Pranav Tejasvi Adiraju (padiraju@umd.edu)
Srikanth Parvathala (psrikant@umd.edu)

Final Project Report

INST 754 - Data Integration & Preparation for Analytics

Topic: Cyber Risk Insight: Analyzing IT Vulnerabilities and Predicting Threat Periods

1. Business Challenge:

Our project aims at understanding how specific IT vendors and products face unforeseen challenges with common vulnerabilities and how we can aim to help these IT vendors strengthen their cyber risk defenses.

2. Dataset Link:

Here is the link of our dataset:
https://www.kaggle.com/datasets/andrewkronser/cve-common-vulnerabilities-and-exposures

3. Dataset Description:

We identified a dataset comprising the Common Vulnerabilities and Exposures (CVE) provided by the MITRE corporation's National Cybersecurity FFRDC. This dataset, which is available on Kaggle, comprises of 4 CSV files (cve, products, vendor_product, vendors) that provide insights into known software vulnerabilities, including severity levels (as determined by the Common Vulnerability Scoring System (CVSS)), and link them to specific IT products and vendors.

Furthermore, CVE corresponds to ***"Common Vulnerabilities and Exposures"***. CVEs are a vital component of many cybersecurity safeguards and have been identified in an extensive variety for the safety of critical IT products and services. They serve as essential to evaluating the threat landscape and guaranteeing that IT infrastructures are safeguarded against recognized vulnerabilities.

4. Research Questions:

Here are the three questions that our dataset would be answering:

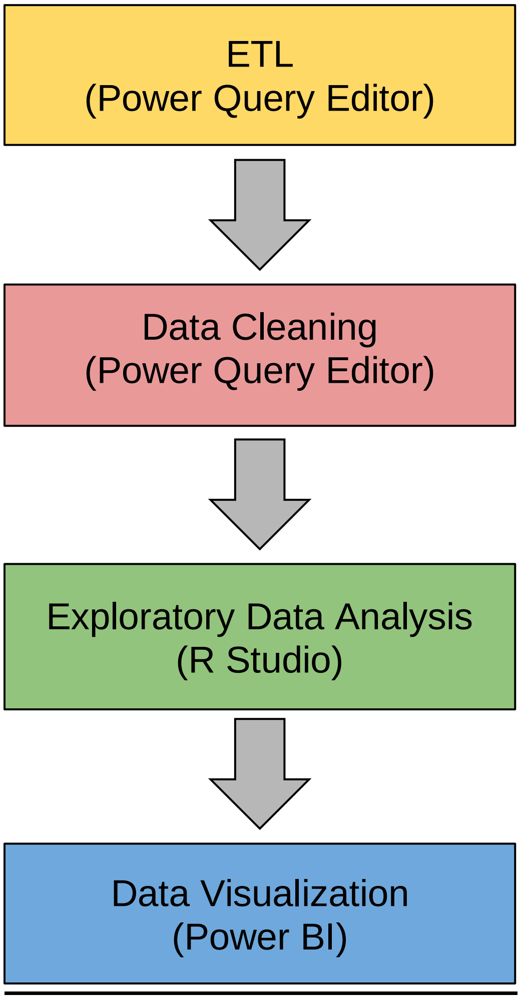
1. ***Time Series Analysis of High/Severe Vulnerabilities over time.***
2. ***Vulnerabilities by Access Complexity, Impact, CVSS scores.***
3. ***Identifying Vendors/Products based on Vulnerability.***

5. Tools/Platforms Used:

We have used the following tools in our project phase:

1. **R Studio** - Data Quality Check, Exploratory Analysis
2. **PowerQuery** - Data Type Updation, Data Curation, Data Integration & Preparation.
3. **Power BI** - Enhanced Visualizations to answer our research questions.

6. Process Involved:



7. Process Steps:

Our process involved the following steps:

1. Data Extraction, Transformation & Loading (ETL)
2. Data Cleaning
3. Exploratory Data Analysis

Data Extraction, Transformation & Load (ETL) Process:

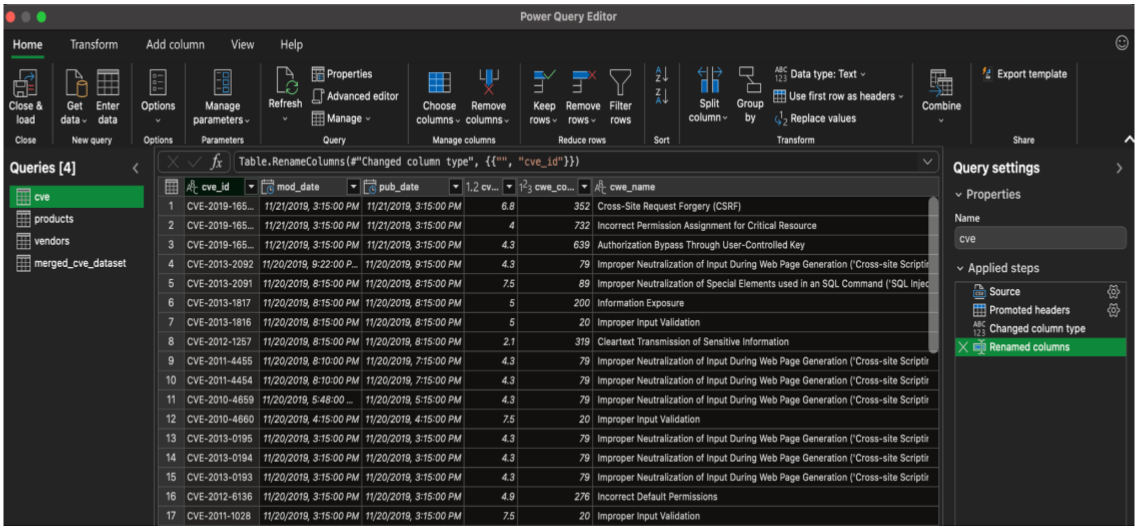
Our project was started with the ETL process where we utilized PowerBI Query Editor to transform each of our dataset by assigning the headers, modifying the column names and fixing the data type references.

We then used joins to merge two datasets Products and Vendors dataset using a common ***cve_id*** (reference column). Subsequently, we combined this dataset with the cve dataset using the ***cve_id*** again (reference column).

Data Cleaning Process:

In the Data Cleaning Phase, we have filtered out all the ***NULL*** values and the outliers in the columns of our merged dataset for a robust analysis and visualizations. Our final merged dataset consisted of ***241, 979 rows*** with ***15 columns.***

Here are the results from our Data Preparation Efforts:



Queries [4]

- cve
- products
- vendors
- merged_cve_dataset

```
Table.RenameColumns(#"Changed column type", {{"", "cve_id"}})
```

	cve_id	mod_date	pub_date	1.2 cv...	1²3 cwe_co...	cwe_name
1	CVE-2019-165...	11/21/2019, 3:15:00 PM	11/21/2019, 3:15:00 PM	6.8	352	Cross-Site Request Forgery (CSRF)
2	CVE-2019-165...	11/21/2019, 3:15:00 PM	11/21/2019, 3:15:00 PM	4	732	Incorrect Permission Assignment for Critical Resource
3	CVE-2019-165...	11/21/2019, 3:15:00 PM	11/21/2019, 3:15:00 PM	4.3	639	Authorization Bypass Through User-Controlled Key
4	CVE-2013-2092	11/20/2019, 9:22:00 P...	11/20/2019, 9:15:00 PM	4.3	79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scriptir
5	CVE-2013-2091	11/20/2019, 8:15:00 PM	11/20/2019, 8:15:00 PM	7.5	89	Improper Neutralization of Special Elements used in an SQL Command ('SQL Injec
6	CVE-2013-1817	11/20/2019, 8:15:00 PM	11/20/2019, 8:15:00 PM	5	200	Information Exposure
7	CVE-2013-1816	11/20/2019, 8:15:00 PM	11/20/2019, 8:15:00 PM	5	20	Improper Input Validation
8	CVE-2012-1257	11/20/2019, 8:15:00 PM	11/20/2019, 8:15:00 PM	2.1	319	Cleartext Transmission of Sensitive Information
9	CVE-2011-4455	11/20/2019, 8:10:00 PM	11/20/2019, 7:15:00 PM	4.3	79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scriptir
10	CVE-2011-4454	11/20/2019, 8:10:00 PM	11/20/2019, 7:15:00 PM	4.3	79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scriptir
11	CVE-2010-4659	11/20/2019, 5:48:00 ...	11/20/2019, 5:15:00 PM	4.3	79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scriptir
12	CVE-2010-4660	11/20/2019, 4:15:00 PM	11/20/2019, 4:15:00 PM	7.5	20	Improper Input Validation
13	CVE-2013-0195	11/20/2019, 3:15:00 PM	11/20/2019, 3:15:00 PM	4.3	79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scriptir
14	CVE-2013-0194	11/20/2019, 3:15:00 PM	11/20/2019, 3:15:00 PM	4.3	79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scriptir
15	CVE-2013-0193	11/20/2019, 3:15:00 PM	11/20/2019, 3:15:00 PM	4.3	79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scriptir
16	CVE-2012-6136	11/20/2019, 3:15:00 PM	11/20/2019, 3:15:00 PM	4.9	276	Incorrect Default Permissions
17	CVE-2011-1028	11/20/2019, 3:15:00 PM	11/20/2019, 3:15:00 PM	7.5	20	Improper Input Validation

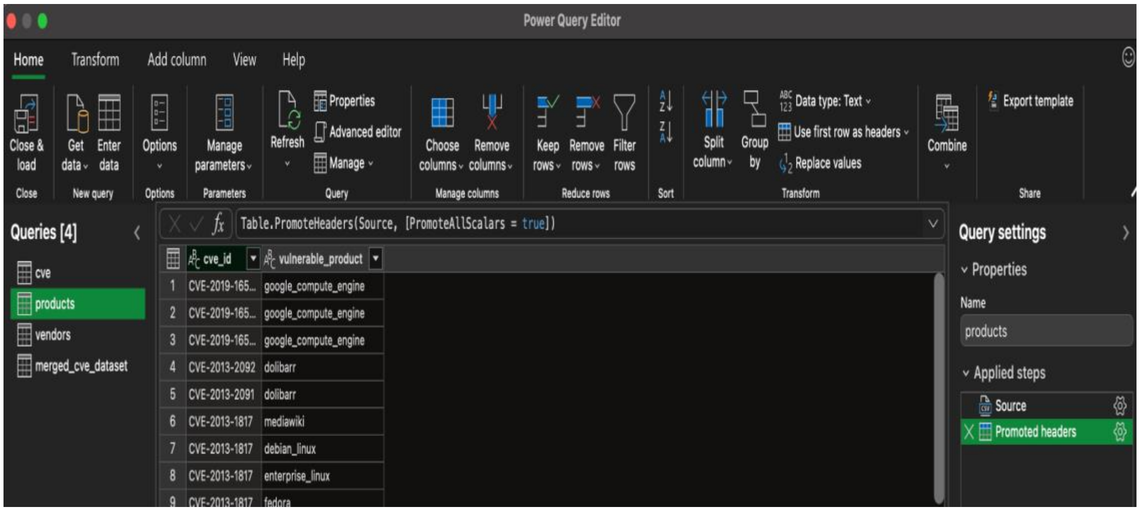
Query settings

Properties

Name: cve

Applied steps

- Source
- Promoted headers
- Changed column type
- Renamed columns



Queries [4]

- cve
- products
- vendors
- merged_cve_dataset

`Table.PromoteHeaders(Source, [PromoteAllScalars = true])`

	cve_id	vulnerable_product
1	CVE-2019-165...	google_compute_engine
2	CVE-2019-165...	google_compute_engine
3	CVE-2019-165...	google_compute_engine
4	CVE-2013-2092	dolibarr
5	CVE-2013-2091	dolibarr
6	CVE-2013-1817	mediawiki
7	CVE-2013-1817	debian_linux
8	CVE-2013-1817	enterprise_linux
9	CVE-2013-1817	fedora

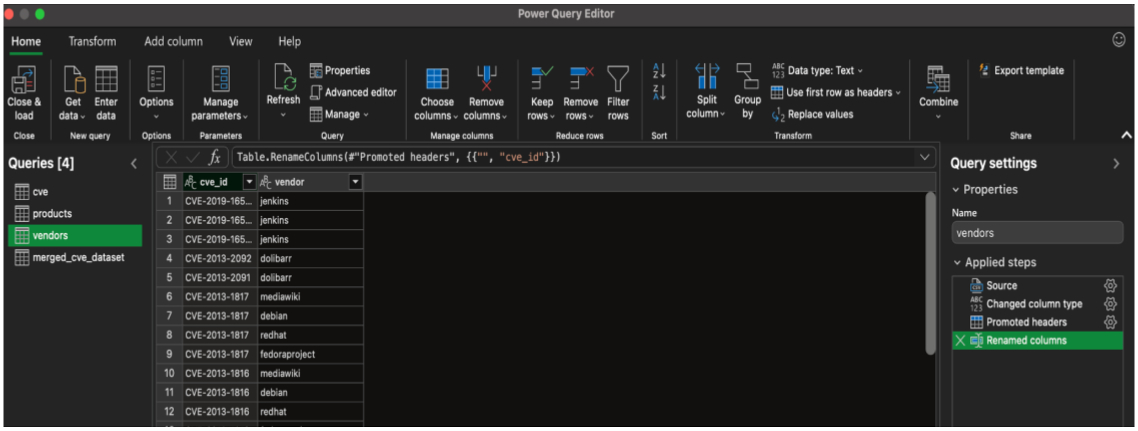
Query settings

Properties

Name: products

Applied steps

- Source
- Promoted headers



Queries [4]

- cve
- products
- vendors
- merged_cve_dataset

`Table.RenameColumns(#"Promoted headers", {{"", "cve_id"}})`

	cve_id	vendor
1	CVE-2019-165...	jenkins
2	CVE-2019-165...	jenkins
3	CVE-2019-165...	jenkins
4	CVE-2013-2092	dolibarr
5	CVE-2013-2091	dolibarr
6	CVE-2013-1817	mediawiki
7	CVE-2013-1817	debian
8	CVE-2013-1817	redhat
9	CVE-2013-1817	fedoraproject
10	CVE-2013-1816	mediawiki
11	CVE-2013-1816	debian
12	CVE-2013-1816	redhat

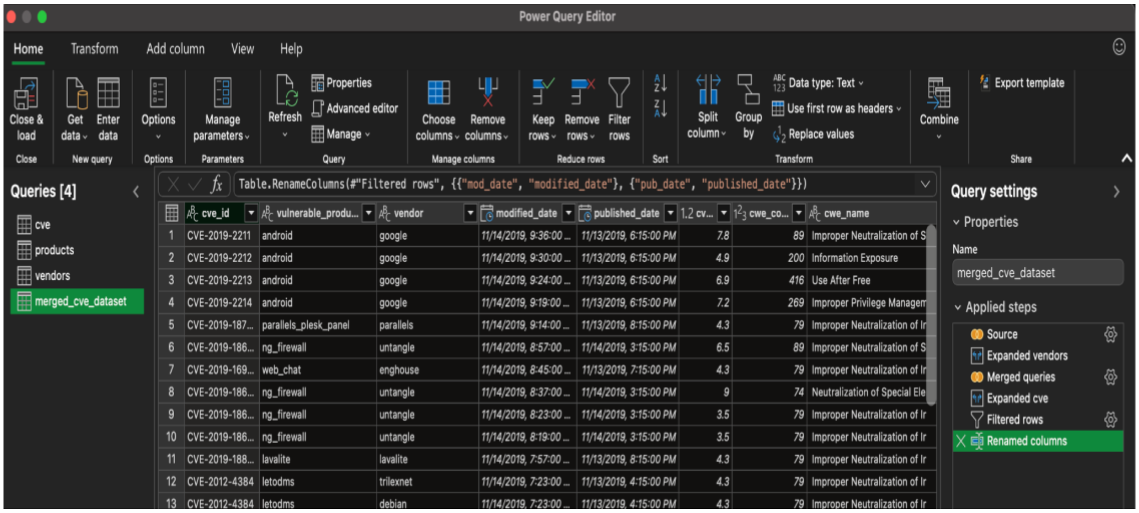
Query settings

Properties

Name: vendors

Applied steps

- Source
- Changed column type
- Promoted headers
- Renamed columns



Queries [4]

- cve
- products
- vendors
- merged_cve_dataset

`Table.RenameColumns(#"Filtered rows", {{"mod_date", "modified_date"}, {"pub_date", "published_date"}})`

	cve_id	vulnerable_produ...	vendor	modified_date	published_date	cv...	cwe_co...	cwe_name
1	CVE-2019-2211	android	google	11/14/2019, 9:36:00 ...	11/13/2019, 6:15:00 PM	7.8	89	Improper Neutralization of S
2	CVE-2019-2212	android	google	11/14/2019, 9:30:00 ...	11/13/2019, 6:15:00 PM	4.9	200	Information Exposure
3	CVE-2019-2213	android	google	11/14/2019, 9:24:00 ...	11/13/2019, 6:15:00 PM	6.9	416	Use After Free
4	CVE-2019-2214	android	google	11/14/2019, 9:19:00 ...	11/13/2019, 6:15:00 PM	7.2	269	Improper Privilege Managem
5	CVE-2019-187...	parallels_plesk_panel	parallels	11/14/2019, 9:14:00 ...	11/13/2019, 8:15:00 PM	4.3	79	Improper Neutralization of Ir
6	CVE-2019-186...	ng_firewall	untangle	11/14/2019, 8:57:00 ...	11/14/2019, 3:15:00 PM	6.5	89	Improper Neutralization of S
7	CVE-2019-169...	web_chat	enghouse	11/14/2019, 8:45:00 ...	11/13/2019, 7:15:00 PM	4.3	79	Improper Neutralization of Ir
8	CVE-2019-186...	ng_firewall	untangle	11/14/2019, 8:37:00 ...	11/14/2019, 3:15:00 PM	9	74	Neutralization of Special Ele
9	CVE-2019-186...	ng_firewall	untangle	11/14/2019, 8:23:00 ...	11/14/2019, 3:15:00 PM	3.5	79	Improper Neutralization of Ir
10	CVE-2019-186...	ng_firewall	untangle	11/14/2019, 8:19:00 ...	11/14/2019, 3:15:00 PM	3.5	79	Improper Neutralization of Ir
11	CVE-2019-188...	lavalite	lavalite	11/14/2019, 7:57:00 ...	11/13/2019, 8:15:00 PM	4.3	79	Improper Neutralization of Ir
12	CVE-2012-4384	letodms	trilexnet	11/14/2019, 7:23:00 ...	11/13/2019, 4:15:00 PM	4.3	79	Improper Neutralization of Ir
13	CVE-2012-4384	letodms	debian	11/14/2019, 7:23:00 ...	11/13/2019, 4:15:00 PM	4.3	79	Improper Neutralization of Ir

Query settings

Properties

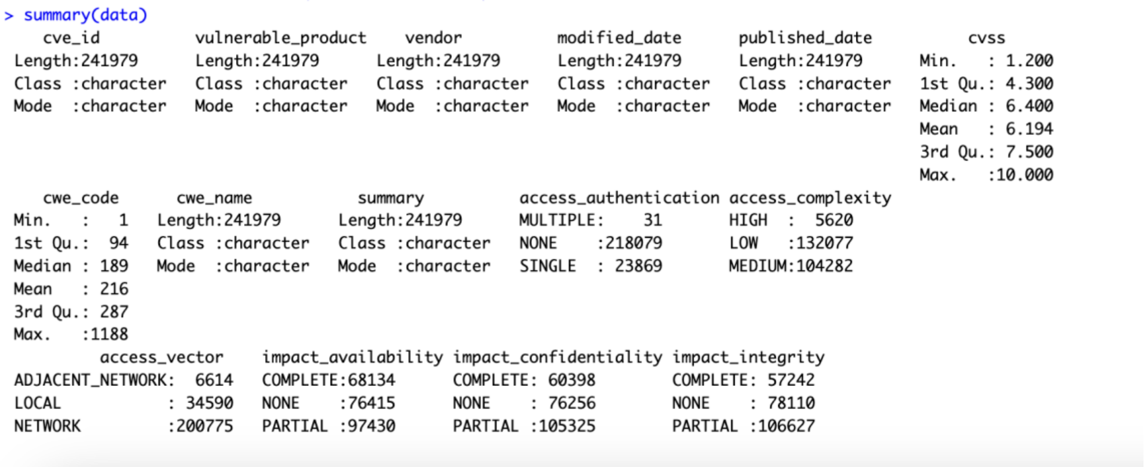
Name: merged_cve_dataset

Applied steps

- Source
- Expanded vendors
- Merged queries
- Expanded cve
- Filtered rows
- Renamed columns

Exploratory Statistical Analysis:

The next step in our process was to conduct a descriptive analysis on our merged dataset. In a more robust analysis we have converted some of the columns ***(access_authentication, access_complexity, access_vector, impact_availability, impact_confidentiality, impact_integrity)*** into categorical variable types as you can identify below in the R studio screenshot.

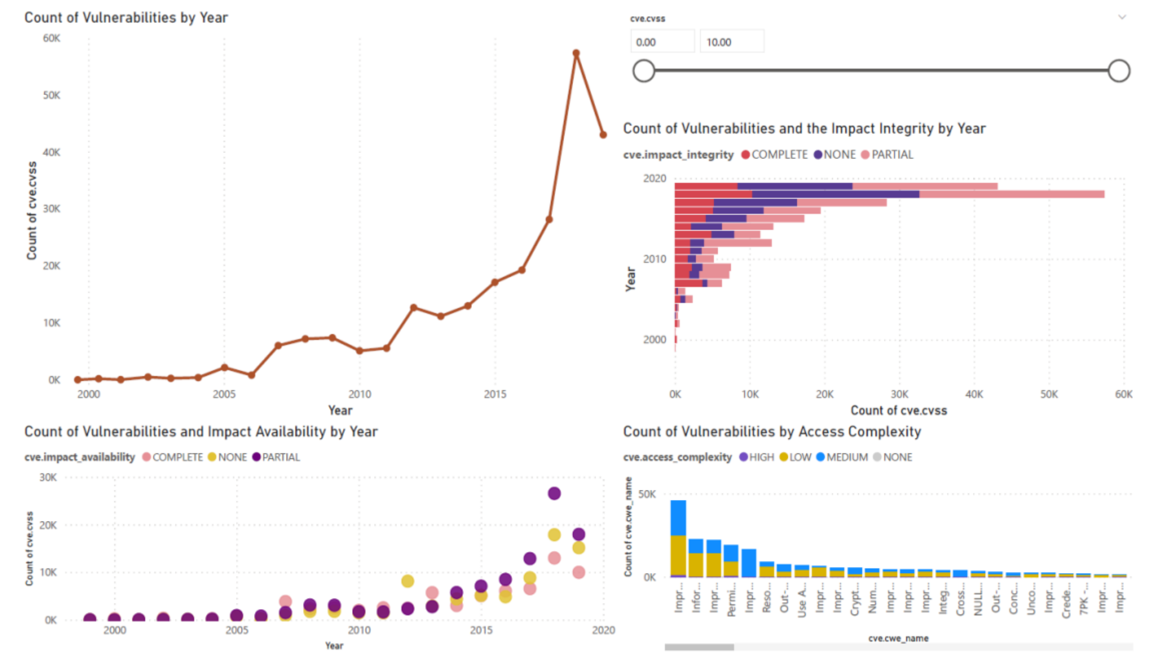


```
> summary(data)
    cve_id          vulnerable_product    vendor          modified_date      published_date          cvss       
 Length:241979      Length:241979      Length:241979      Length:241979      Length:241979      Min.   : 1.200  
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 4.300  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 6.400  
                                                                                                Mean   : 6.194  
                                                                                                3rd Qu.: 7.500  
                                                                                                Max.   :10.000  
    cwe_code      cwe_name            summary          access_authentication access_complexity
 Min.   :   1   Length:241979      Length:241979      MULTIPLE:    31       HIGH  :  5620    
 1st Qu.:  94   Class :character   Class :character   NONE    :218079       LOW   :132077    
 Median : 189   Mode  :character   Mode  :character   SINGLE  : 23869       MEDIUM:104282    
 Mean   : 216                                                                                 
 3rd Qu.: 287                                                                                 
 Max.   :1188                                                                                 
          access_vector    impact_availability impact_confidentiality impact_integrity
 ADJACENT_NETWORK:  6614   COMPLETE:68134      COMPLETE: 60398        COMPLETE: 57242 
 LOCAL           : 34590   NONE    :76415      NONE    : 76256        NONE    : 78110 
 NETWORK         :200775   PARTIAL :97430      PARTIAL :105325        PARTIAL :106627 
```

8. Data Visualization Process:

Our Data Visualization Process involved visualization of our research question to generate trends. We have developed three dashboards, each of which displays the visualizations of our research questions.

- ## Time Series Analysis of High/Severe Vulnerabilities over time:



Inference:

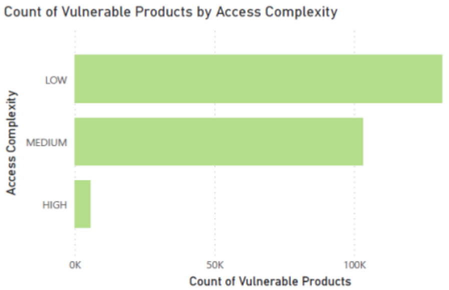
- From the first visualization it can be seen that the maximum number of vulnerabilities were identified in the year **2018** where there is a peak before there was a dip again in 2019.

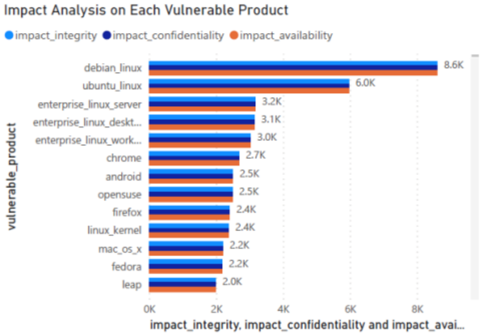
- The second visualization, provides the count of vulnerabilities and their impact integrity on the IT infrastructure over the years.

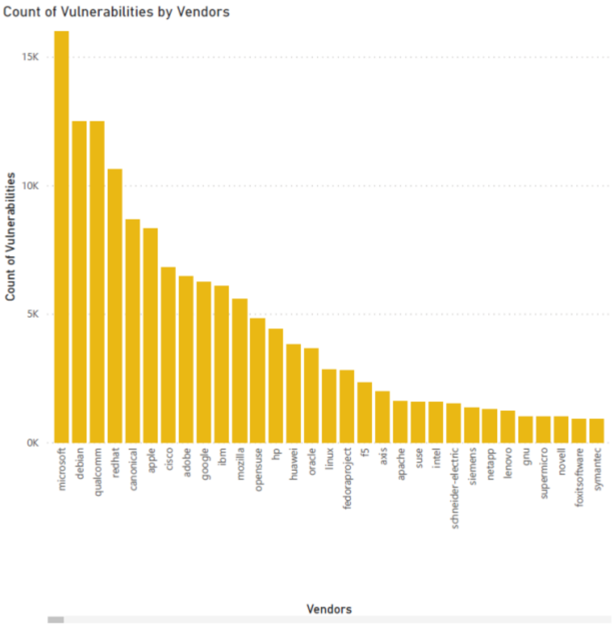
- In the third visualization, we have identified how the vulnerabilities have affected the availability of the IT systems over the years. As we can see from the scatter plot that in the year 2018 there were the highest number of vulnerabilities which have ***"PARTIALLY"*** (Indicated by purple dot) affected the availability of IT systems.

- Coming to the final visualization, it shows which vulnerability had the highest impact along with the count of their access complexities. Here, "***Improper Restriction of Operations within the bounds of memory buffer"*** was the highest common vulnerability as shown by the visualization.

Vulnerabilities by Access Complexity, Impact, CVSS scores:







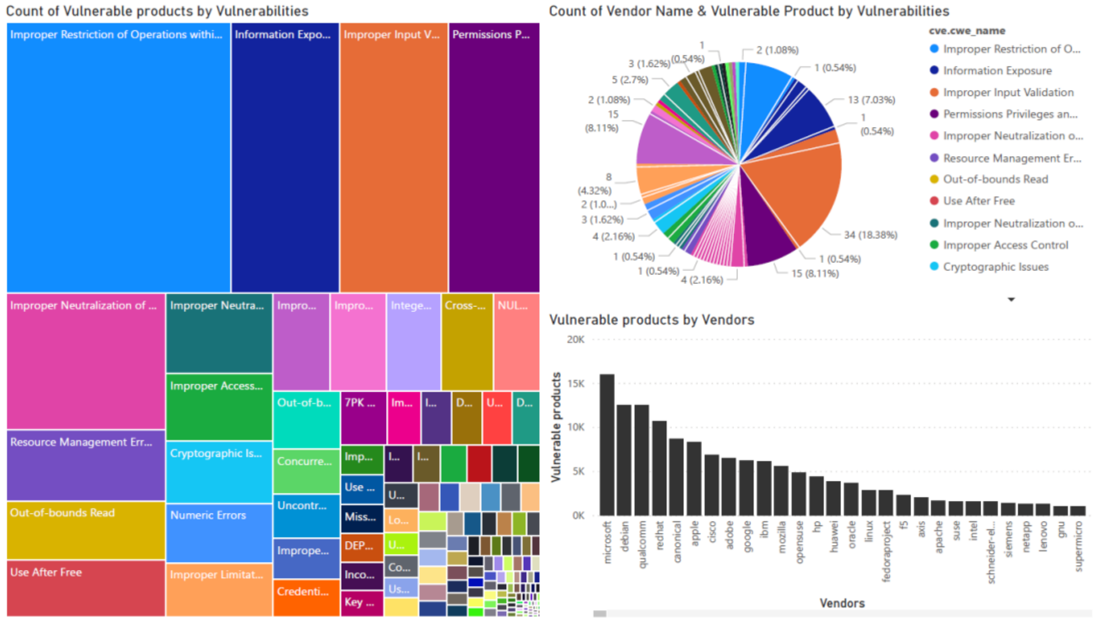
Inferences:

- From the first visualization here, we can see the count of vulnerabilities based on their access complexities. Apparently, there are the highest number of vulnerabilities (around 130K) with a **"low access complexity".**

- The second visualization will help us understand the count of vulnerabilities for each IT vendor and as we can see that Microsoft tops the chart with highest vulnerabilities with more than 15K vulnerabilities while Debian and Qualcomm have joint second highest vulnerabilities with almost around 12.5K vulnerabilities.

- The third visualization provides an "Impact Analysis" ***(impact integrity, impact confidentiality and impact availability)*** for each vulnerable product. Here, as we can see, debian linux had the highest impact among all the vulnerable products.

- ## **Identifying Vendors/Products based on Vulnerability:**



Inferences:

- The first visualization here shows the count of vulnerable products based on the vulnerability names. As we can see, there were the highest number of vulnerable products (around 45K) with the vulnerability "***Improper Restriction of Operations within the bounds of memory buffer".***

- In the second visualization, we have visualized the percentage of vendors based on both vulnerable products and vulnerability identified.

- In the final visualization, we have identified the vendors with the highest number of vulnerable products and we can see that Microsoft again tops this chart with the highest number of vulnerable products.

9. Challenges Faced:

We initially faced a lot of difficulties importing external datasets from our local PCs while attempting to carry out the ETL, cleaning, and descriptive analysis using Microsoft Machine Learning Studio, owing to constraints in the free tier. As a result, we decided to use Power Query Editor and Rstudio for completing our mentioned tasks.

We also encountered a situation where we were hesitant to remove the NULL values from our dataset because it could have an impact on our prediction models and final visualizations, but we later learned that the percentage of NULL values in our dataset is 10% and removing those values may not make a significant difference and thus we made the omission.

10. Conclusion:

Our analysis and visualization would provide a reference for IT Vendors and Customers to identify common IT infrastructure which are prone to the Common Vulnerabilities and also can help understand which vulnerabilities impact the product in which way and develop risk mitigation strategies against these cyber risk threats.

11. Future Scope:

For future-scope of this project, we can use this historical data to make predictive analysis using proper Machine Learning models on these vulnerabilities and uncover various trends which can help the IT Vendors in further improving their cybersecurity posture.

References:

CVE (Common Vulnerabilities and Exposures). (2020b, March 26). Kaggle. Retrieved September 12,2023, from https://www.kaggle.com/datasets/andrewkronser/cve-common-vulnerabilities-and-exposures