**Report Title**

Student name1 and Instructor name1

1 Al Hussein Technical University, Amman 11831, Jordan

Your email@htu.edu.jo, [Instructor](mailto:raneem.qaddoura@htu.edu.jo) email@htu.edu.jo

**Abstract**

**Keywords:** First Keyword, Second Keyword, Third Keyword.

# Introduction

Two paragraphs about your topic (the topic of the dataset).

[1] point to Data breaches or data leak could be one of the serious danger of any business or organization also a source of concern regardless that danger on the data internal or external (hacking attacks). Normally the data consider one of the significant assets of any enterprise. A data breach allows unauthorized individuals to access private, sensitive, or protected information. In a data breach, the files are viewed and/or distributed without authorization.

Data breaches can occur due to various factors, including intentional attacks, oversights, or infrastructure flaws, rather than solely attributed to outside hackers.in this paper study the cases of the breaches that happened to point to the most main reason why these events could have happened. Furthermore, to be this paper as indicator for the enterprise to rearrange their security procedure.

One paragraph about how they use AI in your specific topic.

Jay Trinckes [2] explained that the Cyberattacks are one of the new dangers that AI may bring. Although AI is capable of accurately anticipating and thwarting threats, it may also be used as a weapon for ransomware and sophisticated phishing campaigns. It could be using for in red teaming in blue teaming Cybercriminals are too creating profiling strategies, foreseeing and abusing person behaviors for profoundly personalized assaults. As AI apparatuses gotten to be more complex, organizations must adjust their security measures to ensure delicate information. A comprehensive technique, counting encryption, get to controls, security appraisals, and fix administration conventions, is pivotal for tending to security and protection concerns. AI is being utilized as a vigorous protective instrument in compliance and security instruments, empowering proactive danger chasing, peculiarity location, and prescient approaches in businesses like healthcare. AI is being utilized as a vigorous protective instrument in compliance and security instruments, empowering proactive danger chasing.

One paragraph about your research question, and the objectives for this research, and illustrates your contributions clearly.

One paragraph describes the structure of the document. For example, it could be "Section 2 presents the related work. Section 3 discusses the data collection and description. Section 4 illustrates the proposed approach. Section 5 presents the experiments and results. Finally, Section 6 concludes the work and provides recommendation based on the conclusion."

# Related Work

Two lines introduction about the field of the research papers that will be illustrated in this section. For each paper write 1-3 paragraphs summary that includes the methodology of the paper, its main result, the limitation or weakness of the paper. At the end, add a paragraph about the gap in the literature review that is addressed in your research work.

[3] This paper reviews enterprise data breach causes, challenges, and future prevention strategies, focusing on the threats of intentional or unintentional data leaks that expose sensitive information like employee and customer data, intellectual property, and medical records. It analyzes recent high-profile incidents, such as Target’s 2013 breach and Yahoo’s 2014 breach, showing the substantial financial and reputational damage that data leaks cause. Research indicates that insiders account for over 40% of breaches, with motivations ranging from espionage to accidental errors. The paper explores various state-of-the-art prevention and detection techniques, such as Data Leak Prevention and Detection (DLPD) systems that monitor and control data flows using policies and behavior modeling. However, the big data era presents new challenges, as modern communication tools (e.g., cloud sharing, email, messaging) increase leak risks. Future directions for reducing data breaches include developing privacy-preserving detection systems, cloud service leak detection, and deep learning-based anomaly detection for insider threats, aiming to transition towards more robust and adaptable solutions in enterprise security.

[4] pointed to Computerized change in healthcare has progressed openness to treatment, but has moreover expanded helplessness to information breaches, with hacking/IT occurrences being the foremost common shape, taken after by unauthorized inside divulgence From 2005 to 2019, 249.09 million individuals were influenced and 157.4 million were influenced within the final five a long time alone.In 2018, the industry experienced 536 breaches, compromising 41.2 million records.The normal breach taken a toll in 2019 come to $3.92 million, a worldwide tall, reflecting a 12% increment since 2014. The think about connected straightforward moving normal (SMA) and straightforward exponential smoothing (SES) strategies to foresee data breach patterns utilizing information from the PRC database and HIPAA diaries, and found that SMA was more dependable. The comes about uncovered that 64% of restorative data spills since 2005 were due to hacking, and more than 92% of records have been compromised by hacking in later years.In 2019, hacking occurrences expanded by 73.4%, whereas other spills such as theft/loss and disgraceful transfer sorts diminished. The normal taken a toll per restorative breach rose to $6.45 million, and the fetched per record expanded 5.14%. The consider underscores the require for expanded security measures in healthcare and advocates proactive techniques to address the developing hazard and financial effect of information breaches.

[5] This study examines the prevalence of personal information leaks among Americans in data breaches that lack detailed research, particularly with respect to individual risk. Using a new dataset, the researchers estimated the minimum average number of online accounts compromised per capita by integrating data from a representative YouGov sample with information from Have I Been Pwned (HIBP), which catalogs 293 public information breaches. The results of the analysis of 5,000 e-mails revealed that 14,979 breaches occurred, with an average of three breaches per person and at least 82.84% of Americans experiencing at least one breach. Socioeconomic factors revealed some interesting trends. The frequency of breaches increases with education level, with the average number of breaches for individuals without a high school degree being 2.35, while the more educated are more likely to be breached. The study also found that middle-aged and older accounts face higher risks than younger and older users. Female accounts are 1.12 times more likely to be compromised than male accounts, black accounts are 3.12 times more likely to be compromised than male accounts, and white accounts are 3.16 times more likely to be compromised than white accounts. Of the 15,837 breaches examined, 94.58% were confirmed and one-third were classified as spam listings. The study emphasizes that the relationship between the frequency of breaches and factors such as education and age differs from traditional digital divide concerns, indicating that increased online activity is correlated with a higher risk of breaches, especially among more educated users.

[6] Explain more about Ransomware is a significant cybersecurity threat, causing data breaches and disruptions. However, many studies lack consideration of government strategies, industry guidelines, and cyber intelligence. A study evaluating 212 academic studies found that many were irrelevant to the current reality. The study proposed prioritizing data exfiltration over encryption, considering ransomware in a business-practical manner, and recommending collaboration with the industry to address this evolving threat.

This survey compiled the ransomware evolution history and applied Rogers' Innovation Adoption Curve, predicting the rise of destructive ransomware with espionage. It reviewed 212 academic studies and found that most research has become less relevant in the era of ransomware double extortion with data exfiltration. The survey proposed integrating ransomware risk management into organizational cybersecurity risk management, emphasizing government strategies, industry reports, guidelines, and cyber intelligence. It also discussed innovative research prospects, including generative AI, and suggested future research directions.

[7] Despite significant cybersecurity investments, companies face ongoing data breaches that have significant financial and reputational consequences. To protect digital assets and improve threat visibility, threat intelligence employing AI and machine learning is emerging, moving from reactive to proactive defense strategies. This evolution of threat intelligence aims to predict security threats by analyzing and integrating cyber data to gain insights tailored to an organization's unique risk landscape, enhancing both visibility and incident response. It is critical to distinguish between data loss (unintentional) and data breach (intentional), and data loss prevention (DLP) addresses the former while incident response planning addresses the latter, analyzing system logs for forensic purposes after a breach. Effective machine learning solutions in cybersecurity must be aligned with business goals and security standards, and must focus on specific threat scenarios with high-quality training data sets to maximize predictive accuracy and relevance. However, machine learning also introduces vulnerabilities to hostile attacks, so continuous learning is essential to adapt models to changes in the threat environment and mitigate concept drift. Continuous monitoring, root cause analysis (RCA), and improved mapping of infrastructure and threat interdependencies will continue to be essential in managing organization-specific risks and supporting continuous adaptation to the evolving cybersecurity landscape. Despite significant cybersecurity investments, companies face ongoing data breaches that have significant financial and reputational consequences. To protect digital assets and improve threat visibility, threat intelligence employing AI and machine learning is emerging, moving from reactive to proactive defense strategies.

This evolution of threat intelligence aims to predict security threats by analyzing and integrating cyber data to gain insights tailored to an organization's unique risk landscape, enhancing both visibility and incident response. It is critical to distinguish between data loss (unintentional) and data breach (intentional), and data loss prevention (DLP) addresses the former while incident response planning addresses the latter, analyzing system logs for forensic purposes after a breach. Effective machine learning solutions in cybersecurity must be aligned with business goals and security standards, and must focus on specific threat scenarios with high-quality training data sets to maximize predictive accuracy and relevance. However, machine learning also introduces vulnerabilities to hostile attacks, so continuous learning is essential to adapt models to changes in the threat environment and mitigate concept drift. Continuous monitoring, root cause analysis (RCA), and improved mapping of infrastructure and threat interdependencies will continue to be essential in managing organization-specific risks and supporting continuous adaptation to the evolving cybersecurity landscape.

# Data collection and Description

Define the primary and secondary data and describe the differences between them.

3.1 Primary Data

Describe your primary data (it could include figures and tables).

Describe your experience in collecting your primary data. Include any issues you faced in this process.

Justify your choice.

Describe the merits and limits of primary data.

**Description of primary data:**

Primary data was obtained from an online self-administered survey distributed to cybersecurity workers and cybersecurity specialized people who have dealt with events such as data breach. The survey was shared through two main channels:

Direct Outreach: The emails were sent to cybersecurity enthusiasts who include cybersecurity academicians and cybersecurity industry professionals.

Social Media Groups: The survey was shared in professional cybersecurity groups, cybersecurity conferences, cybersecurity group in LinkedIn and other such platforms where cybersecurity professionals are active.

The data collected aimed to gather insights on:

* More to the point, the frequency and type of data breach incidents are changing.
* Current defense mechanisms.
* People’s perception of the suitability of primitive measures.
* New risks that contribute to the data breach management.

**Describe about experience in collecting the primary data:**

Outreach Issues:

I faced challenge in accessing cybersecurity personnel through email particularly individuals with field experiences in handling real cyber security instances. A number of these experts could not be reached or chose not to respond, thus reducing the range of input from this focus group.

Trust Barrier on Social Media:

Using survey links on social networks was an issue of concern when it came to trust. Some people refrained from clicking the links presumably because of the phishing fears or personal data protection.

Low Response Rate:

In general, the response rate was not as high as desired even though the survey followed a logical design, some key targeted groups such as high rank professionals might not have found time to be more responsive to the survey questions.

**Justify the choice:**

The survey method was chosen for the following reasons:

* Relevance: The most ideal people to seek insight on data breaches, their underlying reasons and mitigation strategies are cybersecurity experts. Direct input from the target audience was provided by surveys.
* Flexibility: Both qualitative and quantitative data in surveys can be captured, thus surveys can be used to analyze cybersecurity incident trends and patterns.
* Cost and Time Efficiency: Unlike interviews or focus groups surveys are more scalable and quicker to collect a large group.

**Merits of Primary Data:**

Specific to the Research Objective:

When tailored to the research problem, primary data is relevant and specific to the problem.

Up-to-Date Information:

The survey acts as an instant source of current challenges and strategies in cybersecurity.

Original Insights:

This is because primary data collection allows the researcher to come up with unique findings by accessing the source directly as you would not necessarily get it in the secondary data.

**Limits of Primary Data:**

Time consuming:

Designing, disseminating and collecting survey responses consumed large amount of time, in particular when response rate was low.

Potential Bias:

This means that respondents may not always give honest or accurate answers when asked and this could lead to biases in the data.

Access Challenges:

One of the biggest constraints was reaching experienced cybersecurity professionals and convincing them to trust us.

Limited Sample Size:

Low participation rates mean that the sample size is no representative of the broader cybersecurity community.

**The questions: note: the yellow highlight is check box question.**

**Section: Data breach trends.**

1. **In your professional experience, what is the most common cause of data breaches in the industries you specialize in?**
   * A) Phishing attacks
   * B) Insider threats
   * C) Misconfigured systems
   * D) Weak or stolen credentials
   * E) Other

**(that is the main cause of the data breaches)**

1. **How frequently do organizations under your expertise report data breaches?**
   * A) Weekly
   * B) Monthly
   * C) Quarterly
   * D) Annually
   * E) Rarely

**(the frequently of the data breaches incidents)**

1. **What type of data is most commonly targeted in breaches you’ve investigated?**
   * A) Personal Identifiable Information (PII)
   * B) Financial data
   * C) Intellectual property
   * D) Healthcare records
   * E) Other

(the consequence of the breaches and most targeted get effected)

**Section: data breach prevention.**

1. **Which security measure do you find most effective in preventing data breaches?**
   * A) Regular security audits
   * B) Multi-factor authentication (MFA)
   * C) Encryption of sensitive data
   * D) Continuous monitoring and detection systems
   * E) Other
2. **In your professional opinion, how effective are organizations at identifying breaches in their early stages?**
   * A) Very effective
   * B) Moderately effective
   * C) Somewhat effective
   * D) Ineffective
   * E) Other

**Section: data breach impact and response.**

1. **What do you observe as the biggest consequence of data breaches for organizations?**
   * A) Financial loss
   * B) Reputational damage
   * C) Regulatory fines
   * D) Loss of customer trust
   * E) Other
2. **How prepared are organizations to respond effectively to a major data breach?**
   * A) Very prepared
   * B) Moderately prepared
   * C) Somewhat prepared
   * D) Not prepared

**Section: data breach impact and response.**

1. **What area should organizations invest in most to reduce the risk of data breaches?**
   * A) Employee training on cybersecurity awareness
   * B) Advanced threat detection tools
   * C) Incident response planning
   * D) Regular vulnerability assessments and penetration testing
   * E) Other

.

3.2 Secondary Data

Describe your secondary data (it could include figures and tables).

Justify your choice.

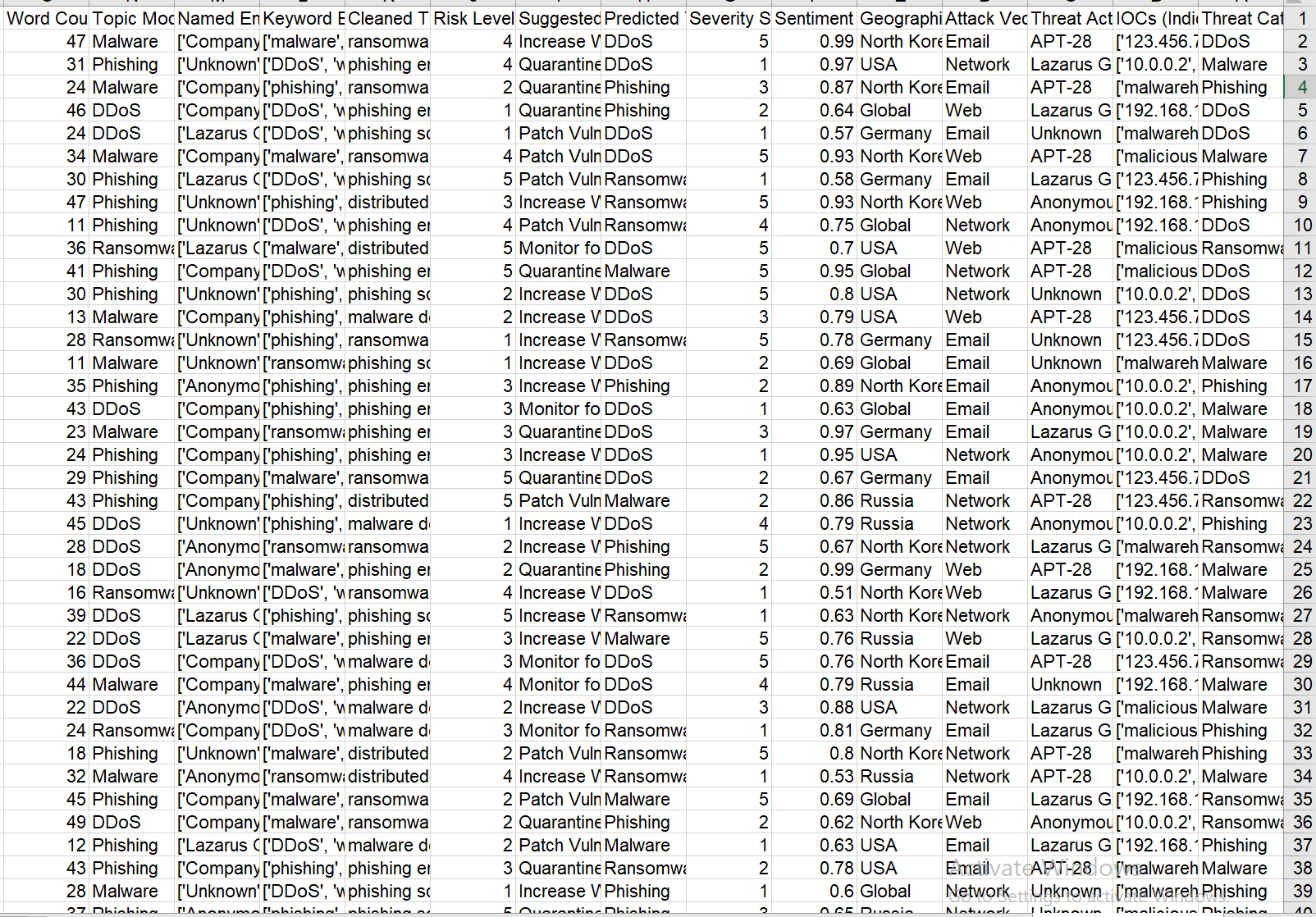
Describe the merits and limits of secondary data.

About the datasets:

[8] explained In order to facilitate Cyber Threat Intelligence (CTI) research, this dataset comprises 1100 examples of cyber threat reports that have been augmented with Natural Language Processing (NLP) features. The collection consists of textual descriptions, indications of compromise (IOCs), and several danger classifications. With features to recognize, categorize, and rate the seriousness of cybersecurity risks, each event is a distinct danger report.

|  |  |  |
| --- | --- | --- |
| Feature Name | Feature Description | Expected Values |
| Threat Category | Describes the type of threat | ['Phishing', 'Ransomware', 'DDoS', 'Malware'] |
| IOCs (Indicators of Compromise) | Lists known indicators of compromise | List of strings, e.g. ['192.168.1.1', 'malicious.com'] |
| Threat Actor | The group or entity behind the attack | ['Unknown', 'APT-28', 'Anonymous', 'Lazarus Group'] |
| Attack Vector | The method used to execute the attack | ['Email', 'Network', 'Web'] |
| Geographical Location | The country or region where the attack originated or was targeted | ['USA', 'Russia', 'Global', 'North Korea', 'Germany'] |
| Sentiment in Forums | Sentiment analysis score (0.5 to 1.0) from hacker forums | Float value between 0.5 and 1.0 |
| Severity Score | A severity rating from 1 (low risk) to 5 (critical) | Integer between 1 and 5 |
| Predicted Threat Category | The predicted category of the threat based on model analysis | Same as Threat Category feature |
| Suggested Defense Mechanism | The recommended action to mitigate the threat | ['Monitor for Phishing', 'Patch Vulnerability', 'Increase Web Security', 'Quarantine'] |
| Risk Level Prediction | A prediction of the overall risk level (1 to 5) | Integer between 1 and 5 |
| Cleaned Threat Description | A cleaned and preprocessed description of the threat | String |
| Keyword Extraction | Important keywords extracted from the description | List of strings, e.g. ['phishing', 'link', 'email'] |
| Named Entities (NER) | Entities identified within the description (organizations, malware names) | List of strings, e.g. ['CompanyX', 'APT-28'] |
| Topic Modeling Labels | Labels assigned based on topic modeling (e.g., LDA) | Same as Threat Category feature |
| Word Count | The number of words in the cleaned threat description | Integer |

Screenshot of the datasets:



[The datasets link](https://www.kaggle.com/datasets/hussainsheikh03/nlp-based-cyber-security-dataset)

**Choosing justification:**

1\_ Diversity of Threats: This dataset also contains many types of threats like DDoS, phishing, malware, and ransomware, which can be used to analyze different attack vector and pattern.

2\_Rich Metadata: It includes those key attributes such as:

* IOCs.
* Attack vector and threat actors.
* Sentiment analysis and geographical impact.
* Predicted risk levels and suggested defense mechanisms.

3\_Actionable Insights:

* It offers clear links between threats to Java programs and suggested mitigation strategies.
* It enables detect threat severity sequences and geographical trends.

4\_Relevance to Research:

* It provides support to understand attack methodologies and organization vulnerabilities.
* It contributes significantly in shaping the practical response to the problem of threat prevention and response.

**Merits of Secondary Data**

1\_Cost-Effective:

Pre collected and organized, secondary data does not have to be gathered at resource intensive levels.

2\_Time-Saving:

Immediate analysis is possible with fast access to a well-structured dataset.

3\_Broad Scope:

Data covers the different aspects of a breach covered in full.

4\_High Reliability:

It looked like some of the data is collected systematically in respect to technical and contextual details.

5\_Data Availability:

It contains a mix of quantitative (severity score, risk levels) and qualitative (descriptions and keywords cleaned).

**Limitations of Secondary Data**

1\_Data Quality Issues:

The fields such as "Named Entities" contain vague entries like "Unknown," that makes analysis less precise.

2\_Lack of Context:

With secondary data, you can never be fully sure to know how the data gets collected, and that can create gaps in your understanding.

3\_Limited Customization:

Specific research needs may not match pre-defined structures and variables.

4\_Outdated Information:

The dataset is updated only if the dataset doesn’t match with recent trends or new attack methods.

5\_Bias in Data Collection:

Possible biases in selection of threat types, sources, or regions; lack of applicability to global scale.

# Research Approach and Methodologies

1 line illustrates what you will include in this section. talk about the onion mode

To present the methodological research approach and tools employed in the study, the Research Onion Model will be used and What, Why, how approaches will be outlined, as well as several layers, such as research philosophy, research strategy, time horizon, and data collection techniques.

### Onion Model

Describe the onion model.

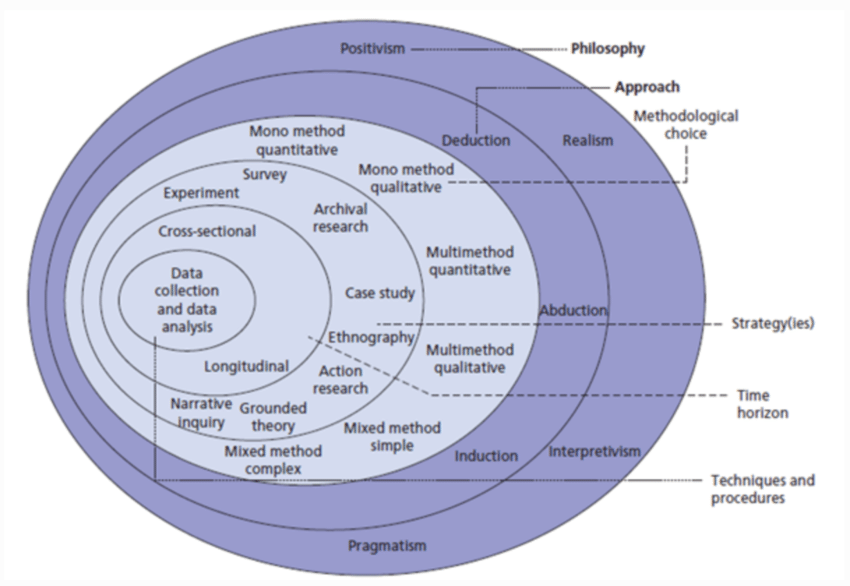


Figure 1: The structure of the Onion Research Model [1].

* + 1. **Philosophy**

Discuss different research approaches and methodology for the Philosophy layer. Provide justification of your choice.

**Positivism**: Designed with objective data collection and the responsibility of measurable outcomes, making it a good candidate for analyzing cybersecurity incidents (such as data breaches).

**Justification and reasons:**

Objective Approach: Positivism is something that seeks data to measure, analyze, and study. This data breach survey is also appropriate because it seeks to gather factual data from cybersecurity professionals.

Quantitative Nature: As is the case of numerical data (e.g., survey results and the performance metrics from machine learning models), positivism supports their usage because it involves statistical analysis and objective truths.

Scientific Testing: Hypothesis testing can be endorsed by positivism (in deploying inquiry into the causes or patterns of data breaches, for example). Collecting data and analysis through research is your holy grail here so you are hoping to validate these hypotheses.

Generalizability: Positivism enables you to analyze data systematically, and reach conclusions of general applicability that can be applied to all of cyber security best practice, like recommendations, for example.

* + 1. **Theory Development Approach**

Discuss different research approaches and methodology for the Development Approach layer.

Provide justification of your choice.

**How is the Deductive Approach used?**

Data breaches are initially fallen into existing studies on cybersecurity vulnerabilities, human errors, or technical flaws with deductive approach, under general theories or frameworks about data breaches.

Then you run these theories against your collected data (your survey responses) to find out which exact causes apply in your case.

**Justification for Deductive Approach in Your Research:**

Existing Theories: There is such a lot of literature and theoretical frameworks about data breaches ranging from the topics area of IT risk management, Information risk management, Security risk management, Information security management, Data security management, Data privacy management, Information privacy management, and Data encryption management.

Human factor causes, as in (e.g., phishing, poor password hygiene).

Related to system vulnerabilities (old software or no encryption).

Testing Hypotheses: For instance, you can postulate that

"Human error is the leading cause of data breaches."  
"Organizations with outdated software are more prone to data breaches."

Survey Data: As the deductive approach allows you to systematically test these hypotheses with your survey responses, it allows you to test them systematically.

Quantifiable Results: You can quantitatively validate or refute the proposed causes of your problem by analyzing your survey results.

**Steps in the Deductive Approach for Your Research:**

Step 1: Approach to Data Breach Causes: Identify theorize existent theories for causes of data breach.

Example: It could be from ISO 27001 standards, Verizon Data Breach Investigations Report or from academic research on cybersecurity incidents.

Step 2: Defining Hypotheses: According to these theories it is possible to define hypotheses, for instance:

"The human errors are the source of most data breaches."

"In fact, the most commonly cited cause of breaches is technical vulnerabilities."

Step 3: Hypothesis: These hypotheses are validated through survey design of a survey aimed at cybersecurity experts.

Step 4: Statistical Methods of Data Collection and Analysis: How to use statistical methods to interpret survey data and focus on which causes are most important.

Step 5: Findings are used to conclude to either confirm or reify the hypotheses and provide a more nuanced understanding of the data breach causes.

* + 1. **Methodological Choice**

Discuss different research approaches and methodology for the Methodological layer. Provide justification of your choice.

**Qualitative vs. Quantitative Research: A Comprehensive Comparison**

Qualitative and doing Quantitative research methods serve as important tools of various fields such as Science, Social Science, and Business. Each provides somewhat different ways to collect and analyze data, strengths and limitations. Here's a comprehensive comparison:

|  |  |  |
| --- | --- | --- |
| Feature | Qualitative Research | Quantitative Research |
| Purpose | Exploration, understanding | Measurement, testing |
| Data Type | Non-numerical (words, images, etc.) | Numerical |
| Methods | Interviews, observations, case studies | Surveys, experiments, statistical analysis |
| Analysis | Interpretation, thematic analysis | Statistical analysis |
| Sample Size | Small | Large |
| Focus | In-depth understanding, context | Generalizability, objectivity |

**Mixed Methods Research:**

However, it's important to note here that the research is not exchangeable or cannot be done using only one or the other types. As a combination of approaches, mixed methods research integrates the two in order to provide a more holistic view of a research problem. This approach promises the depth of the qualitative data as it also offers generalizability of the quantitative data.

**Mono Method Quantitative**: The use of surveys for the purpose of numerical data collection and statistical analysis.

**Justification**: It provides the measured insights of the trends and patterns for cybersecurity incidents.

Because it is efficient and easily scalable, and can collect both quantitative and qualitative data, it is the perfect survey method to research the causes of data breach. It allows for data collection of structured and consistent type from a large crowd thus allowing for diverse and reliable insights. Anonymity in surveys also encourages response to a particularly sensitive topic like that. To overcome trust constraints and accessibility challenges, I distributed surveys online to social media and professional platforms where experts on the topic participated. Despite this, surveys are advantageous due to the fact that they are time saving and generalize more extensively than interviews or case studies and are convenient to do.

* + 1. **Research Strategy**

Discuss different research approaches and methodology for the Strategy layer.

Provide justification of your choice.

**Research Strategy: Survey**

Firsthand data on the causes of data breaches was collected using the survey strategy as the most proper research method from cybersecurity professionals and experts. With a structured and efficient approach to ask questions on a small sample yet a large and diverse audience, surveys enable you to understand common trends, opinions and experiences on data breaches.

**Surveys are particularly effective for this research for several reasons:**

* Efficiency in Data Collection: One of the benefits of surveys over one on one interviews or focus groups is the ability to upscale the questions and have them collected from the same number of people, saving an enormous amount of time. If you want to reach a big group of cybersecurity professionals around the globe, it’s important to do this.
* Anonymity and Honesty: With anonymity, participants are more likely to give honest feedback on such topics as data breaches. With this, we are able to collect unbiased and authentic data.
* Scalability: Surveys can be conducted with even larger sample size, and better data, through online platforms. For example, it’s particularly useful to get insights from across the board in various regions and business areas.
* Structured Insights: If you use predefined questions, you have consistency in how your data was collected and they are easier to analyze and compare the results. In addition, the structured format allows for the capture of quantitative information (e.g., number of breaches), and qualitative information (e.g., causes of breaches) in an efficient manner.
* Cost-Effective: When resource constraints are an issue, conducting surveys (especially online) is far simpler than, say, a case study or an interview.
* Relevance to the Study's Objective: Based on the research aim to identify common causes and patterns in data breaches, the survey directly entails the research aim. It allows the great real world insights from the people that actually have experienced it firsthand.
  + 1. **Time Horizons**

Discuss different research approaches and methodology for the Time Horizons layer. Provide justification of your choice.

**Time Horizons: Cross-Sectional**

In the research, a cross sectional time horizon is assumed which is based upon collecting information in one point in time. That being said, this method of data breach analysis is best suited to determine what is causing data breaches today.

**Justification:**

* Understanding Current Trends: A cross-sectional study captures immediate data breaches experienced by cybersecurity professionals, providing insight into contemporary challenges, attack patterns, and organizational vulnerabilities.
* Efficiency: Longitudinal studies suffer from long observation to correlate with present day and cross sectional data collection reduces time and focus on present day issues.
* Resource Constraints: Cross sectional surveys are less time consuming than longitudinal studies because the later requires time, manpower and capital to accomplish therefore practical for research that has limited capital and time to accomplish research projects.
* Alignment with Survey Methodology: Surveys fit well with cross-sectional research since they establish a point-in-time data point that the participants can discuss, which is the goal of the study since it seeks to identify the causes of data breaches.
* Focused Analysis: Concentrating on a specific timeframe enables research to create clear and definitive, and therefore actionable, results, which correspond to the state of play in the cybersecurity field at a particular point in time. It helps in finding action able patterns and recommendation without complexities of monitoring changes over time.
  + 1. **Techniques and Procedures**

Discuss different research approaches and methodology for the Techniques and Procedures layer.

Provide justification of your choice.

Techniques Used:

Survey Creation and Distribution:

Data breaches are gathered from surveys designed to tap directly into the experiences of cybersecurity professionals and subject matter experts.

It is distributed to platforms like social media and expert groups for very wide reach, and the responses are extremely varied.

Data Exploration with Power BI:

Why Power BI?

The collected data is explored with tools in Power BI to visualize trends, correlations and understand the reasons why data breaches happen. With its easy to use dashboard and interactive features, it’s simple to identify patterns, anomalies and relationships in the data.

Benefits:

It gives detailed charts and graphs to break down factors like geographical trends, common attack vectors and activities of threat actor.

It simplifies the identification of key issues in large datasets and brings immediate visual insights.

Modeling with Orange (Machine Learning):

Why Orange?

Orange is a powerful, user friendly data mining and machine learning tool, which, when applied to survey data analysis, allows one to discover the common causes of data breach. This allows for easy to use visual workflow, with robust machine learning tools for predictive analysis.

Process:

It cleans and pre processes the data.

Common causes of data breach are classified and predicted with various algorithms such as Random Forest, SVM, XGBoost.

Output:

Insights into what are likely causes are offered by models, which help to tell us the things that may trend, and what ways we can defend against them.

Justification of Choice:

Survey Creation and Distribution:

Surveys are considered to be efficient tools for collecting targeted, primary data from the right people, i.e., those with relevant experience.

This all enables fast and scalable data collection, with a wide audience of cybersecurity professionals.

Exploration with Power BI:

* Ease of Use: The visualization is simplified by Power BI, even the users who are not technical can use to visualize.
* Actionable Insights: The immediate identification of common themes through visualizations forms a basis for more in depth analysis.
* Customizable Reports: It can also be filtered and segmented based on attributes such as demographics, attack vector.

Modeling with Orange:

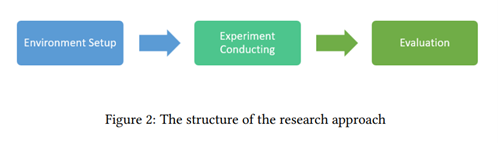
* User-Friendly Interface: Orange adopts drag & drop workflows that allow you to build a machine learning model with no extensive coding knowledge.
* Predictive Power: However, the paper goes beyond descriptive analysis and utilizes machine learning to devise predictions about patterns and causes of data breaches.
* Reproducibility: Orange models are also highly replicable, which means that future researchers can easily built upon this work.

Comprehensive Analysis:

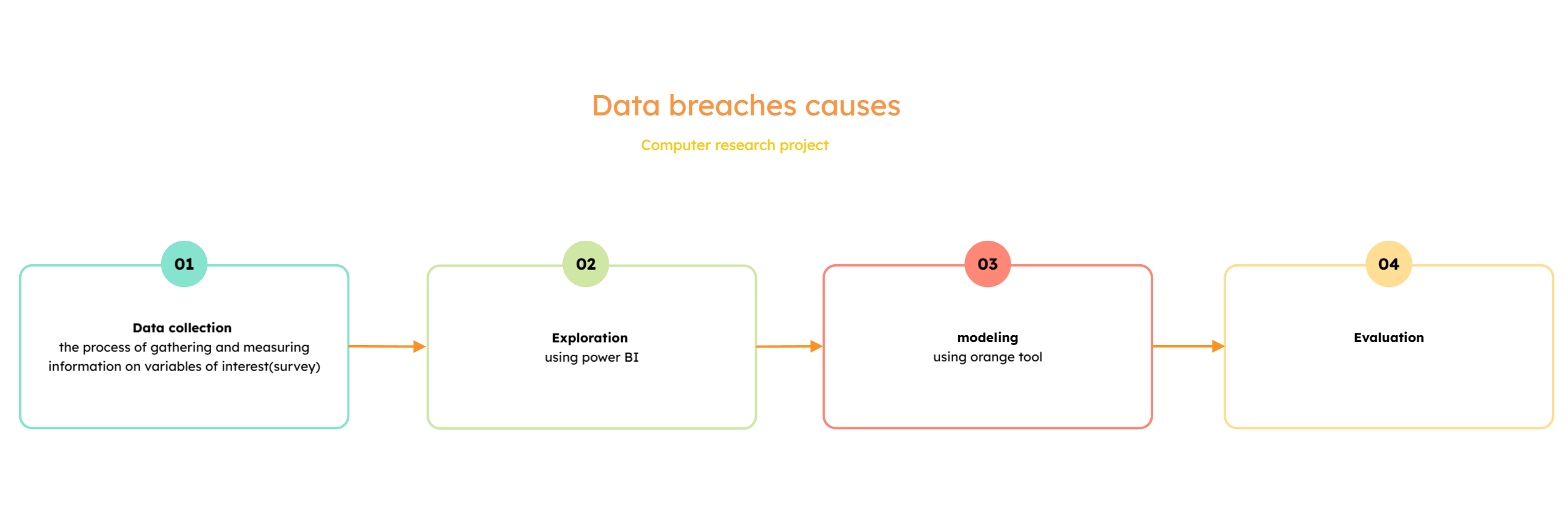
This approach utilizes survey data combined with the exploration techniques powered by Power BI and machine learning with Orange, thereby resulting in a structured and complete methodology. With this, you will get both descriptive and predictive insights on why data breaches exist.

### Research Methodology

Use a chart to show how your work is completed and discuss each part of the chart. An example of a chart is shown in Figure2.



Critically evaluate each part of the chart to justify chosen research methods and analysis.



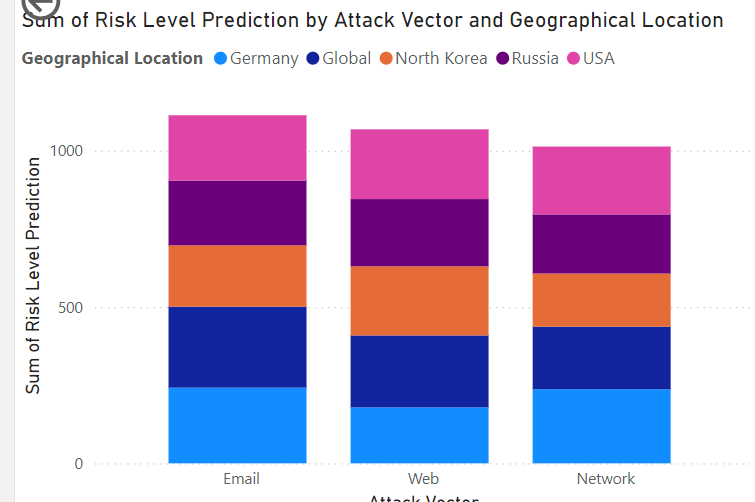
**Stage 1: Data collection**

Go back to section 3.

**Stage 2: exploration**

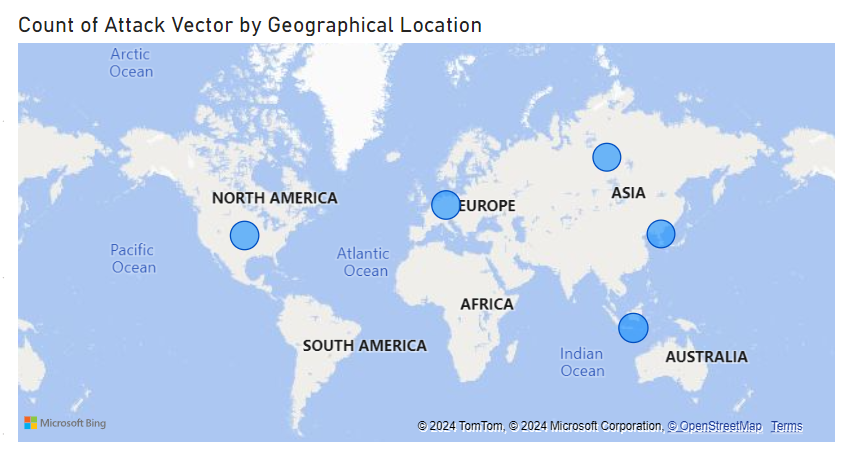
Using power Bi I choose the following charts:

1\_stacked column chart



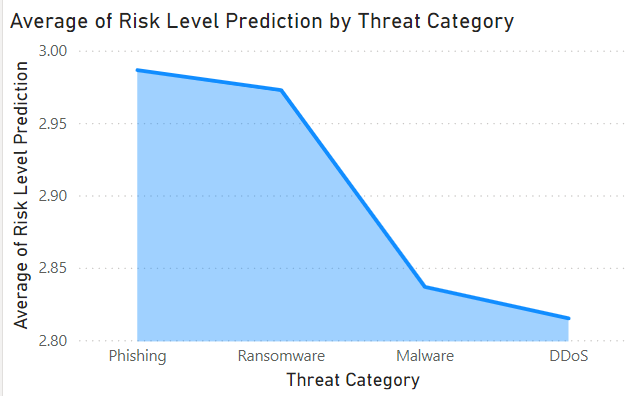
(To know the most location has the most risk level prediction and in which attack vector).

2\_Map



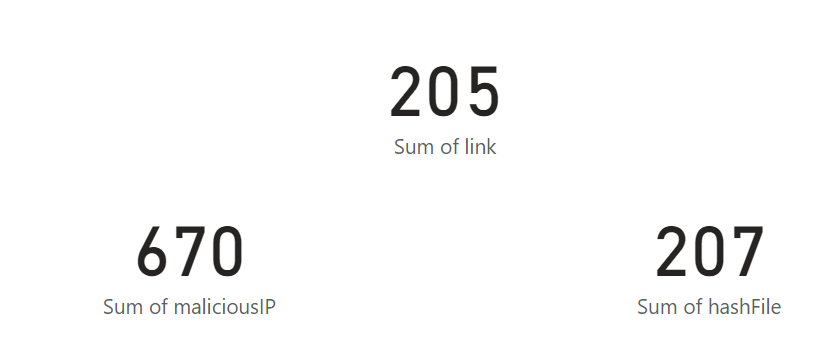
(To know the sum attack vector by the location and the most attack vector).

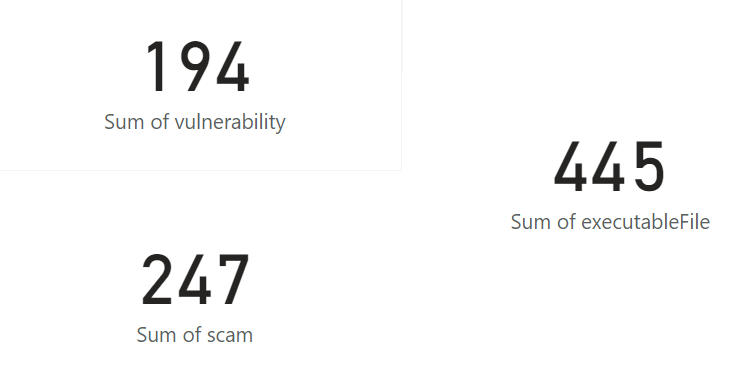
3\_Stacked area chart

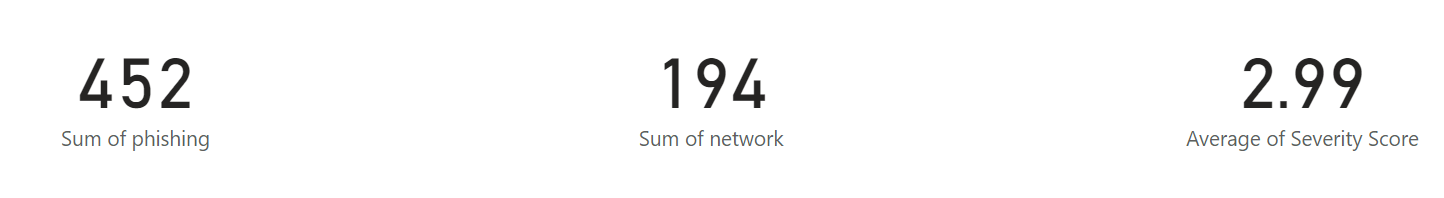


(To know the most average risk level prediction by the threat category).

4\_Card

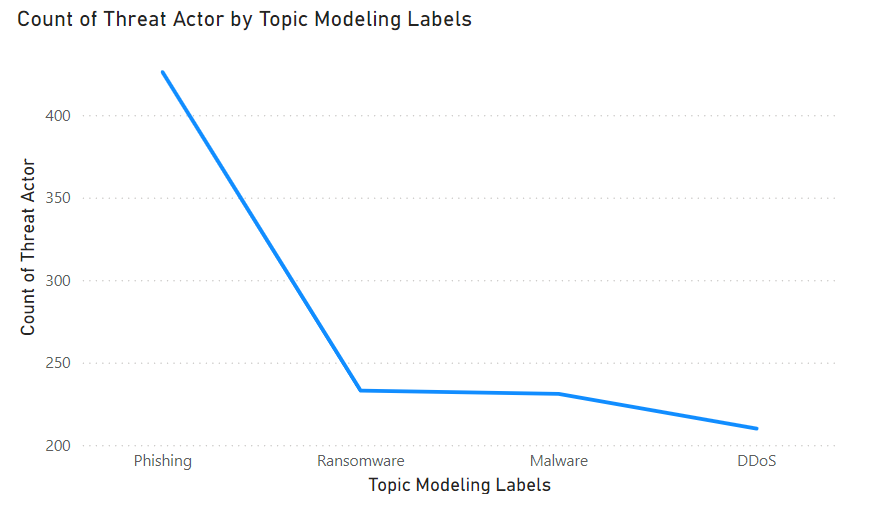






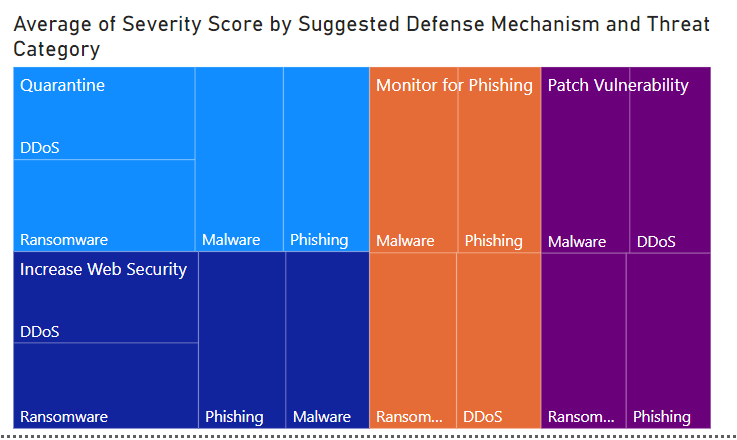
(cards to know the sum of the scams, executable files, vulnerability, malicious IP, phishing, network, Average severity score, hash Files and link).

5\_ line chart



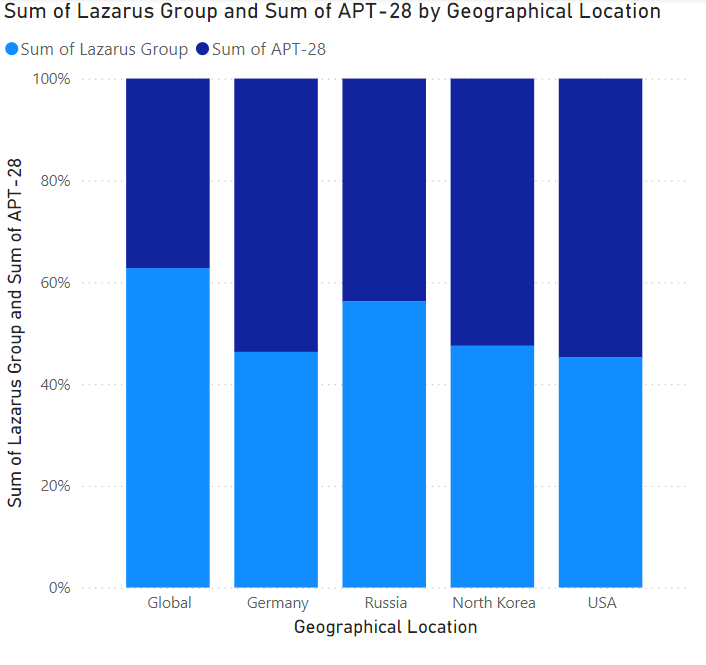
(to know what the topic model is targeted by the count threat actor).

6\_treemap



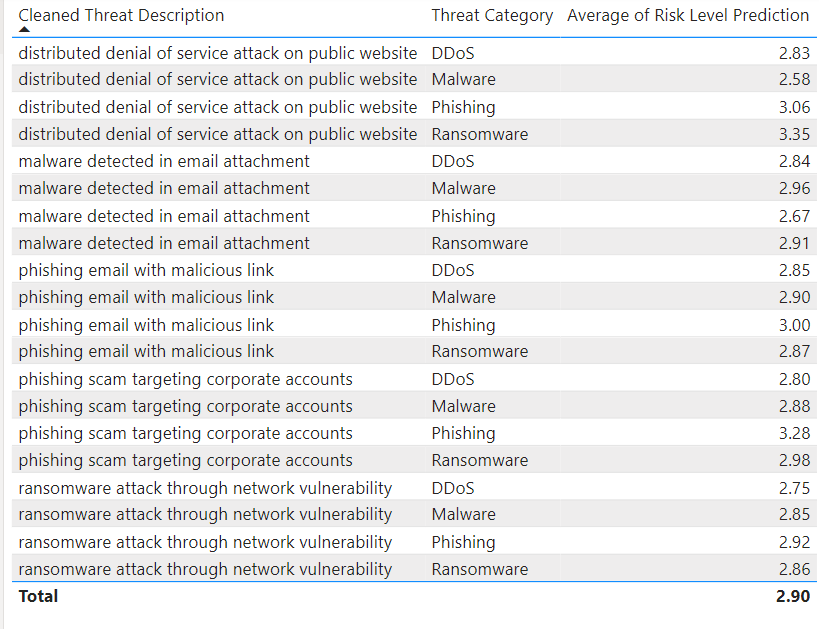
(To know the effected of the defense mechanism and the AVG of the severity score that related to the attack risk level divided base on the threat category)

7\_Stacked bar chart



(to know sum of attacks that both these attacker groups did in different location).

8\_Table



(the Table for the identify the most average risk level prediction with the description of the type of attack).

**Stage 3: modeling**

Using orange tool and here's some explanation the indicators of ML:

1\_ AUC (Area Under the Curve):

What it is: Receiver Operating Characteristic (ROC) curve is the area under the curve. ROC curve shows the relationship (true positive rate) vs (false positive rate) at different threshold levels.

Purpose: A model is assessed for discriminating those classes.

Range:

AUC = 1: Perfect classifier.

AUC = 0.5: Random guess.

AUC < 0.5: Worse than random guessing.

Why it matters: AUC values that are higher indicate that the model is more adept at giving positive cases a higher ranking than negative ones.

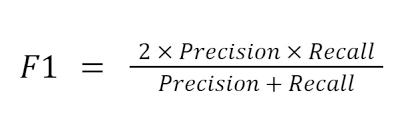
2\_ F1-Score:

What it is: Precision \* Recall / 2 It balances both metrics.

Purpose: This is useful when your dataset has an imbalance and you’re as concerned about the precision as you are about the recall.

Why it matters: It punishes very unbalanced problems (e.g high Precision and low Recall or vice versa).

Formula:



3\_ CA (Classification Accuracy):

What it is: It's the ratio of the correct labelled samples to total number of samples.

Where:

TP = True Positives.

TN = True Negatives.

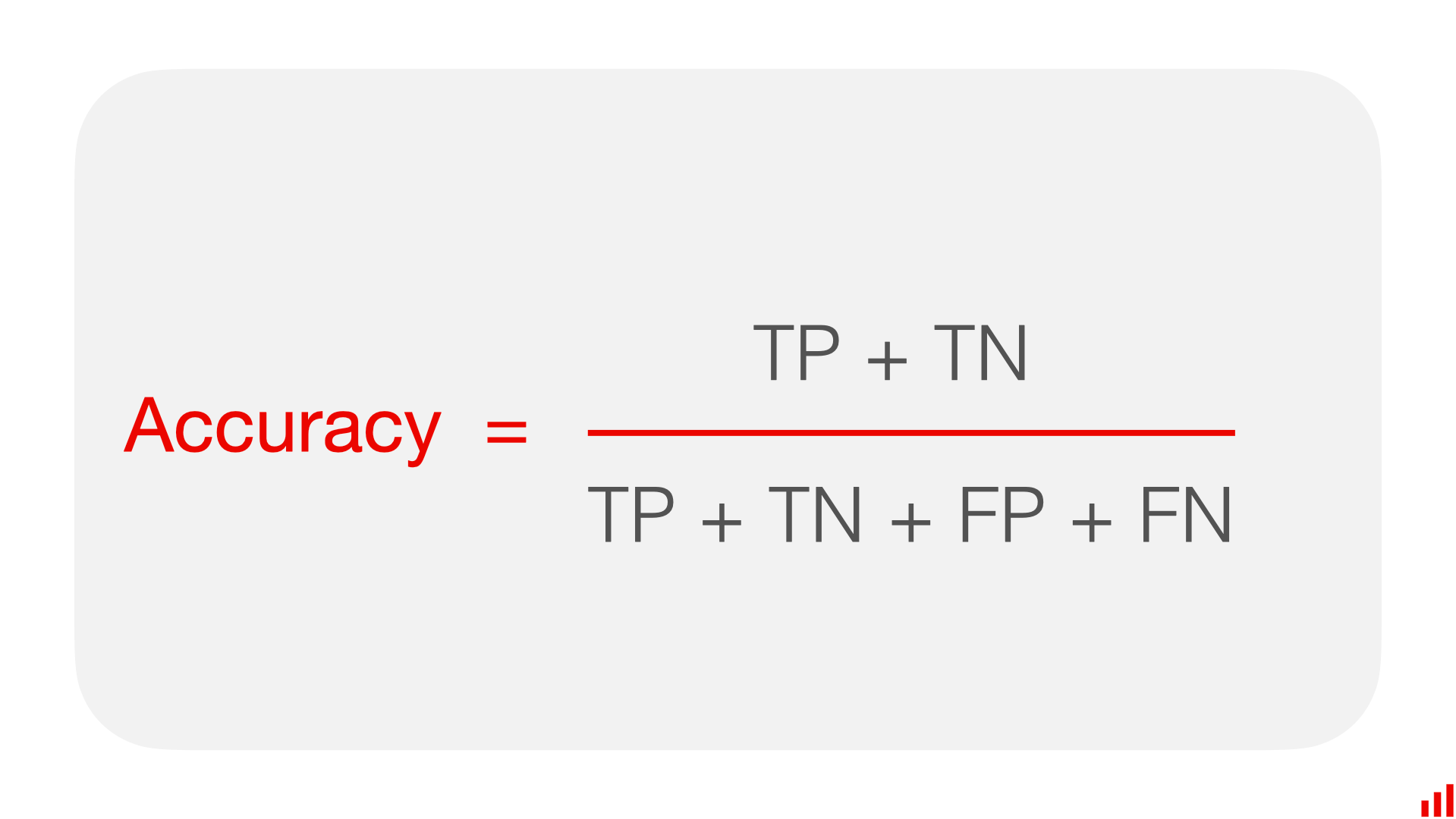
FP = False Positives.

FN = False Negatives.

Purpose: It measures overall correctness of the model.

Why it matters: But relatively simple and intuitive, and can potentially be misleading if the dataset is imbalanced (i.e., high accuracy but poor performance on minorities class).

Formula:



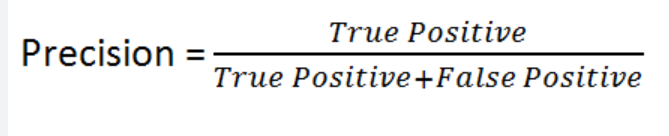
4\_precision:

What it is: Measures how many predicted positive cases are actually positive.

Purpose: Focuses on the relevance of positive predictions.

Why it matters: Important in situations where false positives are costly (e.g., spam detection, fraud prevention).

Formula:



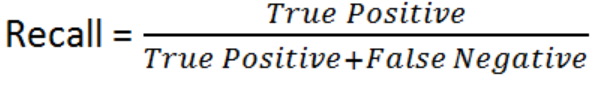
5\_Recall:

What it is: It tells us how many actual positive cases have been identified and we are correct in doing so.

Purpose: It mainly records all positive cases.

Why it matters: Also needed if false negatives have severe consequences (e.g. malicious threats in medical tests or security applications, etc.).

Formula:



6\_MCC (Matthews Correlation Coefficient):

What it is: It is a balanced metric taking all confusion matrix values: TP, TN, FP, FN.

Purpose: It provides a single score for binary classification even for imbalanced dataset.

Why it matters:

MCC = +1: Perfect prediction.

MCC = 0: Random prediction.

MCC = -1: Complete disharmony between actual and prediction.

# Results and Discussion

Discuss the results. Include all tables and figures.

Explain how the results meet the research question and objectives.

Describe merits and limits of the analysis.

# Conclusion and Recommendations

State the aim of the study and summarize the findings.

Based on the findings, discuss the recommendations that you have.

Propose future work (optional).

# Reflections

Avoid generalization and focus on personal development and the research journey in a critical and objective way.

* 1. **Selected Research Methodology**

Reflection of the research process.

Reflection on the merits, limitations, and potential pitfalls of the selected methods.

* 1. **Alternative Research Methodologies**

Alternative research methodologies in view of outcomes.

Lessons learned in view of outcomes.

* 1. **Recommended Actions and Future Considerations**

Use reflection to inform future considerations.

* 1. **Recommended Methodology**

Updated version of paper Methodology (flowchart, block diagram) with discussion.

**References**

[1] KasperSky (2022). How Data Breaches Happen. [online] Kaspersky. Available at: <https://www.kaspersky.com/resource-center/definitions/data-breach>.

[2] Trinckes, J. (2024). AI data breach: Understanding their impact and protecting your data - Thoropass. [online] Thoropass. Available at: <https://thoropass.com/blog/compliance/ai-data-breach/>.

[3] Long Cheng, F. L. D. (. Y., 2017. *Enterprise data breach: causes, challenges, prevention, and future directions.* [Online]   
Available at: <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1211>

[4]Adil Hussain Seh 1ORCID, M. Z. 2. A. 2. K. S. 1. A. 1. K. 1. a. A. K. 1., 2020. *Healthcare Data Breaches: Insights and Implications.* [Online]   
Available at: <https://www.mdpi.com/2227-9032/8/2/133>

[5] Gaurav Sood, K. C., 2019. *Pwned: The Risk of Exposure From Data Breaches.* [Online]   
Available at: https://dl.acm.org/doi/10.1145/3292522.3326046

[6] Timothy McIntosh, T. S. T. L. D. X. P. W. D. L. Y. H. A. N. M. H. I. &. C., 2024. *Ransomware Reloaded: Re-examining Its Trend, Research and Mitigation in the Era of Data Exfiltration.* [Online]   
Available at: <https://dl.acm.org/doi/10.1145/3691340>

[7] Amani Ibrahim, D. T.-G. S. A., 2020. *The Challenges of Leveraging Threat Intelligence to Stop Data Breaches.* [Online]   
Available at: <https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2020.00036/full>

[8] Hussain Afzaal 03 (2024). NLP Based Cyber Security Dataset. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/hussainsheikh03/nlp-based-cyber-security-dataset.

‌