**Cybersecurity Attacks**

Vinay Reddy Jakka

Data Visualization Project

**Introduction:**

A cybersecurity attack is any ill attempt made to steal or reveal secrets of individuals or businesses and compromising CIA (Confidentiality, Integrity, and Availability) of data or information systems. These attacks check for vulnerabilities in the system or network to gain access to sensitive information and causing harm in various forms.

Cybersecurity attacks can occur in any form and any individual, organization or government institute can be a target. Common types of attacks we come across are viruses, ransomware, phishing attacks, and data breaches. The reasons behind an attack can be for financial gain, copying intellectual properties, or competitive advantage among companies.

We have seen a quick escalation of attacks over the years as more individuals and organizations are learning about digital platforms and accessing them. Consequences that occur after cybersecurity are not limited to financial loss, it may also lose reputation in the market, loss of customers trust, and fines levied by government. So, everybody using digital platforms should take some measures countering cybersecurity threats and attacks. By analyzing the dataset, we can find trends and patterns on attacks to boost the security and safety of data and systems. This analysis enlightens us with a broader perspective on attacks and address these effectively.

**Dataset:**

Let’s discuss the project in terms of data life cycle.

**Data Generation:**

Aashray Agur and Uma Venugopal are the authors of dataset. It is mentioned in the source (Kaggle) that data is synthetically generated using cGan.

**Data Collection:**

I downloaded the dataset from Kaggle. The dataset provides information on cybersecurity attacks from 2020-2023. The dataset consists of 40000 records and 25 distinct attributes. The dataset covers basic details on cybersecurity attack scenarios such as time stamps, attack types, severity levels, etc.

Dataset Source: <https://www.kaggle.com/datasets/teamincribo/cyber-security-attacks>

**Data Processing:** Let’s delve deep into dataset and clearly understand all the attributes and their significance.

* **Time stamp:** Date and time of cybersecurity attack.
* **Source IP Address:** IP address of source of the attack.
* **Destination IP Address:** IP address of the target.
* **Source Port:** Port number used by source during the attack.
* **Destination Port:** Port number used by source during the attack.
* **Protocol:** Communication protocol used in attack.
* **Packet Length:** Size of data packets uploaded/downloaded during attack.
* **Packet Type:** Type of packets transmitted.
* **Traffic Type:** Type of network traffic seen during attack.
* **Payload Data:** Data payload transmitted during attack.
* **Malware Indicators:** Indicating if there was any detected malware.
* **Anomaly Scores:** Scores showing a level of deviant behavior.
* **Alerts/Warnings:** Whether alerts or warnings have been given during an attack.
* **Attack Type:** Type of attack used by hackers.
* **Attack Signature:** Associated signatures with previous attackers.
* **Action Taken:** Response taken in order to deal with attack.
* **Severity Level:** Level of severity associated with attack.
* **User Information:** Information about user involved in the attack.
* **Device Information:** Information about the device used by targeted users.
* **Network Segment:** Segment of network affected by attack.
* **Geo-location Data:** Geographic location of user who is attacked.
* **Proxy Information:** Details about IP address used as proxy.
* **Firewall Logs:** Logs generated by firewall systems.
* **IDS/IPS Alerts:** Alerts raised by intrusion detection/prevention systems.
* **Log Sources:** Origin of log data.

After reviewing all the attributes clearly, I concluded that following attributes are more beneficial to cybersecurity engineer rather than data visualization are Source IP, Destination IP, Payload Data, Proxy Information as these contain unique data to every user and doesn’t work well with visualization. I will be focusing on the mentioned attributes to gain insights and patterns, Timestamp, Protocol, Packet Type, Traffic Type, Malware Indicators, Anomaly Scores, Alerts/Warnings, Attack Type, Severity Level, Location.

**Data Cleaning:**

* Trimming only state part of location from Geo-location Data to plot easily.
* Found missing values in Alerts/Warnings attribute, filled with “Alert Not Triggered”.
* Found missing values in Malware Indicators attribute, filled with “IoC not Detected”.
* Found missing values in IDS/IPS Alerts attribute, filled with “Didn’t Alert Data”.

**Tools Used:**

* Python
* Tableau

**Exploratory Data Analysis:**

* Used Python for exploratory data analysis, imported necessary libraries like Pandas, NumPy, Matplotlib, and Seaborn.
* Imported the dataset and performed various analysis metrics such as shape, info, describe, is null, etc., to get a good read on the dataset.
* The dataset I chose has 25 attributes and 40,000 rows. After cleaning and handling missing data the shape of dataset has changed to 18 attributes and 40,000 rows.

Below are the code snippets used for data processing and data cleaning:

A close-up of a message

Description automatically generated

Fig 1. Imported important libraries

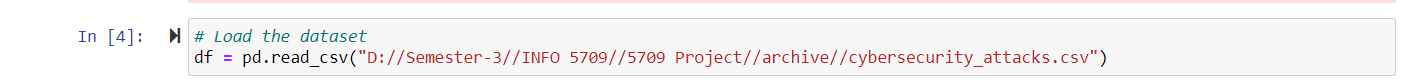


Fig 2. Imported the dataset into jupyter notebook

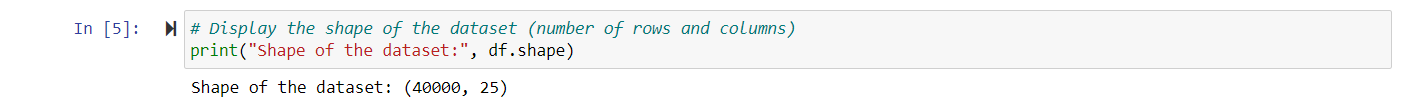


Fig 3. Shape of the dataset before cleaning

A screenshot of a computer

Description automatically generated

Fig 4. Info to check data types of all attributes

A screenshot of a computer

Description automatically generated

Fig 5. Describe feature to see statistics of the data

A screenshot of a computer

Description automatically generated

Fig 6. Is null to check number of missing values in each of the attributes in the dataset

A close-up of a computer screen

Description automatically generated

Fig 7. Removing irrelevant columns and filling the missing values

A screenshot of a computer

Description automatically generated

Fig 8. Final data frame with shape at bottom

A white background with black text

Description automatically generated

Fig 9. Final dataset missing values count with is null function

1. How many attacks happened in total for each type of attack and which type has the highest attacks?

A white background with black text

Description automatically generated

Fig 10. Setting plot canvas

A graph of blue rectangular shapes

Description automatically generated with medium confidence

Fig 11. Count of each Attack Type

Result: Above figure shows that DDoS attacks are highest among other 3 attacks with 13,428 attacks in total.

1. Show a graph displaying distribution of various severity levels in the dataset, which severity levels have lowest attacks with number of attacks.

A white card with black text

Description automatically generated

Fig 12. Count plot of severity levels

A graph showing different blue rectangular shapes

Description automatically generated with medium confidence

Fig 13. Distribution of Severity Levels

Result: The above plot shows the count of attacks for each severity level, we can see ‘Low’ severity attacks were less than other two types with a value of 13,183.

1. How many attacks triggered an alert?

A close-up of a computer screen

Description automatically generated

Fig 14. Count plot of Alert/Warnings

A graph showing a number of blue squares

Description automatically generated

Fig 15. Count of Alert/Warnings

Result: The output shows that 19,933 attacks have given alerts for the users to act.

1. Explain the distribution of packet length variations with attack type.

A close-up of a computer code

Description automatically generated

Fig 16. Box plot of Attack Type code snippet

A graph with blue squares

Description automatically generated

Fig 17. Box plot of Attack Type

Result: We can see from the box plot that; the distribution of packet length isn’t varying much. But most of the attacks happen between the packet lengths 400 and 1200.

1. For how many attacks did the system detect IoC?

A close-up of a white background

Description automatically generated

Fig 18. Count plot of Malware Indicators code snippet

A graph with blue squares

Description automatically generated

Fig 19. Count plot of Malware Indicators

Result: IoC is an indicator of compromise, that means it shows there has been breach in the network endpoint. There are a total of 20,000 attacks where IoC is detected.

1. Show the anomaly scores distribution and their variation with severity levels.

A white card with colorful text

Description automatically generated

Fig 20. Box plot of Anomaly Scores Code

A diagram of a diagram

Description automatically generated with medium confidence

Fig 21. Box plot of Anomaly Scores

Result: We can see from the box plot that; the distribution of anomaly scores isn’t varying much. But most of the attacks happen between the anomaly scores 25 and 75.

1. Show year wise attack distribution and which year had more attacks in total?

A blue and white bar

Description automatically generated with medium confidence

Fig 22. Year wise Total Attacks

Result: From the horizontal bar graph, we can see the distribution of total attacks for each year. We can see the highest number of attacks happened in the year 2022 with a value of 10,750.

I downloaded the cleaned dataset with the below command, to be able to upload the dataset to Tableau and visualize the results of hypothesis.

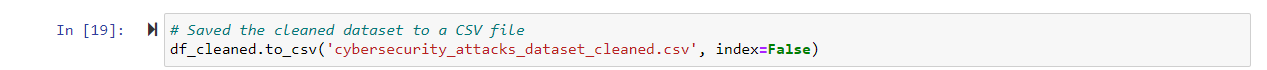


Fig 23. Code snippet to download dataset

**Hypothesis:**

1. Does the cyber-attacks distribution vary based on severity level and protocol type? Which protocol type is easy to attack?
2. How are the different types of attacks varying over time? Is there any peculiar behavior in line graph observed?
3. Did malware indicators help detect the attacks happening? What % of attacks had IoC detected? From the above spike in 2022 Q2, how many attacks in intrusion has IoC not detected in 2022?
4. How are action taken attributes distributed? How did it vary through the years? Does the action taken depend on severity level? Show distribution of actions taken with respective to attack types with a % total.
5. Show the geographical distribution of attacks? Which state was most vulnerable and had the highest number of attacks in total?

**Data Visualization:**

1. Does the cyber-attacks distribution vary based on severity level and protocol type? Which protocol type is easy to attack?



Fig 24. Total Attacks under various Protocols

* I used a stacked bar chart here to see the distribution of attacks based on protocols used for different severity levels. I felt this chart is suitable because we are dealing with 3 kinds of protocols and their total can be grouped by color variations. Each bar represents total attacks for a severity level, then with use of different colors we can breakdown attack count by protocol used (ICMP or TCP or UDP).
* I provided a year filter to see how attacks are varying across the years, severity level filter to see total number of attacks for a specified severity level, and protocol feature to see total number of attacks based on protocol.
* From the above visualization, I understood that ICMP protocol has been attacked more frequently than other protocols. Even when compared with total attacks for each protocol, ICMP took first place indicating ICMP protocol is vulnerable and more susceptible to cyber-attacks. This helps cybersecurity engineers understand where the weakest link is and what type of network protocol they are using when an attack occurred and makes them strengthen the network protocols to stay safe.

1. How are the different types of attacks varying over time? Is there any peculiar behavior in line graph observed?

A graph of a graph

Description automatically generated with medium confidence

Fig 25. Attack Type variation over time

* To answer the above hypothesis, I chose a lines (discrete) graph in Tableau. This chart is appropriate for a time series graph, which shows variation over a certain period of time. Each line represents variation of particular attack type from 2020 through 2023. Also used color variation for each attack type to distinguish among themselves.
* The only filter I used here is attack type, through which I can see a line graph for any type of attack at once.
* From the above visualization, I found that there is a sharp decline in the number of attacks from 2023 Q2 to 2023 Q4. After thoroughly seeing the dataset, I concluded that dataset generation has only been considered till September month and not any further. That’s the reason we see a sudden drop in the number of attacks.
* The highest spike is seen during 2022 Q2, it shows intrusion attack type with a value of 938 attacks.

1. Did malware indicators help detect the attacks happening? What % of attacks had IoC detected? From the above spike in 2022 Q2, how many attacks in intrusion has IoC not detected in 2022?

A screenshot of a computer screen

Description automatically generated

Fig 26. Malware indicators bubble chart

* To investigate this kind of problem, I used a bubble chart to explore the distribution of attacks with malware indicators. In malware indicator attribute, “IoC Detected” and “IoC not Detected” are used to alert the users about breaches that may have happened without knowing. This shows evidence of compromise in the organization’s network.
* The bubble chart allows us to visually interpret how each of the attack type is distributed with malware indicator. The size of bubble chart is explained with count of attacks.
* I used a year filter to see how attacks are varying across the years, an attack type filter to see total number of attacks for a specified attack, and malware indicator feature to see total if the attack was detected or not.
* From the visualization, there are a total of 20,000 attacks where IoC is detected, showing that those systems are vulnerable. That implies a 50% success in indicating compromises of total.

A diagram of a type of attack

Description automatically generated with medium confidence

Fig 27. Malware indicators bubble chart for Intrusion in 2022

* In the above visualization, I applied filters to get intrusion attack type distribution in 2022.
* We can see that IoC is detected for 1,824 attacks in total.
* To answer the hypothesis question, it is undetected for 1,739 attacks.

1. How are action taken attributes distributed? How did it vary through the years? Does the action taken depend on severity level? Show distribution of actions taken with respective to attack types with a % total.

A pie chart with a number of numbers

Description automatically generated with medium confidence

Fig 28. Pie chart of Action Taken distribution

* Used pie chart to clearly understand the distribution. The action taken attribute has options like blocked, ignored, logged. Blocked – the team has identified the threat and defense mechanism has already blocked the attack from happening. Ignored – The team has detected the attack and chose not to take any action, this maybe because of low severity or other reasons. Logged – It means the incident has been recorded in system logs, to further analyze the details of attack.
* I implemented filters year and action taken to be able to see variation in pie chart over years.
* We can interpret that the users have blocked most of the attacks with a value of 13,529 attacks.

A graph with red and blue lines

Description automatically generated

Fig 29. Line chart of Action Taken distribution over time

* The above line graph tells us the exact behavior of actions taken over the years. All the actions peaked in 2022 and dropped a bit in 2023.

A pie chart with a few different colored circles

Description automatically generated with medium confidence

Fig 30. Action taken distribution for High severity attacks

* I used an additional filter here, severity level. This helps us understand what actions have been taken against several severity levels.
* It seems like there is no direct correlation between the action taken and severity level, as the numbers seem around same for all severity levels.
* The concerning fact from this visual is that there are 4,460 attacks ignored from severity level ‘high’. This shows systems have been lacking features to properly categorize attacks to take appropriate actions.

A screenshot of a graph

Description automatically generated

Fig 31. Pie chart distribution of action taken for various attack types

* Here I utilized attack type filter to learn how to react to upcoming attacks. The filters I used here are year, action taken, and attack type.
* In a collective purpose we can see that around 33% of the attacks are ignored. This is what causes a large range of ransomware attacks or breaches.
* The companies or users need to implement systems which better classify the attacks with accurate severity levels to implement appropriate actions to keep the data safe.

1. Show the geographical distribution of attacks? Which state was most vulnerable and had the highest number of attacks in total?

A map of india with different colored states

Description automatically generated

Fig 32. Geographical Representation of attacks

* The dataset collected showed geographical location as India, covering most of the states. The dataset had City and State as format. But with the help of python, I got rid of city as they may not be recognized by Tableau and included only state part. This helped me easily map them onto India map in Tableau.
* I used filters State – to see total number of attacks in particular state and year filter to read how the attacks trend has changed over years.
* The most vulnerable state is Manipur with a total number of attacks of 1,498 over all the years.

**Dashboards:**

Dashboard 1:

A screenshot of a data presentation

Description automatically generated

* Dashboard can accommodate more than 2 sheets and will be a great tool to visualize the data and explain the insights easily. The sheets I used in this dashboard are Geographical distribution, Malware indicators, Attack type variation, and action taken.
* The active filters are state – if I click on any state, we get all other sheets displaying data from that particular sheet, severity level – this can be applied on top of state or independently as well to see data show various severity levels, attack type – to see data about various attack types.
* With the help of this dashboard, we can explain how attacks are being handled and also know about malware indicators, and state wise statistics as well.

Dashboard 2:

A screenshot of a data presentation

Description automatically generated

* In this dashboard we can see how various attack types are taken care of in each of the state available and also understand pattern of attacks on various protocols.

**Conclusion:**

The analysis of cybersecurity attacks data has provided valuable insights on cyber threats. By exploring all the hypothesis and visualizations, I have found some patterns and the most vulnerable users, systems, states, network protocols, etc. From the geographical point of view, we could find some hotspots for cyber-attacks and these kinds of insights help users/organizations strengthen their defense applications and use resources properly. As a whole, this might also help form better strategies to tackle cyber threats and safeguard critical data/assets.

**References:**

*Cyber Security Attacks*. (n.d.). Www.kaggle.com. <https://www.kaggle.com/datasets/teamincribo/cyber-security-attacks>

*Differences Between TCP, UDP, and ICMP | PingPlotter*. (n.d.). Graphical Network Monitoring and Troubleshooting | PingPlotter. <https://www.pingplotter.com/wisdom/article/packet-type-differences/>

*The 12 Most Common Types of Cybersecurity Attacks Today*. (n.d.). Https://Blog.netwrix.com/. <https://blog.netwrix.com/types-of-cyber-attacks>

*‌What Are Indicators of Compromise (IOC)?* | Microsoft Security. (n.d.). Www.microsoft.com. <https://www.microsoft.com/en-us/security/business/security-101/what-are-indicators-of-compromise-ioc#:~:text=An%20indicator%20of%20compromise%20(IOC>