Exploratory Data Analysis and Visualization Global Cybersecurity Threats (2015-2024)

ISDA 111 – Information and Data Science
Spring 2025
By: George Triplitt

Synopsis

Cybersecurity threats are as prevalent as ever due to the constant expansion of technology and global connectivity. Every day, different facets of the business industry face the grim reality of data management and security. Companies rely on various security measures to remain steps ahead of the opposition. As a student pursuing a potential cybersecurity career, it is only fitting to choose an exploratory analysis along with a visual report on the field that I intend to enter. This report aims to find valuable insights into cybersecurity-related threats that have been reported worldwide in the last nine years. I aim to apply descriptive, diagnostic, and predictive analysis combined with the Python programming techniques acquired from Dr. Pramod Gupta's Information and Data Science class to analyze an omnipresent cybersecurity topic.

The dataset utilized for this project is called "*Global Cybersecurity Threats 2015-2024*" and was acquired from *Kaggle.com*. The dataset explores various metrics, consisting of ten columns and almost three thousand rows, ranging from targeted industries to time of incident resolution. This dataset should provide ample information to develop meaningful conclusions for improvements in business applications.

Dataset Preview

Quickly scanning through the dataset gave me two distinct benefits—it allowed me to process the overall data condition and begin conceptualizing how to approach the data analysis. I began by sorting the entire set of data and unlocking all rows and any missing columns. Applying "display.max_rows" in pandas gave me all the available data that was not provided by the initial input of "df.head()."

	Country	Year	Attack Type	Target Industry	Financial Loss (in Million $)	Number of Affected Users	Attack Source	Security Vulnerability Type	Defense Mechanism Used	Incident Resolution Time (in Hours)
0	China	2019	Phishing	Education	80.53	773169	Hacker Group	Unpatched Software	VPN	63
1	China	2019	Ransomware	Retail	62.19	295961	Hacker Group	Unpatched Software	Firewall	71
2	India	2017	Man-in-the-Middle	IT	38.65	605895	Hacker Group	Weak Passwords	VPN	20
3	UK	2024	Ransomware	Telecommunications	41.44	659320	Nation-state	Social Engineering	AI-based Detection	7
4	Germany	2018	Man-in-the-Middle	IT	74.41	810682	Insider	Social Engineering	VPN	68

2993	Germany	2017	SQL Injection	Education	54.98	786577	Insider	Unpatched Software	Firewall	70
2994	Germany	2019	Ransomware	Government	58.60	76066	Insider	Unpatched Software	AI-based Detection	8
2995	UK	2021	Ransomware	Government	51.42	190694	Unknown	Social Engineering	Firewall	52
2996	Brazil	2023	SQL Injection	Telecommunications	30.28	892843	Hacker Group	Zero-day	VPN	26
2997	Brazil	2017	SQL Injection	IT	32.97	734737	Nation-state	Weak Passwords	AI-based Detection	30
2998	UK	2022	SQL Injection	IT	32.17	379954	Insider	Unpatched Software	Firewall	9
2999	Germany	2021	SQL Injection	Retail	48.20	480984	Unknown	Zero-day	VPN	64

Data Cleaning

It was crucial to establish a clean dataset in order to move forward. My first task in data cleaning was establishing different data types. This was accomplished with the input "df.info()." With this command, I determined that I had three types of data: objects, floats, and integers.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Country                           3000 non-null   object
 1   Year                              3000 non-null   int64
 2   Attack Type                       3000 non-null   object
 3   Target Industry                   3000 non-null   object
 4   Financial Loss (in Million $)     3000 non-null   float64
 5   Number of Affected Users          3000 non-null   int64
 6   Attack Source                     3000 non-null   object
 7   Security Vulnerability Type       3000 non-null   object
 8   Defense Mechanism Used            3000 non-null   object
 9   Incident Resolution Time (in Hours)  3000 non-null   int64
dtypes: float64(1), int64(3), object(6)
memory usage: 234.5+ KB
```

Moving on from confirming data types to locating any duplicates, the input list "[df.duplicated() ==True]" showed all the possible duplicates within the rows. Additionally, I reconfirmed the results with "df.duplicated().sum()".

Checking for duplicates:

```
df[df.duplicated() == True]
```

Country	Year	Attack Type	Target Industry	Financial Loss (in Million $)	Number of Affected Users	Attack Source	Security Vulnerability Type	Defense Mechanism Used	Incident Resolution Time (in Hours)

```
df.duplicated().sum()
```

```
0
```

The next objective in the data cleaning process was to check for missing values in our dataset. I accomplished this task by creating a separate column to consolidate any missing values per category.

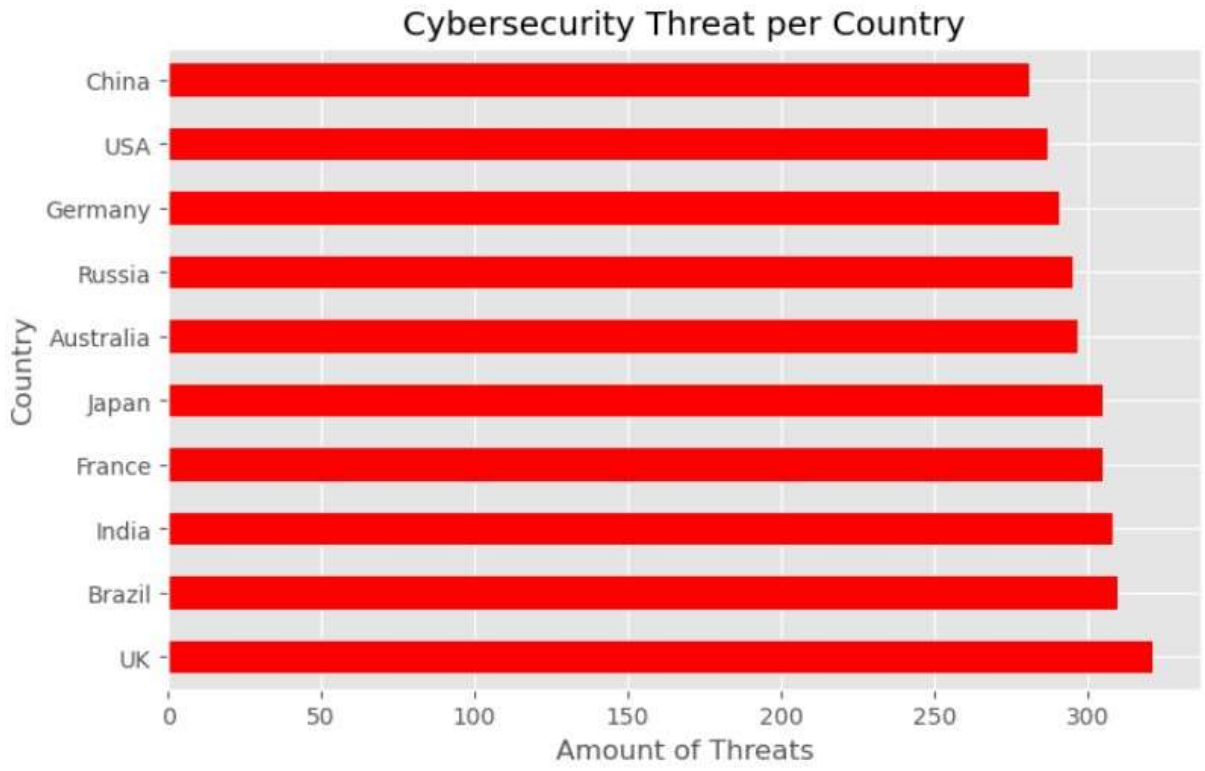
	Columns	Missing Values
0	Country	0
1	Year	0
2	Attack Type	0
3	Target Industry	0
4	Financial Loss (in Million $)	0
5	Number of Affected Users	0
6	Attack Source	0
7	Security Vulnerability Type	0
8	Defense Mechanism Used	0
9	Incident Resolution Time (in Hours)	0

Data Cleaning Conclusion

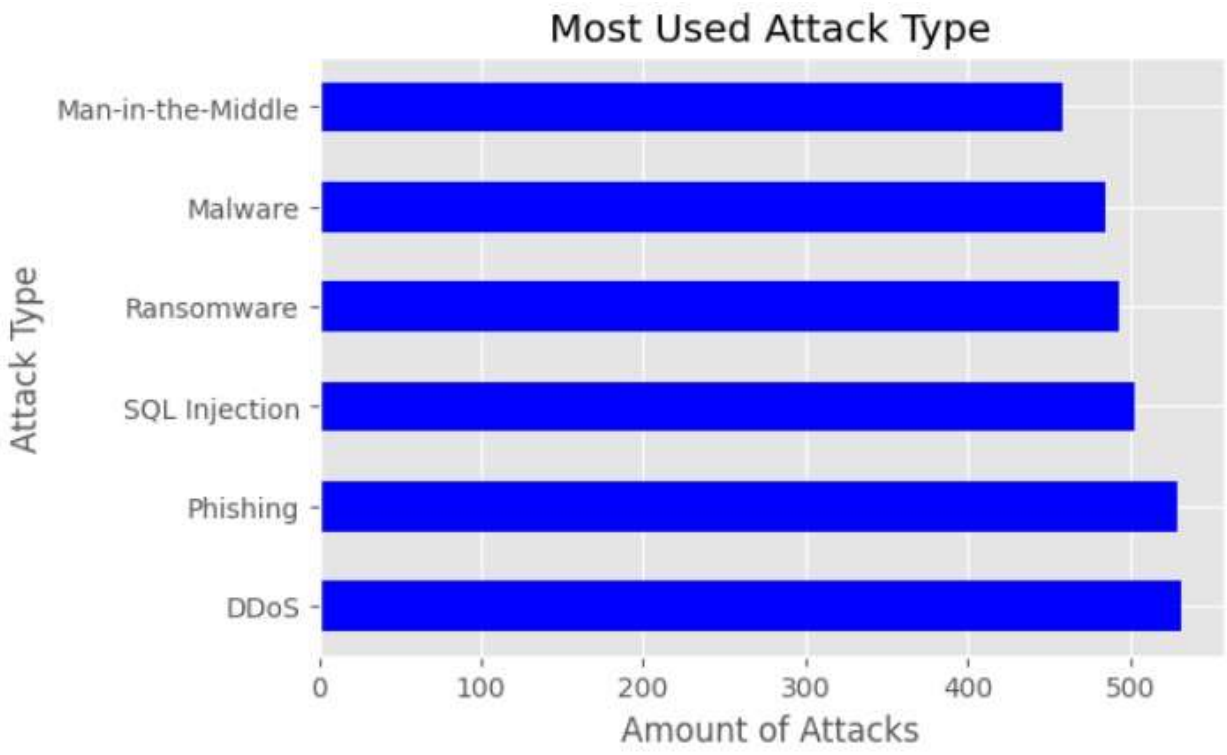
This dataset is expected to be time-consuming and will take half of my time to develop this project. However, after categorizing the columns and rows, filtering any possible duplicates, and finding the various data types, it was clear that the data set had been appropriately vetted for analysis. Therefore, no further data cleaning was necessary on my end. Now, we shall proceed to the data analysis section.

Data Analysis

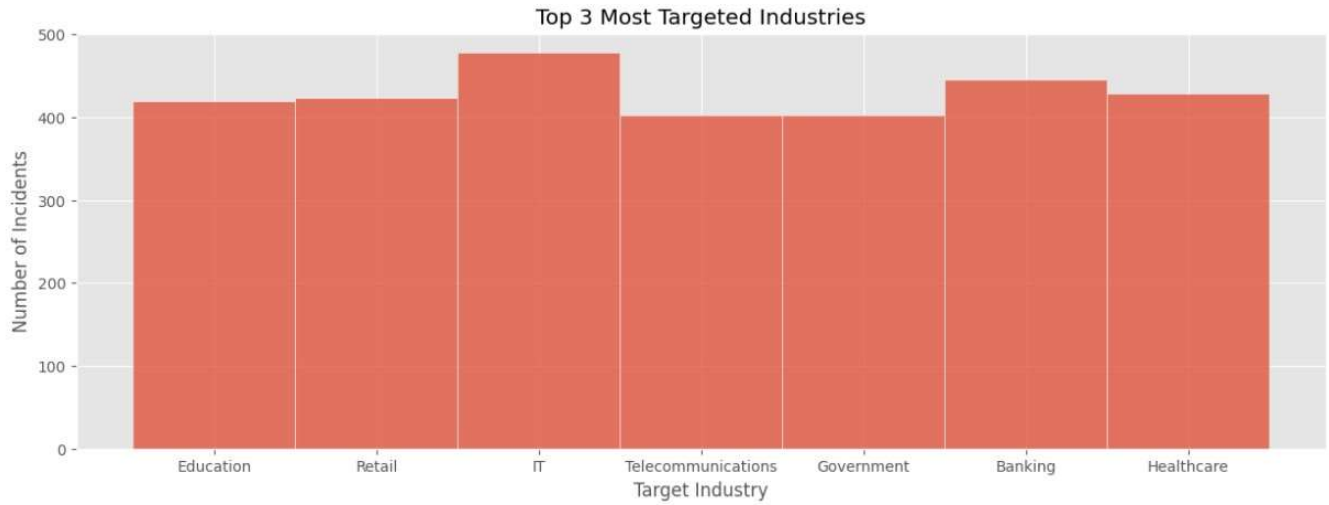
With the formatted dataset that I prepared for study, I developed several infographics to better understand the data and gain valuable insights. Different graphs were necessary to proceed. The first task was to determine which country has the most cybersecurity threats. The United Kingdom led the world in compromised security, with China and the USA having the fewest attacks.



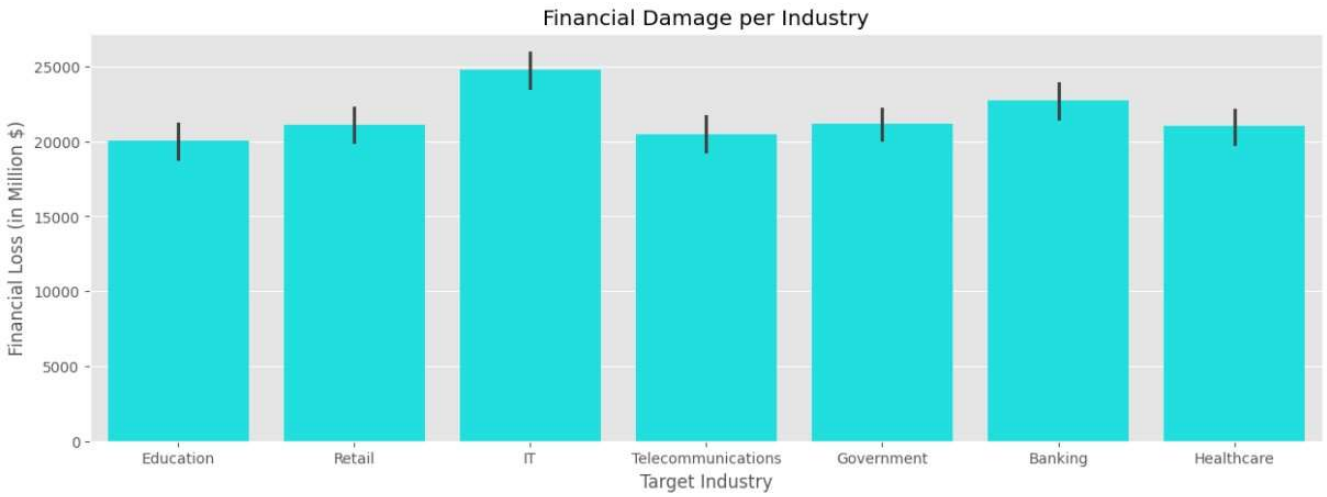
The next task is to determine what weapon cybercriminals use to infiltrate their targets. By combining the “Attack Type”, I can filter the different approaches to hacking and the number of occurrences. As shown below, Distributed Denial of Service (DDoS) and Phishing attacks are the top two threats, with DDoS taking the top spot and Man-in-the-Middle attacks having the least attempts.



Now that we have established the most targeted country and type of attack, the next task is to find the top 3 most targeted industries. The graph I developed below shows that IT is the top industry with the most incidents. It is not surprising as IT is the backbone of technology, data management, and communications for many companies worldwide. Banking comes in second, as expected, as the banking infrastructure holds many financial accounts that can be marked as easy and viable targets for cyber attackers. Then, Healthcare comes in at a close third. There has been a growing trend of attacks on the Healthcare sector in the past several years; this may require a separate and deeper study.

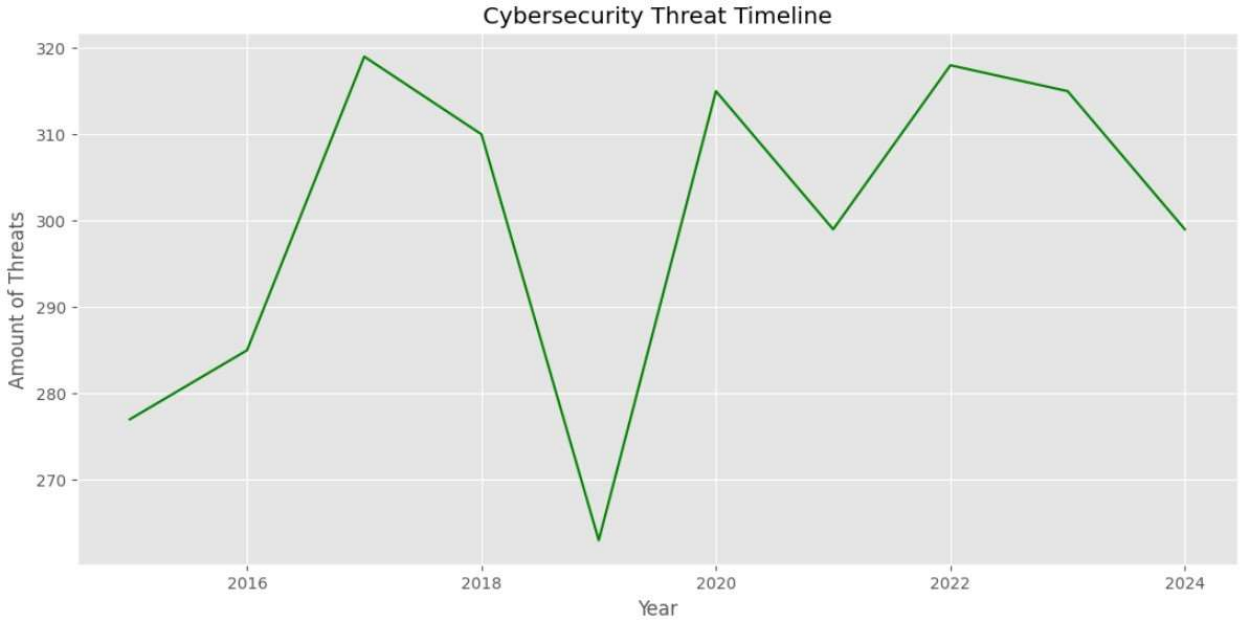


As I dig deeper into the subject of industries, it is vital to consider the financial impacts per category of target. The results showed IT remaining at the top of the rankings, followed again by Banking. Surprisingly, the Government sector and Healthcare are closer in terms of financial damages.

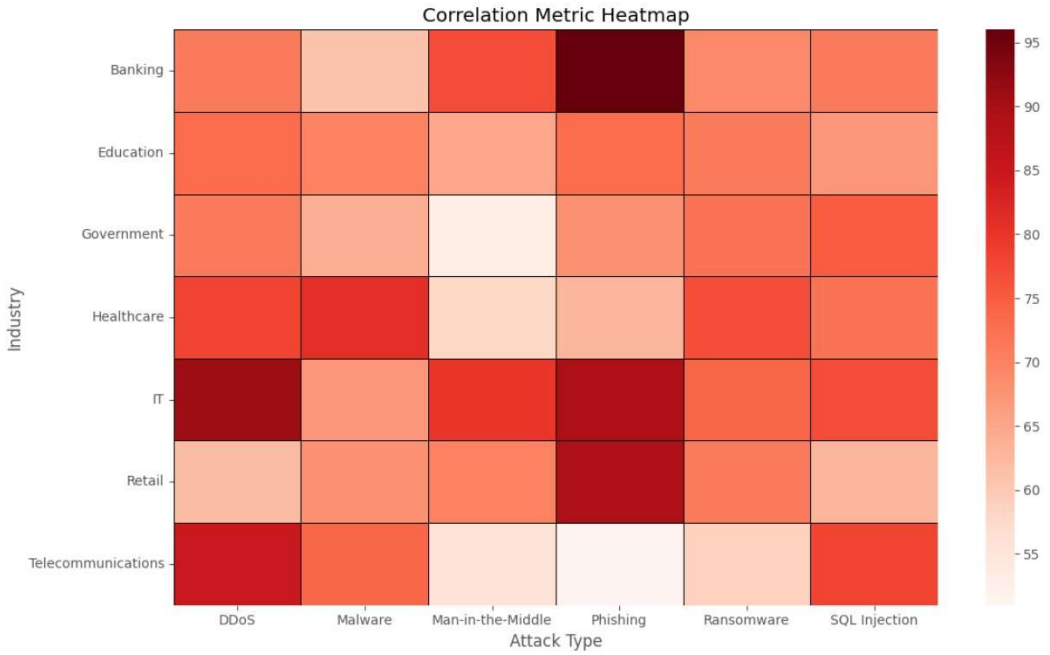


Gaining an overall perspective on yearly changes can be beneficial for predictive analysis. I developed a year-to-year timeline of cybersecurity threats from 2015 to 2024 using a graphical line chart. With this chart, I can visualize trends, patterns, and fluctuations across variables such as time. As shown in the image below, there was a significant uptick in security threats from 2015 to 2017. However, this threat increase was followed by a massive drop from 2017 to 2019. Entering 2020, another increase in attacks occurred. The COVID pandemic may have impacted this immense surge in 2020; further investigation is

required to produce a reliable diagnostic analysis. In 2022, it is crucial to note that the number of threats reached almost matches 2017's numbers.



After gaining insights into the yearly activities of cyber threats, I decided to jump back to industry-related investigations. This time, I wanted to see if I could find any patterns and compare relationships between industries and attack types. Implementing a heatmap method should aid in spotting certain patterns and discrepancies between two categories across a matrix.



The results were interesting, as IT remains the front and center regarding impact related to industries. With the IT industry in mind, DDoS and Phishing were the most prominent attack types. The Banking industry also remains consistent among the targets that have its share of issues with Phishing attacks. The heatmap further confirms the two dominant types of attacks discovered in an earlier bar chart: DDoS and Phishing.

Based on the overall results so far, IT remains the top target industry. The output below displays a correlation between the Target Industry and Security Vulnerability Types. The IT industry again leads almost all the categories of Security Vulnerability Types, with Banking coming in close second place.

Security Vulnerability Type	Social Engineering	Unpatched Software	Weak Passwords	Zero-day
Target Industry				
Banking	117	107	111	110
Education	115	101	94	109
Government	96	97	98	112
Healthcare	112	105	115	97
IT	107	112	122	137
Retail	97	109	102	115
Telecommunications	103	107	88	105

Additional analysis was done to discover an optimum solution for resolving incidents by combining parameters of the Defense Mechanism Used and the least amount of Resolution Time in the IT industry. Despite programming to display at least the top 10 DFU, the results only produced 5. Encryption is the most reliable defense application for IT finance resolution.

	Target Industry	Defense Mechanism Used	Incident Resolution Time (in Hours)	Usage Count
23	IT	Firewall	34.370370	81
21	IT	Antivirus	35.155556	90
20	IT	AI-based Detection	35.485149	101
24	IT	VPN	36.640449	89
22	IT	Encryption	38.427350	117

Conclusion

In summary, the United Kingdom has led in vulnerability to cyber-attacks during the past nine years. Malicious hackers prefer DDoS attacks with phishing as a runner-up. Business fields such as Information Technology, Banking, and Healthcare were impacted the most as they led the categories of targeted industries and financial damages. Through the years, the number of cyber threats fluctuated and did not remain consistent, indicating the lack of a steady or linear trend.

This final analysis culminates with several realizations that can benefit future network security assessments. It makes sense that the same industries that cybercriminals focused on were also the same areas that suffered the most financial loss. The year 2019 in the timeline truly piqued my interest, additional research is required to reveal insights into the drastic drop in cyber threats just a year before the Covid pandemic began. Security professionals must focus on intercepting DDoS and phishing attacks as they continue to cause substantial damage. Since IT was consistently the primary target, enhancing security policies on social engineering, unpatched software, weak passwords, and zero-day should be paramount. Furthermore, adding robust annual employee training and enhancing network security standards should help mitigate these issues. One area that was missing representation was Artificial Intelligence and the impact it has on cyber security. Perhaps that is saved for a separate research project in the future.