**HDSC August’22 Premiere Project**

**TEAM TABLEAU**

**Cyber Security Risk (2022 CISA Vulnerability)**

Cyber attacks are on the increase in recent time. Efforts have been made to mitigate cyber security risk in order to protect vital data. It can be quite difficult to be protected from cyber security risks if one lacks vital information about recent vulnerabilities, or fails to understand the meaning of available information on cyber security risk. The dataset gathered from CISA Known Exploited Vulnerabilities catalog for 2022 provides important information on the latest cyber security risk across the United States such as specific vulnerability name, the vendors attacked, the products attacked, the severity and complexity of the vulnerabilities, and other helpful information about cyber security risks. This available information will be very helpful in averting cyber security risk, but it must be nicely prepared and properly presented to be useful.

**Aim and Objectives**

The aim of this project is to prepare the available and analyze it to gain useful insight for data driven decision making.

The following objectives will be achieved from the project.

* Analyze the severity of different security vulnerabilities over time in order to measure effectiveness of cyber security efforts.
* To carry out exploratory data analysis that will show severity distribution, average CVSS scores of products, vendor projects and products names associated with security vulnerabilities, and vectors of vulnerability in the US.
* To create visualization dashboard of different metrics from the dataset that will help to make insightful decision that will aid in avoiding cyber security risks.

**FLOW PROCESS**

**Data Sourcing**

The dataset for this project was obtained from Kaggle via the link below

<https://www.kaggle.com/datasets/thedevastator/exploring-cybersecurity-risk-via-2022-cisa-vulne>

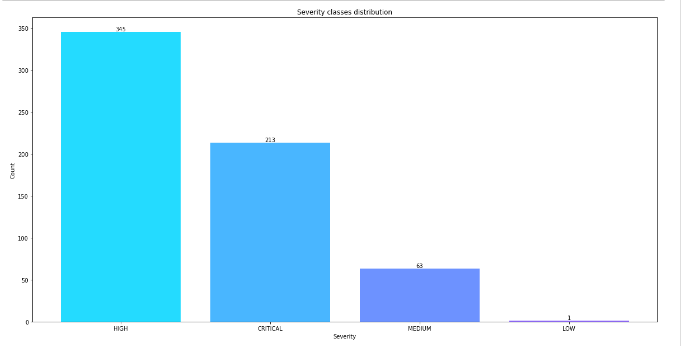
**Data Preparation:**

* **Data collection:** At this stage, data on USA cyber security risk for 2022 were taken kaggle, which is free data source platform. The collected data had 5 different dataset.
* **Data discovery and profiling:**Pandas profiling was done on each of the dataset. It was discovered that all the dataset had exactly the same number of columns and column heading. It was also discovered that most rows are exactly the same across the 5 dataset. The profiling also showed that rows of a particular id may have missing values in one dataset but no missing values in another dataset.
* **Data cleaning:**The entire 5 dataset were combined into one dataset. This gave rise to many duplicate rows. The combined dataset was sorted in order of rows without missing values. Furthermore, all duplicate rows were removed. It was noticed that the ‘notes’ column had value in only one row in the entire dataset, hence it was removed.
* **Data structuring:**The CSV (comma-separated values) format which was the default format of the data was maintained.
* **Data transformation:**The data was further transformed to make it more useful for analysis. The ‘severity’ and ‘complexity’ columns where duplicated and transformed into whole number data types. New columns for days, months, and years were created from the date\_added column.

**Exploratory Data Analysis**

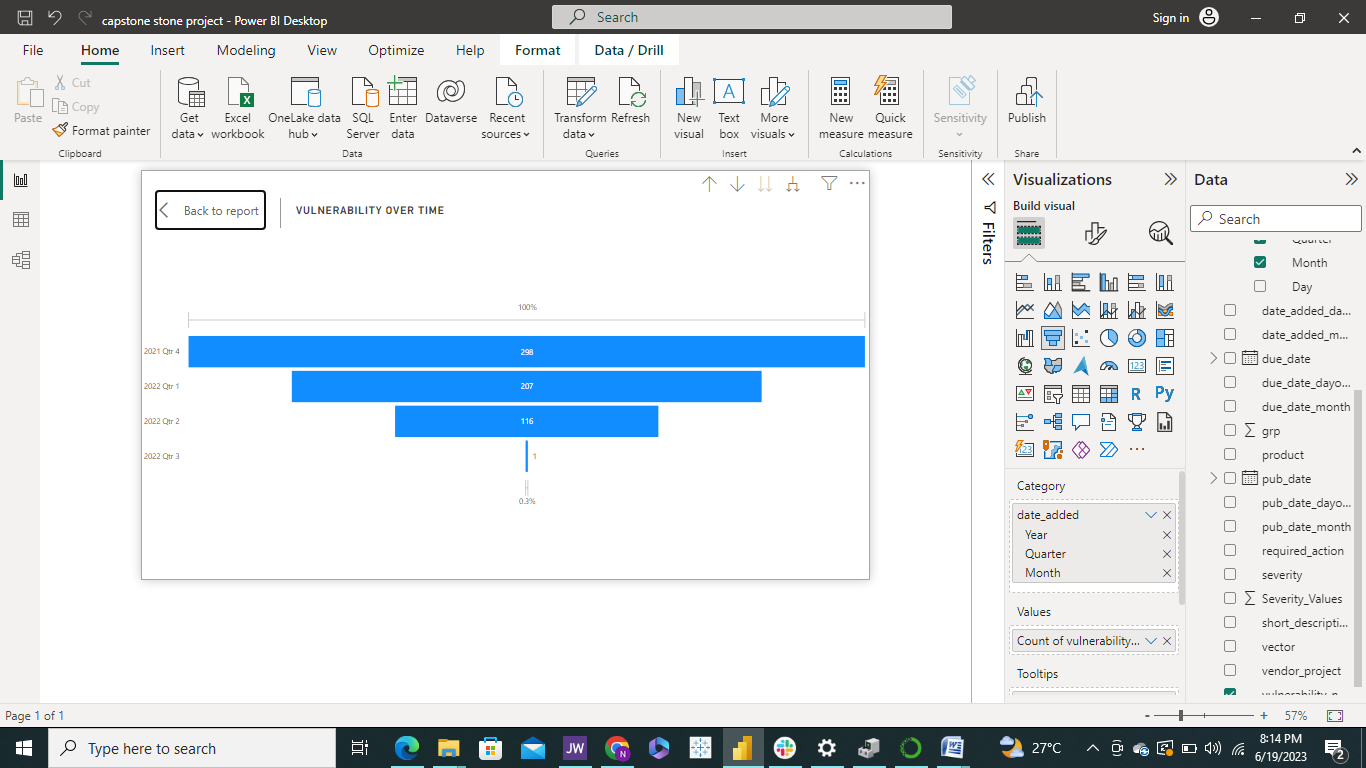
The data was further explored to gain insight on the effectiveness of the efforts made to overtime to mitigate the vulnerabilities.

Firstly, a severity distribution bar chart was created to display the quantity of public exploits that are critical, high, medium, or low in terms of severity. It was clearly seen that most are highly severe.

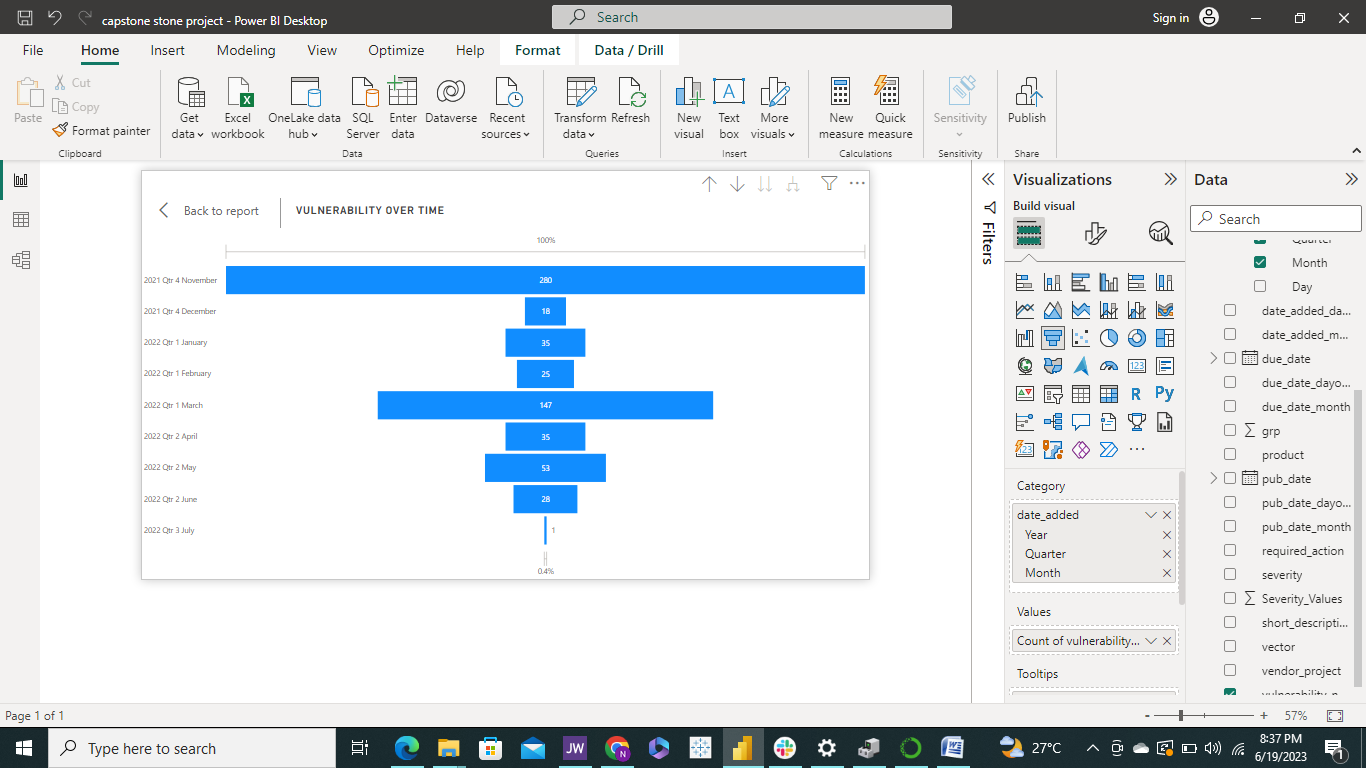
****

*Figure 1: Severity Distribution*

A funnel was used to view the spread of the vulnerability overtime and it was seen that in the fourth quarter of 2021 which was when the data was initially collected, there were 298 vulnerabilities, but it reduced to 207 and 116 in the first and second quarter of 2022 respectively. As at the third quarter of 2022, one vulnerability was untreated.

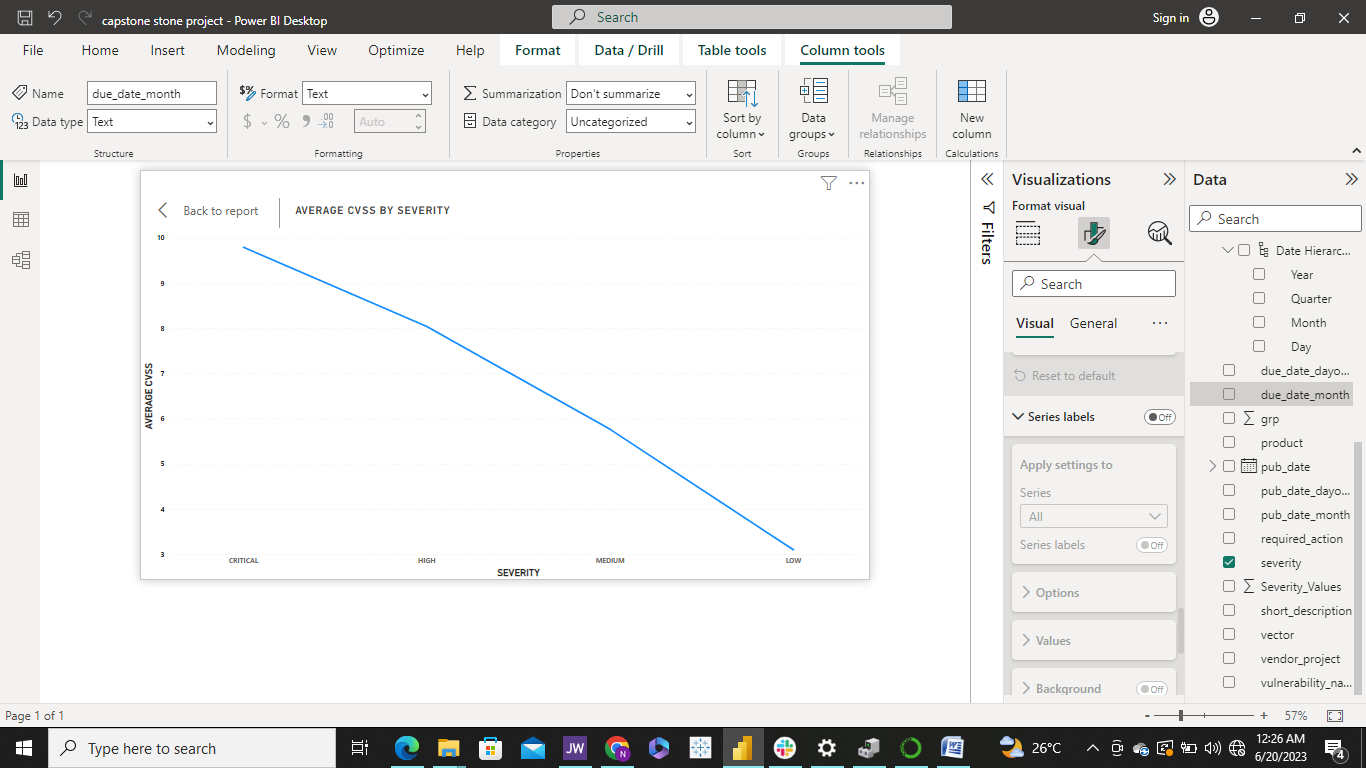


*Figure 2: Vulnerability over time by Quarter*

**

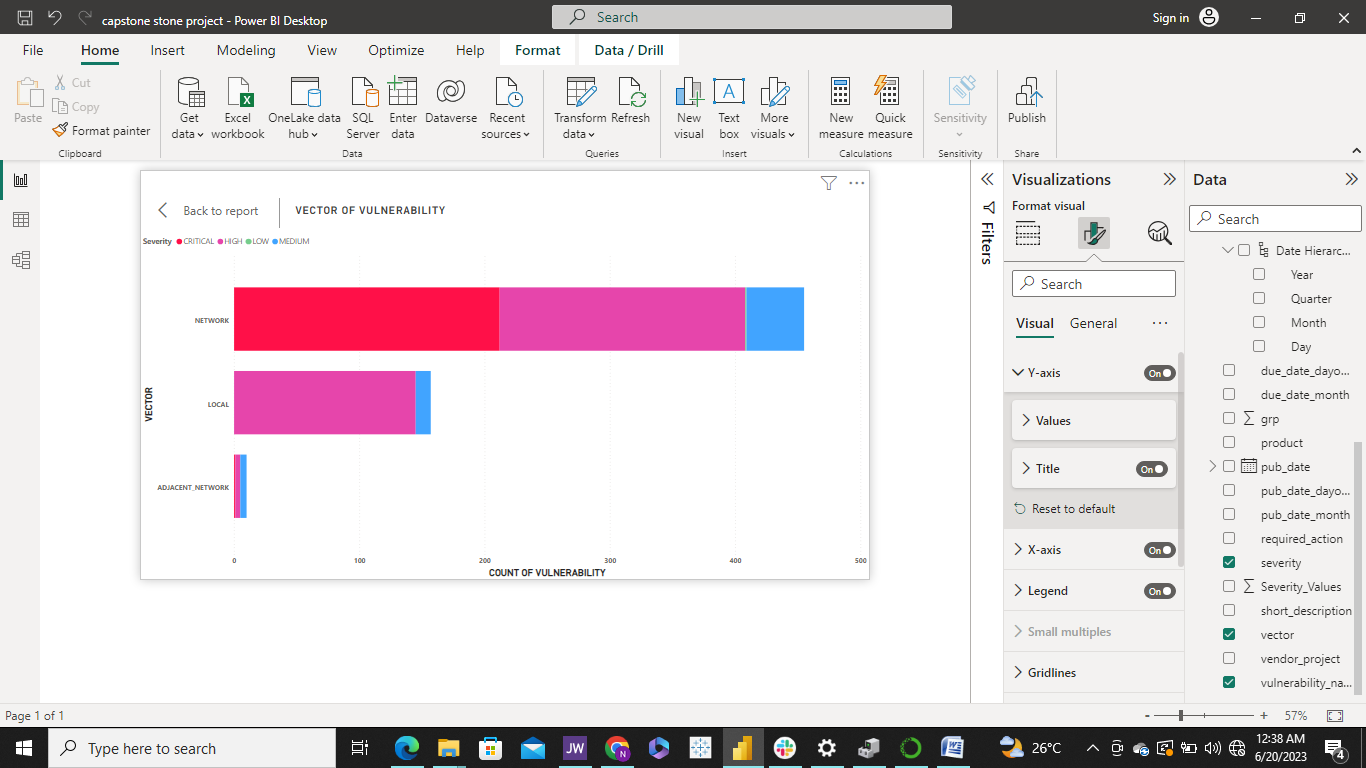
*Figure 3: Vulnerability over time by months*

A line graph of average CVSS by severity shows that the higher the CVSS score the more severe the vulnerability is.



*Figure 4: Vulnerability over time by months*

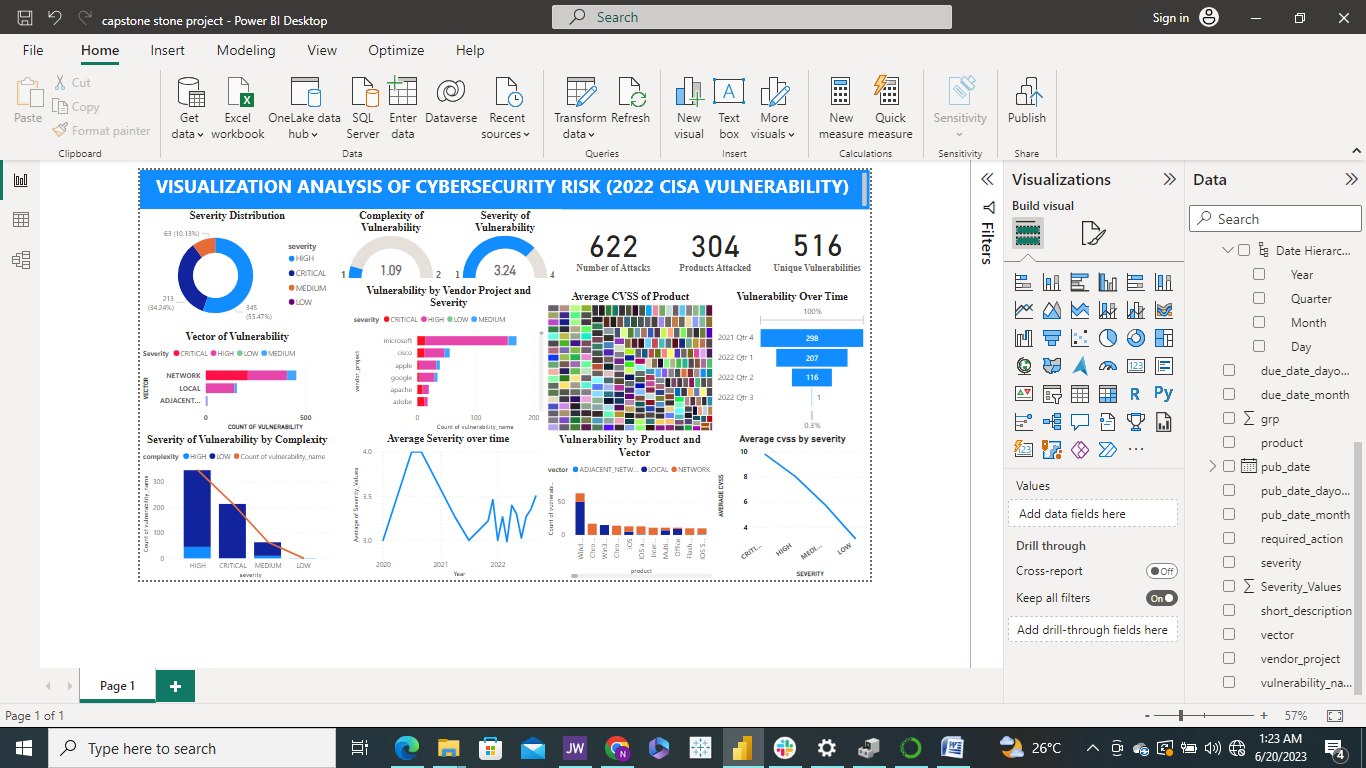
Further insight into the data by plotting vector against the count of associated vulnerabilities showed that network is the number one source of vulnerabilities, and most critical and highly severe public exploit are through the network.

**

*Figure 5: Vectors of vulnerability*

**Data Visualization Dashboard**

A visualization dashboard to display different data metrics and relationships was achieved using Power BI.

**

*Figure 6: PowerBI Visualization Dashboard*

Results

The project has successfully shown that actions taken again vulnerabilities between 2020 and 2022 were very effective as the vulnerability reduced drastically over time. The project has successfully explored severity distribution, average CVSS scores of products, vendors and products associated with vulnerabilities, and vectors of vulnerability in the US. Lastly, the project was able to produce a live visualization dashboard of vital metrics for drawing insights from the dataset.

Conclusion and Recommendation

The analysis showed clearly that Microsoft products are the most vulnerable product, but this was not a big issue as most of the exploits are of low complexity. The higher the CVSS of an exploit, the more severe the public exploit will be. Debian\_Specific Redis servers have the highest average CVSS of 10.0, and were therefore exposed to exploit with most critical severity. Cyber security risk is most likely to happen via network as this was shown to be the highest vulnerability vector. Finally the analysis showed that vulnerabilities reduced overtime drastically, showing the all applied required actions were very effectively in mitigating cyber security risks