HDSC August'22 Premiere Project

TEAM TABLEAU

CYBERSECURITY RISK ANALYSIS

Project Description

Something no one wants is to be exposed or feel insecure about personal information while surfing the web. The task is to analyze the severity of different security vulnerabilities over time in order to measure effectiveness of cybersecurity efforts, build a data visualization tool to showcase the different types, CVSS scores, vendor projects and products names associated with security vulnerabilities over time

About the Dataset

The dataset presents an in-depth exploration of the security vulnerabilities across the United States from the CISA Known Exploited Vulnerabilities catalog for 2022. You'll find a number of vulnerability types and severity levels, as well as CVSS scores, vendor projects, product names, and other pertinent information.

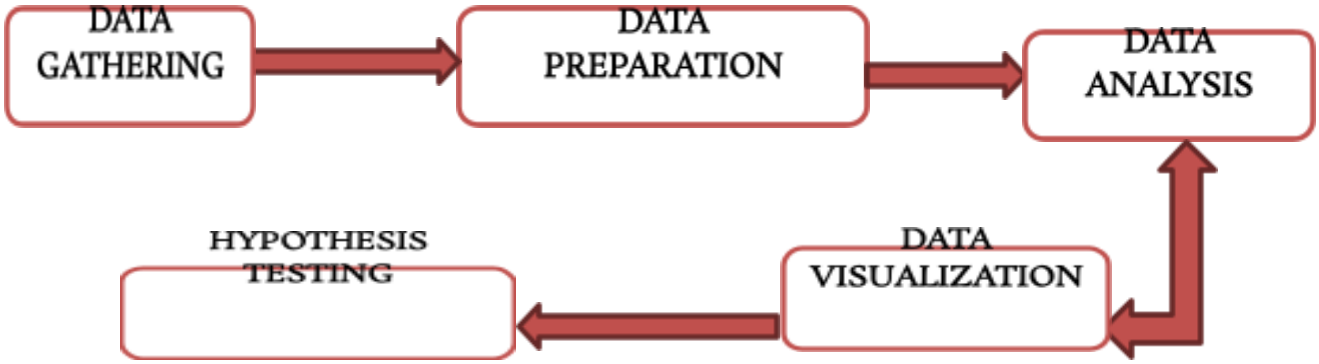
Aim and Objective

The goal of this project is

1. Analyze the severity of different security vulnerabilities over time in order to measure effectiveness of cybersecurity efforts.
2. To build data visualization graphs to showcase the different types, CVSS scores, vendor projects and products names associated with security vulnerabilities over time in the US.
3. Perform statistical analysis to identify the relationship between vulnerabilities associated variables

Tools: During the course of analysis, we made use of Jupiter Notebook, Microsoft Excel and Microsoft Power BI

FLOW PROCESS



Data Gathering

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 Compilation: we were provided with five datasets for different dates which were compiled together into one single dataset
- Data Discovery: we noticed early on that the five datasets contain the same entries with few differences causing the compiled dataset to have duplicate 'cve_id's.
- Data Cleaning: This involves grouping the compiled dataset by the 'cve_id' and filling in missing values with the corresponding value found in the same column and later dropping rows with duplicated 'cve_id'. The remaining missing values were dropped as well as data that do not contribute to the overall performance of the analysis
- Data Transformation: The date columns were converted into datetime objects and from them new columns were created for days, years and months

Data Analysis

During our exploration, we were able to answer questions such as

1. What are the top 5 products with the most vulnerabilities

Product	**No of Vulnerabilities**
Windows	63
Chromium V8 Engine	17
Win32k	15
Chrome	14
iOS	13

Table 1: Top 5 Products with the most vulnerabilities

2. What are the top 5 vendors with the most vulnerabilities

Vendor	**No of Vulnerabilities**
Microsoft	171
cisco	56
apple	39
google	35
apache	20

Table 2: Top 5 vendors with the most vulnerabilities

3. How many vulnerabilities were added in the year 2022 and 2021

Year	***No of Vulnerabilities***
2022	2429
2021	1555

Table 3: No of vulnerabilities added in each year

4. What day was the highest number of vulnerability added
   '2021-11-03' - 276

And so on

Data Visualization

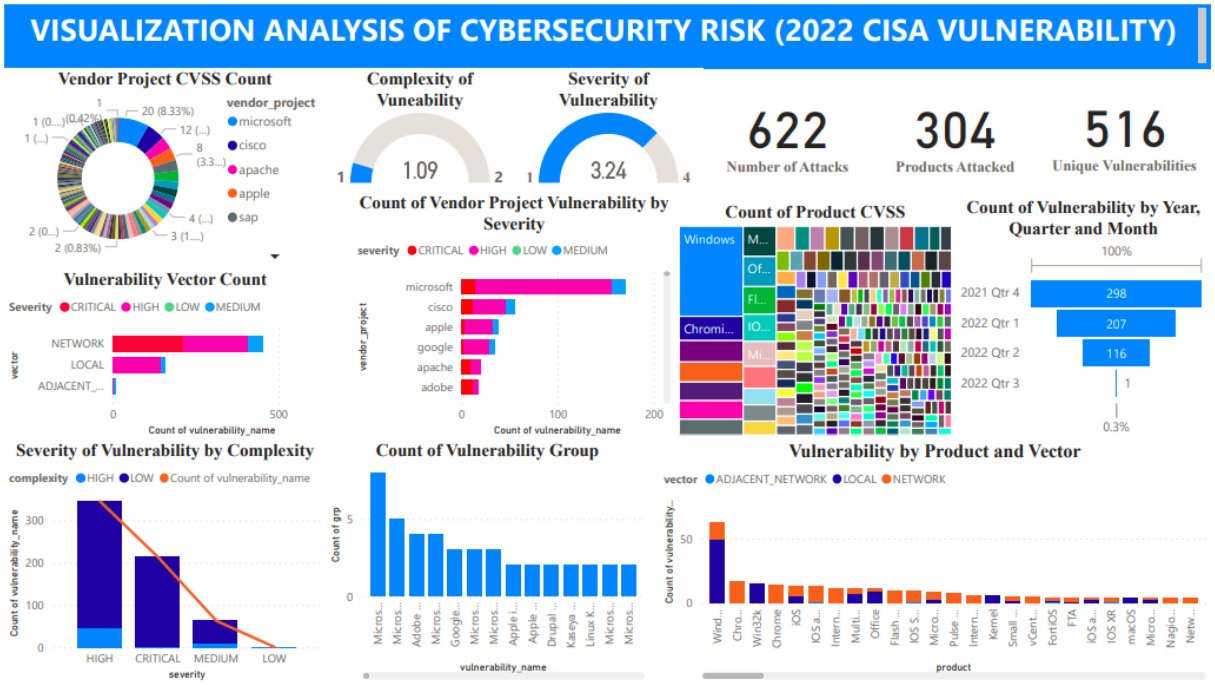


Figure 1: PowerBI dashboard

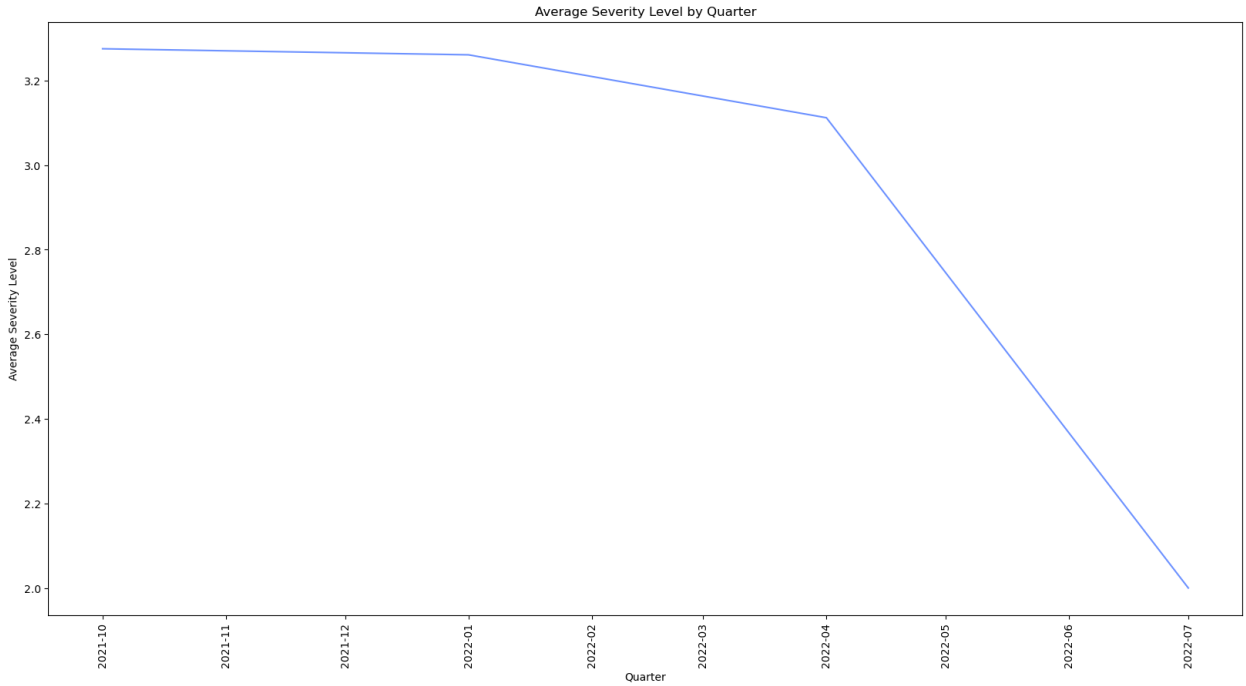


Figure 2: Average severity level by Quater

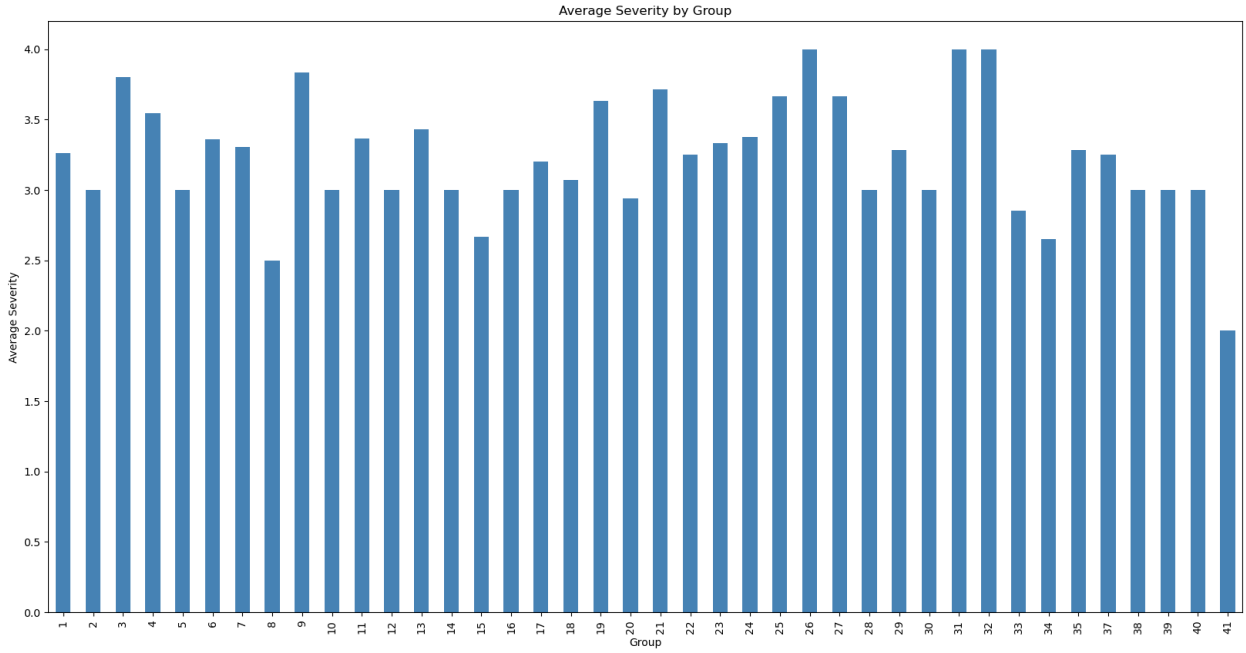


Figure 3: Average severity level by group

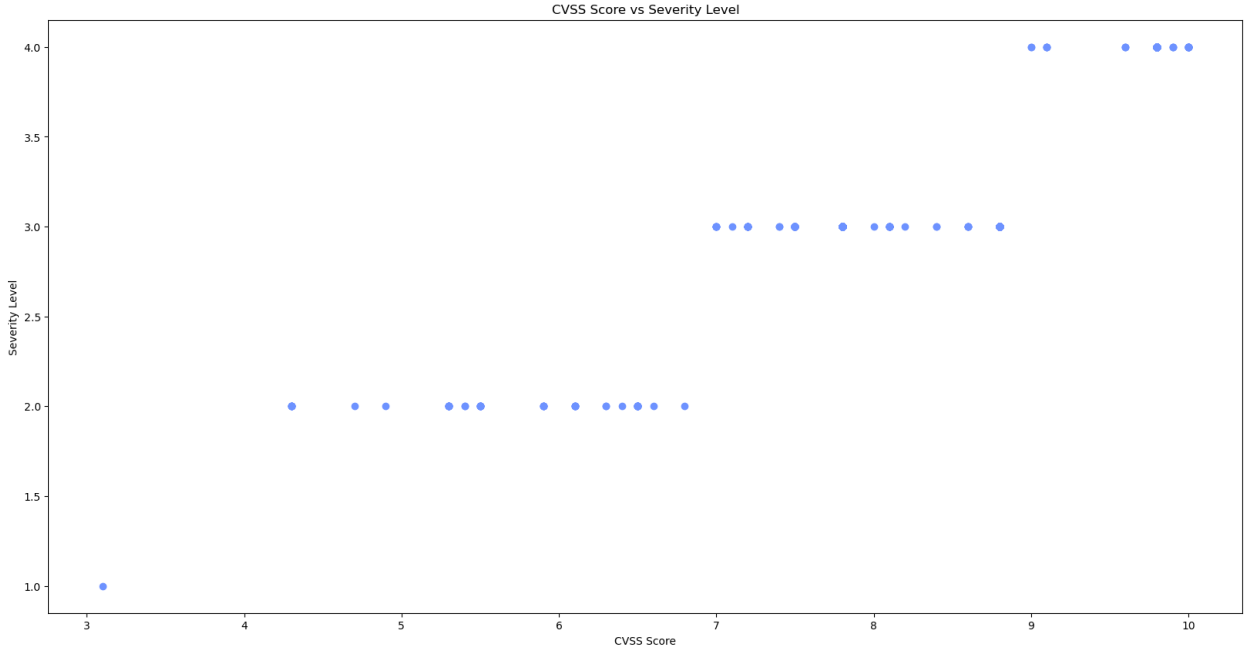


Figure 4: Severity level VS CVSS score

Hypothesis Testing

We performed some statistical analysis to test the relationship between the associated variables of the vulnerabilities

H1: Vulnerabilities with higher CVSS scores are more likely to have a higher severity level.
Correlation coefficient between CVSS score and severity level is 0.9270402556785979

This indicates a strong positive correlation between CVSS score and severity level. This means that as CVSS score increases, severity level increases. In other words, more severe vulnerabilities tend to have higher CVSS scores.

H2: Vulnerabilities discovered earlier (date_added) are more likely to have a higher severity level.
Performing one-way ANOVA to test for differences between groups gives the result
One-way ANOVA Results:
F-Statistic: 3.328409606331009
p-value: 0.019316634896811 24

This means that there is a statistically significant difference between the average severity levels of the four quarters. In other words, the data suggests that the average severity level of vulnerabilities increased from 2021Q4 to 2022Q3.

H3: Vulnerabilities in certain CWE categories are more likely to have a higher severity level.
By performing a chi-square test for association between CWE category and severity level
Chi-Square Test Results:
Chi-Square: 441.3452724662253
p-value: 3.329756781851147e-16

The analysis suggests that the CWE category is related to the severity levels of vulnerabilities. The average severity levels vary across different CWE categories, and the chi-square test indicates a significant association between CWE category and severity

H4: Vulnerabilities with a shorter time frame between discovery (date_added) and the due date are more likely to have a higher severity level.

By performing ANOVA to evaluate the differences in severity levels across time frames

ANOVA Results:

F-Statistic: 2.557321207214212

p-value: 0.07833084700446755

Based on the analysis we conducted to test the hypothesis that vulnerabilities with a shorter time frame between published and the due date are more likely to have a higher severity level, the results suggest that there is no strong evidence to support this hypothesis.