A logo of a globe with a graduation cap

Description automatically generated**1. Introduction**

**30/9/2024**

|  |  |
| --- | --- |
| cybersecurity attacks project |  |

* **Project Title**: Cybersecurity Attacks Analysis & Visualization
* **Team Members**:
  + Youssef Ahmed – Data Cleaning & Preprocessing (Python)
  + Ahmed Mohamed Gamal – Preprocessing (Python)-Validation & Extra Analysis (SQL)
  + Rabie Taha – Data Visualization (Power BI)
  + Youssef Abdel\_Elhalim – Data Visualization (Power BI)
* **Supervisor**: Dr. Doaa Mokhtar
* **Objective**: To analyze a large cybersecurity dataset to uncover insights on network vulnerabilities, attack trends, and the effectiveness of security measures. Our analysis aimed to provide actionable data for improving cybersecurity protocols in organizations.
* **Significance**: In an era of increasing digital dependence, cyber-attacks are a growing threat to organizational and national security. This project focuses on identifying key trends in attack behaviors to better understand vulnerabilities and defense strategies.

**2. Challenges and Solutions**

* **Challenges Faced**:
  + **Limited Cybersecurity Knowledge**: The team initially had limited exposure to the domain of cybersecurity, which was critical to fully understanding the dataset and performing in-depth analysis.
  + **Missing Values in the Dataset**: The dataset included missing values in key columns, potentially leading to incomplete or inaccurate results if left unaddressed.
* **Solutions Implemented**:
  + **Learning & Research**: Each team member took the initiative to self-study cybersecurity concepts and familiarize themselves with the nature of attacks, types of protocols, and common vulnerabilities. This helped improve our understanding and guide the analysis.
  + **Data Imputation**: Missing values were handled using various techniques, such as replacing missing categorical values based on data trends or context. In some cases, records with extensive missing values were excluded.

**3. Workflow**

This section outlines the step-by-step process we followed to complete the project, from data collection to final insights.

1. **Data Collection**:
   * The dataset was sourced from **Kaggle** and consisted of cybersecurity logs, attack signatures, anomaly scores, and other relevant attributes.
   * The dataset contained various types of data, including numerical, categorical, and time-based information.
2. **Data Exploration**:
   * We used Python (Pandas, NumPy) for an initial exploration to understand the dataset’s structure and content. Key exploration tasks included:
     + Checking the shape of the dataset.
     + Summarizing key statistics.
     + Identifying missing and duplicated data.
3. **Data Cleaning & Preprocessing**:
   * We dealt with missing values using imputation and removed duplicates to ensure a clean dataset for analysis.
   * Additional features were engineered based on cybersecurity knowledge, such as creating new columns to capture trends in devices being attacked
   * A final check was performed to ensure data integrity before moving to SQL validation and Power BI.
4. **Formulating Key Questions**:
   * Based on the data and cybersecurity knowledge, we formulated several questions to guide our analysis, including:
     + Which countries have the highest cyber-attacks?
     + What is the daily trend of malware detection?
     + Which devices and operating systems are frequently targeted?
5. **SQL Validation and Extra Analysis**:
   * The cleaned dataset was imported into SQL for further analysis and validation.
   * SQL queries were executed to validate anomaly scores, calculate packet lengths, and analyze attack trends by year/month and traffic type.
6. **Data Visualization**:
   * Power BI was used to visualize the data, presenting attack patterns, malware trends, and target distribution.
   * Several dashboards were created to display critical insights interactively.
7. **Insights & Interpretation**:
   * The visualizations and SQL analysis helped derive actionable insights, including the identification of vulnerable systems, attack frequency patterns, and correlations between malware detection and anomaly scores.
8. **Presentation and Reporting**:
   * The results were compiled into a PowerPoint presentation and summarized in this report.
   * The presentation highlighted key findings and actionable recommendations based on the analysis.

**4. Data Cleaning and Preprocessing**

* **Technology Used**: Python (Pandas, NumPy)
* **Steps Followed**:
  1. **Dataset Loading**: The raw dataset was loaded into Python for exploration.
  2. **Initial Exploration**: We performed basic data exploration (e.g., head(), info(), and describe()) to understand the data structure and content, including its types, missing values, and ranges.
  3. **Handling Missing Values**: The dataset had missing values in several key fields, which were addressed using:
     + Mode imputation for categorical variables (e.g., malware indicator).
     + Dropping columns after extracting the information from them
  4. **Duplicate Data Removal**: We checked for and removed any duplicated records to ensure the data's uniqueness and reliability.
  5. **Feature Extraction**: Based on the research into attack patterns, we created new features such as "Browser" and "Device\OS" to derive more meaningful insights.
  6. **Final Data Check**: A thorough review was performed to ensure all issues were resolved before starting the analysis phase.
  7. **Outcome**: The cleaned and preprocessed dataset was ready for SQL validation and visualization in Power BI. We ensured that the data was consistent, accurate, and well-structured for analysis.

**5. Questions and Their Answers**

**Key Questions Investigated**:

1. **Which countries experience the highest number of cyber-attacks, and what insights can be drawn from their geographic distribution?**
   * **Findings**: Countries with advanced economies and large technology infrastructures (e.g., the U.S., China, and Russia) experienced the highest frequency of attacks. Cities with more corporate or government infrastructures were particularly vulnerable.
2. **How have malware indicators evolved over time, and what trends can be observed in their frequency?**
   * **Findings**: There was a noticeable increase in malware indicators during specific periods, often correlating with global events like political tensions or economic sanctions, which could trigger politically motivated cyber-attacks.
3. **Which devices and operating systems are most frequently targeted by attacks?**
   * **Findings**: Windows-based systems were disproportionately targeted, potentially due to their prevalence in the corporate environment and vulnerabilities in older systems that had not been patched.
4. **What is the daily trend of anomaly scores, and how does it correlate with malware detection?**
   * **Findings**: A strong correlation was observed between high anomaly scores and increased malware detection on certain days. This indicated that spikes in anomaly scores could be used as early warning signs for potential malware activity.
5. **Which protocols and destination ports are most frequently targeted in attacks?**
   * **Findings**: TCP and UDP were the most commonly used protocols in attacks, with certain ports (e.g., port 80 for HTTP and port 443 for HTTPS) being frequently targeted, suggesting attackers' preference for web-based vulnerabilities.
6. **What patterns can be seen in protocol usage based on specific firewall logs or attack signatures?**
   * **Findings**: Specific attack signatures were associated with known vulnerabilities in firewall logs, helping pinpoint which firewall rules were being bypassed or needed enhancement.

**6. SQL Validation and Extra Analysis**

* **Database Setup**: We created a relational database in SQL named "Cybersecurity" and loaded the cleaned dataset into it for validation.
* **Validation & Analysis Performed**:
  1. **Average Anomaly Scores**: Using SQL queries, we calculated the average anomaly score for each day, validating the results against the Python-calculated scores.
  2. **Malware Indicator Trends**: We analyzed the count of malware indicators per day and cross-referenced this with the anomaly scores to validate the correlation between the two.
  3. **Average Packet Length**: By calculating the average packet length per attack type, we identified which types of attacks were associated with larger or more data-intensive packets.
  4. **Count of Attacks by Year/Month**: This analysis revealed temporal trends, highlighting specific months and years that experienced spikes in attack numbers, which could indicate coordinated attack campaigns.

**7. Driving Insights**

**Major Insights Derived from Analysis:**

1. **Attack Frequency by Country: The data showed that certain countries experienced significantly more cyber-attacks than others, often due to higher levels of industrialization, technological infrastructure, or political instability.**
2. **Severity of Attacks by User Types: Users from large corporations or government entities were targeted more frequently by sophisticated attacks, often involving high-severity malware and targeted exploit attempts.**
3. **Response Action Efficiency: By analyzing the actions taken in response to attacks (e.g., blocked, logged, ignored), we assessed how prepared the systems were to handle different types of threats. A high number of "ignored" attacks suggested a need for more proactive defense mechanisms.**
4. **Protocol Vulnerabilities: TCP-based attacks were among the most common, highlighting the need for strengthening network security protocols and ensuring that firewalls and intrusion detection systems are optimized for TCP traffic.**

**8. Conclusion**

* **Key Takeaways: This project demonstrated the value of analyzing cybersecurity data to identify key trends in attack patterns and system vulnerabilities. Through a combination of Python-based preprocessing, SQL validation, and Power BI visualization, we uncovered critical insights that could help improve the resilience of organizations against cyber threats.**
* **Future Work: Potential future analyses could include deeper exploration of attack vectors and further segmentation of user types to understand specific risks posed to different sectors (e.g., healthcare, finance, education).**

**Implementation of the work :**

1. **Python :**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generatedA white background with a black border

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A white rectangular object with black border

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A blue rectangle with white text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A pie chart with different colors

Description automatically generated

A screenshot of a computer

Description automatically generated

**For accessing the Python Notebook:** [**Click Here**](https://colab.research.google.com/drive/1fU3Qgb-VWQVbb65bygut-KwUpUN-sl6w?usp=sharing)

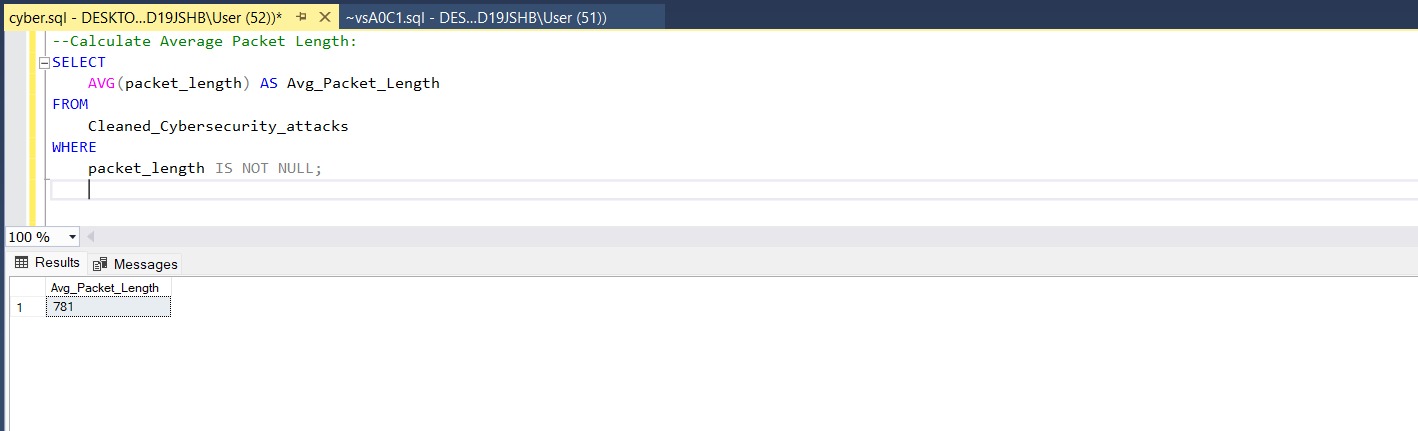
**2-SQL:**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

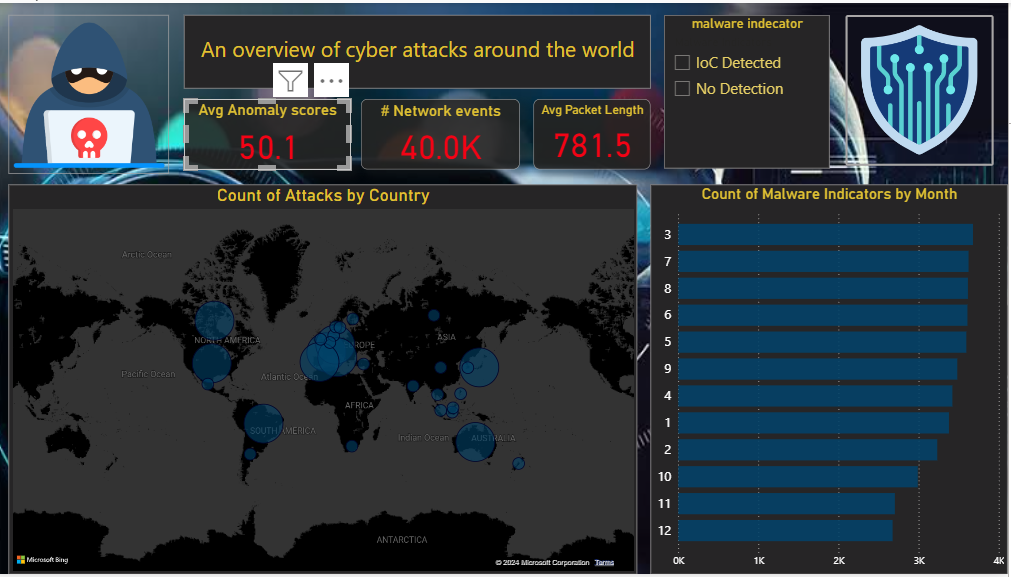
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**3-Power Bi:**

****

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer screen

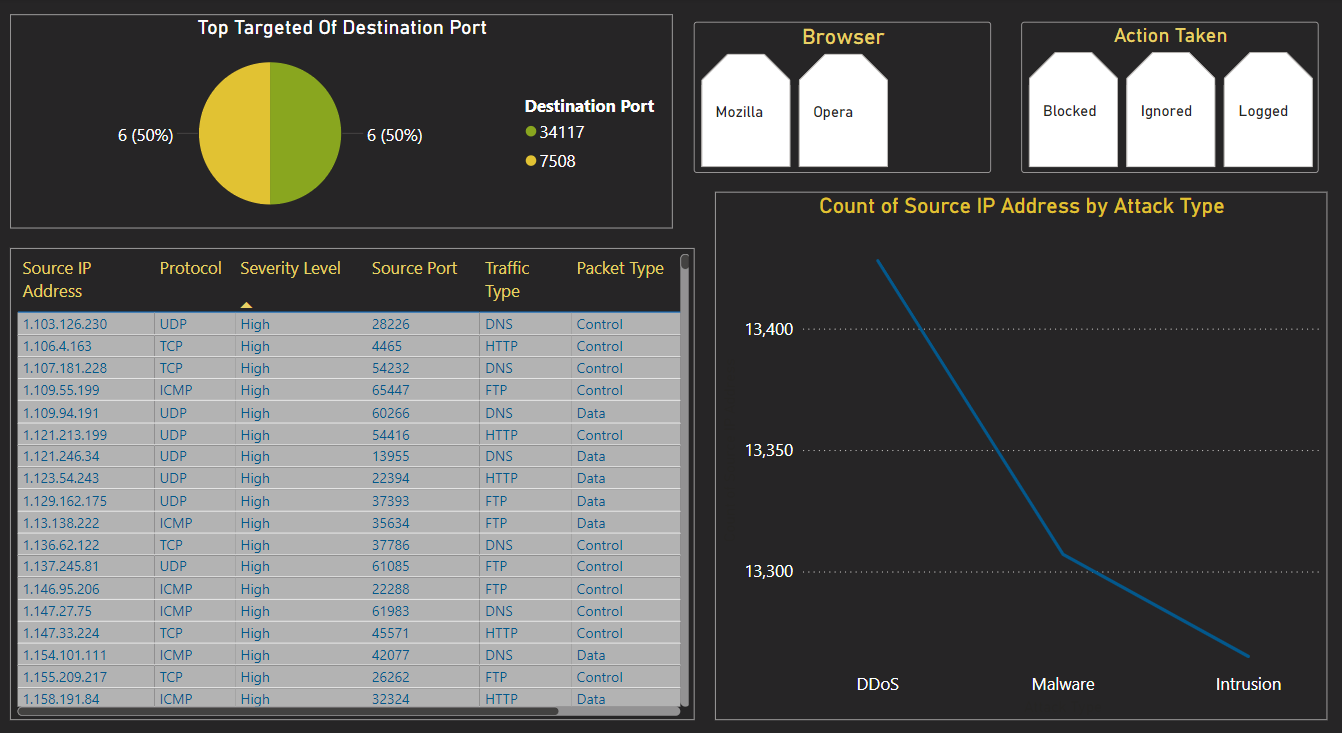
Description automatically generated**

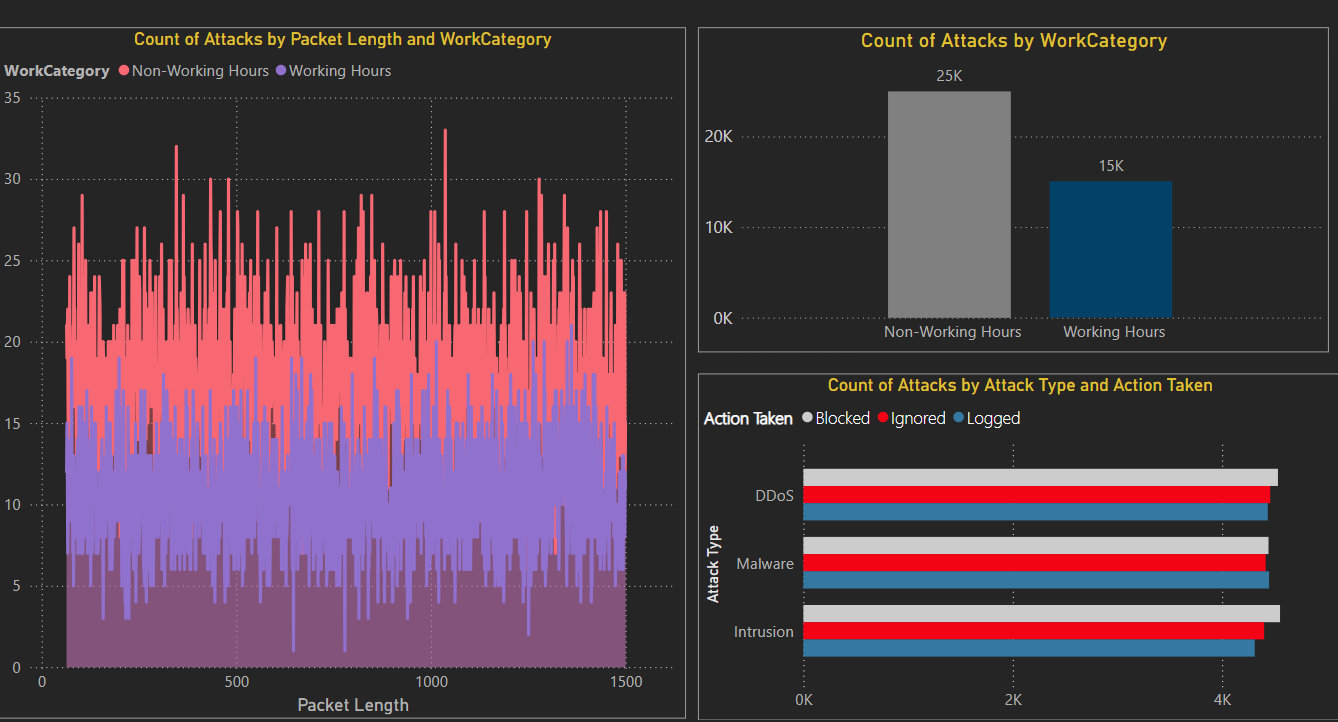
**A screenshot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

****

****

****

**THANK YOU**

You can access all the data from our drive : [Click Here](https://drive.google.com/drive/folders/1EAxVvqyImNfxG7zIFRm3YsM-979GwDXf?usp=sharing)