust the position of the legend
        plot.title = element_text(hjust = 0.5), # Center the plot title
        plot.subtitle = element_text(hjust = 0.5)) # Center the plot subtitle
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
# Print the plot
print(gg)
```

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



References

Alberta Government. 2023. "Air Indicators – Fine Particulate Matter." https://www.alberta.ca/air-indicators-fine-particulate-matter.

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.

Environment, Ministry of, and Protected Areas. 2021. "Status of Air Quality in Alberta." https://open.alberta.ca/dataset/9b00aab3-c37d-4080-854e-5f329c621b92/resource/057c65ac-7837-49bb-9528-38c2611540c4/download/epa-alberta-air-zones-report-2019-2021.pdf.

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/.

government, Alberta. 2023a. "Air Quality Index by Municipality." https://open.alberta.ca/opendata/air-quality-index-by-municipality#detailed.

———. 2023b. "Leading Causes of Death." https://open.alberta.ca/opendata/leading-causes-of-death.

———. 2024. "Alberta." https://www.alberta.ca/.

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.