Commits Fixing Vulnerabilities

A Comparative Analysis on World of Code (WoC)

Clay Alan Shubert
cshubert@vols.utk.edu
The University of Tennessee-Knoxville
Knoxville, Tennessee, USA

Robert Grady Williams
rwill116@vols.utk.edu
The University of Tennessee-Knoxville
Knoxville, Tennessee, USA

Zach Williams
zwilli13@vols.utk.edu
The University of Tennessee-Knoxville
Knoxville, Tennessee, USA

Figure 1. Imperva graphic illustrating CVE-2023-22524: an RCE Vulnerability in Atlassian Companion for MacOS. 1

Abstract

This paper reviews commit messages containing the keyword "CVE" from over three billion World of Code (WoC) commits. Each commit message is then parsed to create a dataset of only commit hashes fixing or resolving vulnerable code. Each CVE number in the commit is verified with the National Vulnerability Database (NVD) to determine if it is a legitimate CVE. We then perform a comparative analysis of the projects. Our goals of this analysis included: learning more about software supply-chain issues by providing data of the number of repositories that had to be fixed in order to correct a CVE, comparing the size of the project to the response time (time between CVE post on NVD and the commit fixing the CVE), and comparing the severity of the CVE to the number of commits that fix the CVE (ranking). The results of this analysis provide insight into the challenges and breadth of software supply-chain vulnerability management and resolution.

CCS Concepts: • **Security and privacy** → **Software security engineering**; *Domain-specific security and privacy architectures*; • **Software and its engineering** → Software creation and management.

Keywords: software supply-chain, cybersecurity, vulnerabilities

ACM Reference Format:
Clay Alan Shubert, Robert Grady Williams, and Zach Williams. 2024. Commits Fixing Vulnerabilities: A Comparative Analysis on World of Code (WoC). In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym '24).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym '24, June 03–05, 2024, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In today's interconnected world, cybersecurity has become a critical concern for organizations of all sizes. With the increasing number of cyber threats, it is essential to identify and mitigate vulnerabilities in software systems promptly. One of the key frameworks used to manage and address these vulnerabilities is the Common Vulnerabilities and Exposures (CVE) system. The CVE program provides a standardized method for identifying and tracking security vulnerabilities, which is vital for maintaining the integrity and security of software applications.

Each CVE instance is assigned a unique identifier (CVE ID), facilitating the tracking and resolution of issues across various systems. The CVE system provides a standardized identifier for a given vulnerability, making it easier for organizations to communicate and address security issues consistently. Managed by the MITRE Corporation under the oversight of the Cybersecurity and Infrastructure Security Agency (CISA), the CVE program is a cornerstone of cybersecurity.

Fixing CVEs is crucial for several reasons. Firstly, it helps prevent attackers from exploiting known vulnerabilities, thereby reducing the attack surface and overall security risk within a project. Addressing CVEs is also often required to comply with security standards and regulations, avoiding potential legal and financial penalties. Finally, regularly updating systems to address CVEs is a critical component of proactive risk management and security best practices.

In our study, we utilized the World of Code (WoC) database, which contains 3.1 billion commits. We wrote a script to pull all commits in WoC that contained 'CVE' in the commit message, resulting in over 500,000 commits to analyze against known CVEs. This allowed us to examine both organizational and commercial commits using the a2c mapping, providing valuable insights into how vulnerabilities are addressed in different contexts. We then performed a quantitative analysis of the data we collected to make comparisons and draw conclusions about current trends and practices in software security vulnerability management.

2 Methodology

2.1 Project Goals

Our study aims to deepen the understanding of software supply-chain issues by analyzing various aspects of CVE management. We will provide data on the number of repositories that required fixes to correct a CVE, offering a clear picture of the extent of these vulnerabilities. Additionally, we will compare the response times between organization-affiliated and commercial users, specifically looking at the time elapsed between a CVE being posted and the corresponding commit that addresses it. Another goal is to compare the severity of CVEs to the number of commits needed to fix them, ranking these CVEs based on this data. Through this analysis, we aim to gain valuable insights into the efficiency and effectiveness of different entities in addressing security vulnerabilities.

2.2 Technical Approach

The first step of our analysis involved pulling all of the relevant commits from the World of Code version S. We selected this version because we were worried about the overall run time and wanted to use a smaller dataset. The query script used to perform this iterates through all the commits stored in World of Code that contain the text string 'CVE' in the commit message. The script is shown in Appendix A.1. This query took over 15 hours to complete and the final output file was 4 GB.

Following the initial query, we began data pre-processing in Python. This script can be found in Appendix A.2. We can only analyze commits that have a valid CVE number, so we removed all cases where 'CVE' was part of the commit message by coincidence or where the CVE number was not valid. Additionally, we manually pulled the severity scores of the top 25 most committed CVEs due to there not being any easily accessible datasets for all CVE severities. The Common Vulnerability Scoring System (CVSS) is used as a method to measure a vulnerability's impact on a scale from zero to ten for these CVEs. Along with pulling the severity scores of the top 25 most committed CVEs we also categorized them based on the general type of vulnerability it was. Finally, we also formatted the data to make it as simple to import into Microsoft's PowerBI as possible for final analysis. This pre-processing allowed our PowerBI edits and analytics to run smoother and handle less of the overall processing load. We made the decision to use PowerBI due to our team's familiarity with it as well as its ability to easily make complex graphics combining multiple datasets to communicate our findings. We began by analyzing basic information such as the distribution of commits by year, vulnerabilities with the most commits fixing them, and any trends among those top vulnerabilities. We then obtained a dataset with all CVEs from 1999 to date so that we would have access to a description and update date for each one. We then merged that dataset with our commit dataset so we could analyze the time between the latest CVE update and the commit fixing the vulnerability. We used this information to analyze the average time to fix a given vulnerability compared to the number of commits, organization, and CVE year.

3 Results

In the beginning of our analysis, we looked at the general dataset to see the distribution of commits that we were analyzing. Figure 2 shows the number of commits by year and shows that 2016 through 2018 had the highest number of commits with a downward trend in more recent years. This

could partially be attributed to the World of Code version cutting off during 2020.

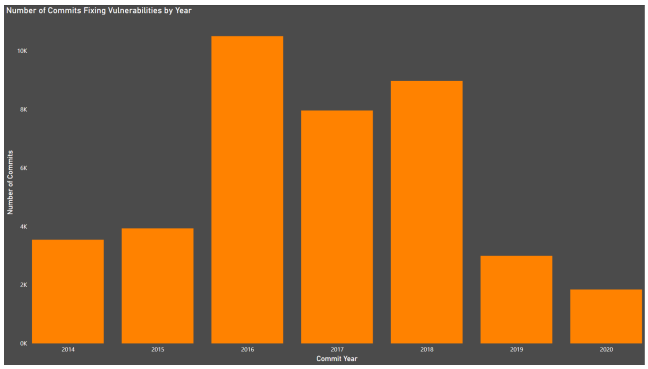

Figure 2. Commits that fixed a vulnerability by commit year.

We then analyzed the number of distinct authors for the top 25 vulnerabilities by number of commits fixing them and noticed that they generally trend together aside from one outlier. Figure 3 shows this downward trend. This specific CVE was a 7.5 severity, which is not a clear reason for it being such an outlier.

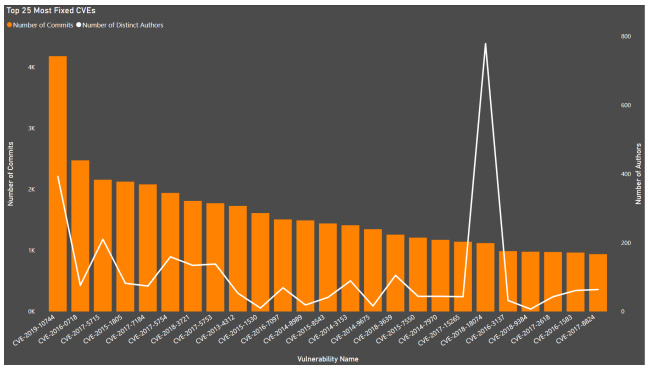

Figure 3. Top 25 most fixed CVEs compared to the number of unique authors fixing the vulnerability.

Next, we looked at the severity scores of the top 25 most fixed CVEs to see if there was a trend in why they were fixed. Figure 4 shows that there is no correlation between the CVE severity and the number of commits.

We then wanted to analyze the types of vulnerabilities that were being fixed most frequently. We compared the number of commits to the severity score and vulnerability category. Figure 5 shows the top 25 most fixed CVEs and shows that there is again no clear trend in the number of commits based on the category of vulnerability.

After merging our datasets, we were able to analyze the average time between the most recent CVE update and the commit fixing the specific vulnerability. Figure 6 shows this

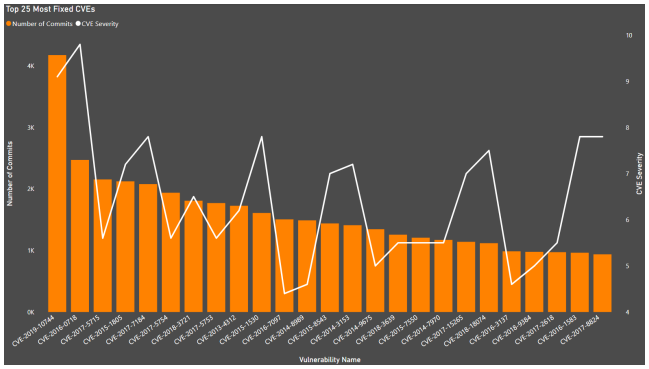

Figure 4. Top 25 most fixed CVEs compared to the severity of the vulnerability.

Name	Vulnerability Category	Number of Commits	Severity
CVE-2019-10744	Prototype Pollution Vulnerability	4176	9.10
CVE-2016-0718	Denial of Service (DoS) Vulnerability	2472	9.80
CVE-2017-5715	Information Disclosure Vulnerability	2155	5.60
CVE-2015-1805	Denial of Service (DoS) Vulnerability	2124	7.20
CVE-2017-7184	Privilege Escalation Vulnerability	2079	7.80
CVE-2017-5754	Information Disclosure Vulnerability	1939	5.60
CVE-2018-3721	Prototype Pollution Vulnerability	1809	6.50
CVE-2017-5753	Information Disclosure Vulnerability	1771	5.60
CVE-2013-4312	Linux Kernel Vulnerability	1727	6.20
CVE-2015-1530	Denial of Service (DoS) Vulnerability	1610	7.80
CVE-2016-7097	Denial of Service (DoS) Vulnerability	1506	4.40
CVE-2014-8989	Linux Kernel Vulnerability	1490	4.60
CVE-2015-8543	Information Disclosure Vulnerability	1440	7.00
CVE-2014-3153	Denial of Service (DoS) Vulnerability	1411	7.20
CVE-2014-9675	Other Vulnerability	1346	5.00
CVE-2018-3639	Information Disclosure Vulnerability	1257	5.50
CVE-2015-7550	Denial of Service (DoS) Vulnerability	1208	5.50
CVE-2014-7970	Denial of Service (DoS) Vulnerability	1172	5.50
CVE-2017-15265	Privilege Escalation Vulnerability	1141	7.00
CVE-2018-18074	Information Disclosure Vulnerability	1119	7.50
CVE-2016-3137	Other Vulnerability	987	4.60
CVE-2018-9384	Privilege Escalation Vulnerability	978	5.00
CVE-2017-2618	Other Vulnerability	973	5.50
CVE-2016-1583	Privilege Escalation Vulnerability	964	7.80
CVE-2017-8824	Privilege Escalation Vulnerability	938	7.80

Figure 5. Type of vulnerability for the top 25 most fixed CVEs

average time by the CVE release year. There is a clear downward trend due to the number of commits fixing any vulnerabilities increasing towards 2018, and very old CVEs have a very long average time to fix due to the World of Code dataset primarily having data from 2014 to 2020.

Figure 7 then compares the average time to fix to the number of commits for the top 25 most fixed vulnerabilities. It shows that there is no clear correlation between the number of commits and the average time it took to make those commits.

Next, figure 8 compares the average time to fix a vulnerability by CVE publish year to the number of commits by the commit year in our dataset. This graph shows that, as the number of commits increases, the average time to fix goes down. This does not seem to hold true for 2020, but this is

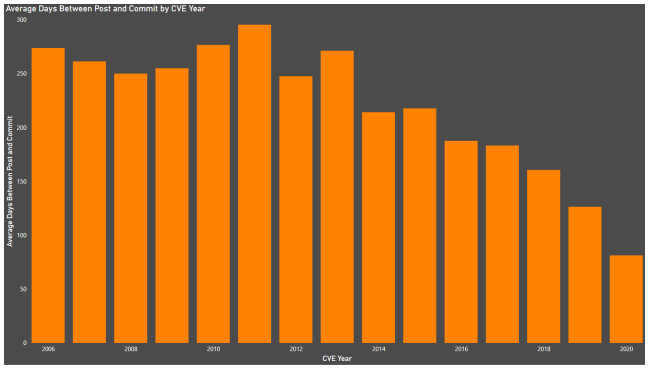

Figure 6. Average time to fix vulnerabilities by CVE year.

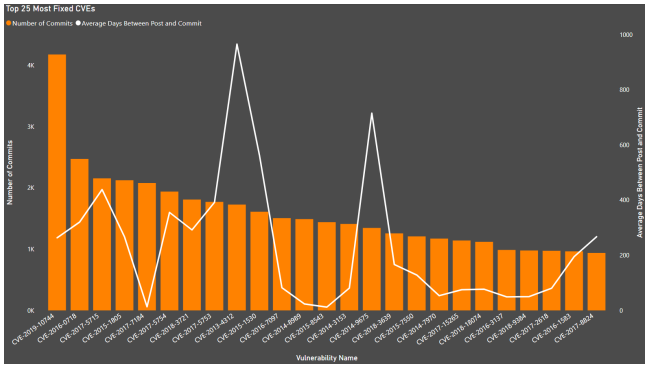

Figure 7. Top 25 CVEs compared to the average time to fix them.

likely also due to the World of Code version ending in this year.

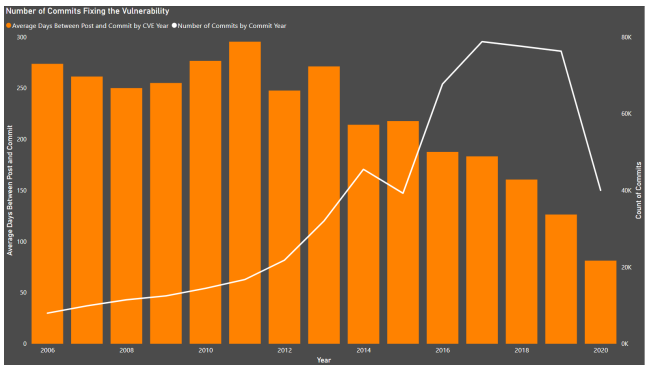

Figure 8. Average time to fix a vulnerability by CVE update year compared to the number of commits by commit year.

Finally, figure 9 compares the author's email domain to the number of commits and average time to fix a vulnerability. This shows that organizational users are more likely to fix a vulnerability quickly as opposed to individuals even though they make up the majority of the commits being made.

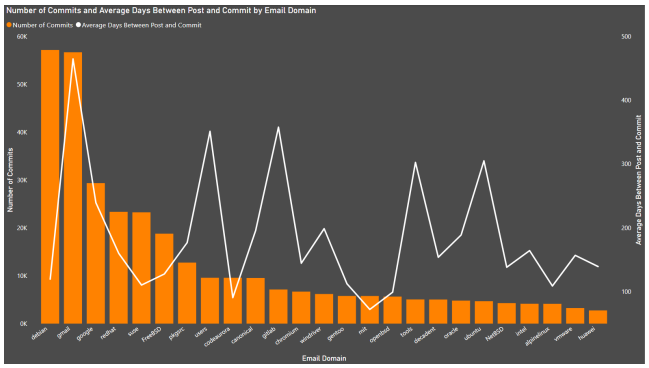

Figure 9. Number of commits and average time to fix by email domain.

4 Discussion

One important thing to note in terms of data collection is that the only dataset that we could find did not include the original release date of the CVE. Because of this, any CVEs that were updated after a majority of the fixes had already been committed had a negative average time to patch and were not included in our analysis. We also would have liked to have a dataset that had the severity scores of every CVE rather than just the top 25. This would have allowed us to see if there was a broader correlation among all CVEs as opposed to just the most fixed ones.

For our actual analysis, in looking at our results we had a couple of surprising results. Firstly the lack of correlation between the severity and number of commits seemed odd. Our initial speculation was a higher severity would trend toward a higher number of commits, however, this proved to not be the case. Further analysis could be done, but we believe a possible explanation could be that the popularity of the code with the vulnerability is more likely a driver for the commits fixing it than the severity.

We also found a couple of interesting results when looking at the number of commits and average days between post and commit based on email domain. We found that debian, codeaurora, and mit were a couple of the domains that were the most responsive, whereas gmail, users, and gitlab were domains that had a relatively high average of days between the post and commit. This seemingly shows that organizational users are more likely to quickly respond to a known vulnerability, which we expected.

A correlation that we did anticipate was the average time to fix a vulnerability with the number of commits. When there were a greater number of commits, the average time to fix a vulnerability went down. This indicated that the commits were reflective of the work done on a particular CVE and this lead to a more rapid fix, rather than merely the passing of time.

Overall there were multiple findings that were in line with

our expectations and others that surprised us. We noted interesting outliers and unanticipated correlations or the lack thereof.

5 Conclusion

Our dataset, which includes over 500,000 commits that fix vulnerabilities, provides a valuable resource for future work and research. To enhance the robustness and relevance of our analysis, the metrics of our dataset can be consistently updated as new World of Code data becomes available, ensuring that it remains current. This will involve recollecting and parsing data using the scripts in Appendix A. Additionally, more detailed data could be pulled from each commit, enabling further comparisons and trend analysis. By expanding the scope of the dataset and refining our metrics, we could offer deeper insights into software supply-chain issues and CVE management.

References

[1] Ron Masas. 2023. Imperva Uncovers CVE-2023-22524, A RCE Vulnerability. https://www.imperva.com/blog/cve-2023-22524-rce-vulnerability-in-atlassian-companion-for-macos/. [Accessed 10-04-2024].

6 Appendices

A Data Collection and Parsing Scripts

A.1 WoC Commit Hash Collection Script

```
$ for i in {0-31}
$ do
$ zcat /da0_data/basemaps/gz/a2cFullS.$i.s |
awk -F ';' '{print $2}' | ~/lookup/showCnt
commit 2 | grep --line-buffered "CVE" >>
cveCommits{i}.txt
$ done
```

A.2 Data Parsing Script

dataparser

May 11, 2024

1 Data Parsing

1.0.1 Remove dummy entries with no CVE number or commits that are not patches, resolutions, or fixes for CVEs

```
[ ]: #Read in csv to a pd dataframe from inside a gzip file and print out the first␣
     ↪10 lines
     import gzip
     import csv
     import sys
     import os
     import numpy as np
     import pandas as pd
```

```
[ ]: with gzip.open("data/CVECommits.csv.gz", 'rt') as f:
         df = pd.read_csv(f, dtype={'timezone': np.character})
         print(df.head(10))
```

```
/Users/clayshubert/Documents/GraduateClasses/COSC525-
DL/.venv/lib/python3.12/site-packages/pandas/core/dtypes/common.py:1645:
DeprecationWarning: Converting `np.character` to a dtype is deprecated. The
current result is `np.dtype(np.str_)` which is not strictly correct. Note that
`np.character` is generally deprecated and 'S1' should be used.
  npdtype = np.dtype(dtype)
                                commit_hash  \
0  72ff8767fbf4bd189192cd8240887324c5e4600b
1  01c809193e4c9f11986460b213b1a9f7c5a7d38a
2  24501a2997d98305e6f7ac2eb82762184dcb2b37
3  9b4f2fd81560c0b37166df4286cdc436cd2a963b
4  a993b06ba8991ea9369ac252c261798e6af8d138
5  b5f2960231864ce7c8298d1a6b8e498e92d09676
6  b9892be17370c654d0f09bd9ed2595952aaaa844
7  cfd1664dc19d09fdf688757e5b9bd0ad9e09911b
8  db9e27de9c0e8747287411461178898fc6eee46
9  213322cbc3ca9863a522f08b0430ea2488cccd26

                       commit_author   timestamp timezone  \
0  ./snoww0lf <codescyber@yahoo.co.id>  1414425713    +0700
```



```
1        0pc0deFR <0pc0deFR@gmail.com>  1406361195    +0200
2        0pc0deFR <0pc0deFR@gmail.com>  1436596973    +0200
3        0pc0deFR <0pc0deFR@gmail.com>  1436597061    +0200
4        0pc0deFR <0pc0deFR@gmail.com>  1406360941    +0200
5        0pc0deFR <0pc0deFR@gmail.com>  1406960535    +0200
6        0pc0deFR <0pc0deFR@gmail.com>  1432668395    +0200
7        0pc0deFR <0pc0deFR@gmail.com>  1428851348    +0200
8        0pc0deFR <0pc0deFR@gmail.com>  1434908930    +0200
9           0x023 <lagetseb@gmail.com>  1523268248    +0200

                                      commit_message
0  Update README.md\n\nShellshock ( Bash CVE-2014…
1                                  Add CVE-2014-5034
2                Add CVE ID in IBS Mappro Exploit
3                 Add CVE ID in Swim Team Exploit
4                                  Add CVE-2014-5034
5                                  Add CVE-2014-5072
6  Add CVE Identifier in WP_Fastest_Cache_0_8_3_4…
7                    Add CVE-ID & EDB-ID Identifiers
8  Add CVE ID for Zip Attachments 1.4 Arbitrary F…
9  CVE-2018-0171\n\nEmbedi PoC for Cisco vuln CVE…
```

```
/Users/clayshubert/Documents/GraduateClasses/COSC525-
DL/.venv/lib/python3.12/site-packages/pandas/core/dtypes/common.py:1645:
DeprecationWarning: Converting `np.character` to a dtype is deprecated. The
current result is `np.dtype(np.str_)` which is not strictly correct. Note that
`np.character` is generally deprecated and 'S1' should be used.
  npdtype = np.dtype(dtype)
```

```
[ ]: #drop rows where the commit_message column does not contain 'CVE-', 'fix',␣
     ↪'resolve', 'patch', fixing', 'resolving', 'patching', or any numbers in it
     df_parsed = df[df['commit_message'].str.␣
     ↪contains('CVE-|fix|resolve|patch|fixing|resolving|patching|[0-9]',␣
     ↪case=False, na=False)]
```

```
[ ]: df_parsed.head(50)
```

```
[ ]:                                commit_hash  \
     0   72ff8767fbf4bd189192cd8240887324c5e4600b
     1   01c809193e4c9f11986460b213b1a9f7c5a7d38a
     4   a993b06ba8991ea9369ac252c261798e6af8d138
     5   b5f2960231864ce7c8298d1a6b8e498e92d09676
     6   b9892be17370c654d0f09bd9ed2595952aaaa844
     7   cfd1664dc19d09fdf688757e5b9bd0ad9e09911b
     8   db9e27de9c0e8747287411461178898fc6eee46
     9   213322cbc3ca9863a522f08b0430ea2488cccd26
     10  347a103a28832ec6d5c4181f9fbd41d9617ef72d
     11  646d100343e1f301ed44bedc5a19d3d1d9bfeed6
```

```
12  bb388c136740530f63784082c093405efd11c649
13  9c38377618130ffe957357e8ed15a45d809c2ce4
14  d4069b761d1151ac19cec9f7a43f509a1f374a65
15  2509d33f9a66f143ebf5f38ec8f35b80cc5b11a8
16  1fc6bff4727412ba32bf43d0109f2e7ecf7e4a55
17  7fa423abd6cad64a7829f9bd54a2e2930f731c9e
18  328fc733561a7f1e67367d407528ed1a9e15ae5f
19  f2253d87832493f44a2831d42828d618d952210c
20  fbaf9890c792a8e5f479289adae68e6ba9cdf7c1
21  331bb982df2c919c7b591433d643f0a9ecf93b30
22  2d700b116c0e05c4ba63b33d8815084d424ad844
23  da4f795d5074e52b088428228aaea2ea8a405cde
24  6a4fbe1bd9da21684eecd77a92f168af8d12d950
25  8950e3fa5a241449c720c8c40ac56f77b9aa0505
26  2c70c535915f621fe51508076804b4575171a23d
27  969f3c39a2d4e1a7fb5ffaf18b8c939921375206
28  03e77b3063e3ae503e9def221cd71bc53c4f499e
29  66321cd75294509f684fe8394fdc18f14a0ba124
30  6f2caf7cea5047782fbefb3baf0f0cc98a247487
31  ab9fd4ce7524da4d91ed6e8f68d334ebbd62ba9b
32  bccd42bf4e9102ef10eff3724881679485dd0430
33  f58e8227040f2d3f95a12a3127096a7bc07d4a32
34  ada56fbbb922bd25406e85e5f34d7bacd3beb2e0
35  3a01877a22a5511f7543eb92f9d84f2d6a02a1c3
36  ceebe0743fc30fc45dd676fe4c7755d83cdfe89d
37  ce549a8dbd0130491aadb5a645f88084510e6f92
39  22f48b9b8248473ffc6d350f672fab7595992a60
42  6854bb24bd49b848f95e3334c9f5f948895d87e1
45  7a842a9f287c61f986ac592a3483b9648969fa8b
46  7db9c92bbfaf94a12e87cc739e6c60bdddd18b98
47  8218cd6c23490615040cbe4b2fee8acbead36587
48  99b7b1c40f3455dde2ee8bdb199b4d9016409572
49  9a914c46862e7326821d57629968bf28ca5db188
51  ec93d18f6fadce02bd399c9b381dbd77c280570d
52  0177bd3f46b8b5f8f32d930553da8fed68fd627d
53  06b8d9f5dd9ba708f4729cee6d052210fa583da7
54  0789540462d717c8119232edcd79e248a679797d
55  0a4f9befd953235c61ecf60a67a1c2c73e1b905c
56  0c93b76a56072233ad8a3480afe74ceaff5f72b5
57  11e968aefa545a13a25d2940d184b94346156914

                                        commit_author   timestamp timezone  \
0                  ./snoww0lf <codescyber@yahoo.co.id>  1414425713    +0700
1                       0pc0deFR <0pc0deFR@gmail.com>  1406361195    +0200
4                       0pc0deFR <0pc0deFR@gmail.com>  1406360941    +0200
5                       0pc0deFR <0pc0deFR@gmail.com>  1406960535    +0200
6                       0pc0deFR <0pc0deFR@gmail.com>  1432668395    +0200
```



```
0   Update README.md\n\nShellshock ( Bash CVE-2014…
1                                    Add CVE-2014-5034
4                                    Add CVE-2014-5034
5                                    Add CVE-2014-5072
6   Add CVE Identifier in WP_Fastest_Cache_0_8_3_4…
7                       Add CVE-ID & EDB-ID Identifiers
8   Add CVE ID for Zip Attachments 1.4 Arbitrary F…
9   CVE-2018-0171\n\nEmbedi PoC for Cisco vuln CVE…
10                               Create CVE-2018-12290
11                               Create CVE-2019-16920
12  Apply patch for CVE-2019-11358.\n\nThis is a p…
13  Workaround dep bug, jsdoc->marked@0.3.7\n\nhtt…
14            Updated yard to address CVE-2017-17042
15                    Patch for CVE-2018-3760
16  request package.lock version, resolve CVE-2018…
17  Update activerecord due to CVE-2012-2661 (bund…
18  Merge commit '681f798f4138e0ebc91896ea8d036752…
19  Merge commit '681f798f4138e0ebc91896ea8d036752…
20  Merge commit '681f798f4138e0ebc91896ea8d036752…
21  Merge pull request #10626 from WhiteCat22/CVE-…
22      Update Pipfile.lock to resolve CVE-2019-14806
23      Update Pipfile.lock to resolve CVE-2019-14806
24  Reference exact CVE in HISTORY for 5.6.6 (#165…
25          Reference exact CVE in HISTORY for 5.6.6
26  Don't allow the stack to grow into hugetlb res…
27  i40e: Changed maximum supported FW API version…
28  update fedora latest for CVE-2017-5461\n\nSign…
29  update rawhide, 23, and latest for openssl CVE…
30  update fedora 23 and rawhide for glibc: CVE-20…
31   update fedora 22 and 23 for glibc: CVE-2015-7547
32       update 22 - 20160218 - glibc: CVE-2015-7547
33  Updated version of paramiko for security patch…
34  Bumping version of requests because of CVE-201…
35  CLI: implent update downloading\n\nThis turned…
36  Backport fixes for openQA input issues after C…
37                                   CVE-2014-1695 PoC
39  ginkgo: Update ADSP/CDSP/SCVE blobs from laure…
42  ginkgo: Update CVP and add missed SCVE blobs f…
45  ginkgo: Update ADSP/CDSP/SCVE blobs from laure…
46  ginkgo: Update CVP and add missed SCVE blobs f…
47  ginkgo: Update ADSP/CDSP/SCVE blobs from laure…
48  ginkgo: Update ADSP/CDSP/SCVE blobs from laure…
49  ginkgo: Update CVP and add missed SCVE blobs f…
51  ginkgo: Update CVP and add missed SCVE blobs f…
52  Update to webmin-1.170nb2 to address: \t http:…
53  Full and proper fix for CVE-2007-5135 PKGREVIS…
54  Fix for CVE-2008-3530 from matt@ Implement imp…
```



```
7                    0pc0deFR <0pc0deFR@gmail.com>  1428851348    +0200
8                    0pc0deFR <0pc0deFR@gmail.com>  1434908930    +0200
9                      0x023 <lagetseb@gmail.com>  1523268248    +0200
10  7hang <48209216+SmithEcon@users.noreply.github…  1593501160    +0800
11  7hang <48209216+SmithEcon@users.noreply.github…  1593497595    +0800
12    Aaron Gray <aarongray@users.noreply.github.com>  1568401551    -0500
13      Aaron J. Lang <aaronjameslang@googlemail.com>  1515625314    +1300
14                 Aaron M. Bond <ambond@gmail.com>  1519769725    -0600
15             Aaron Michal <amichal@greenriver.org>  1529446153    -0400
16                      Aaron Rose <github@acdr.co>  1552004239    +1300
17                Aaron Stone <aaron@brightroll.com>  1338495882    -0700
18            Abhijith Desai <desaia@codeaurora.org>  1554391923    +0530
19            Abhijith Desai <desaia@codeaurora.org>  1554391923    +0530
20            Abhijith Desai <desaia@codeaurora.org>  1554391923    +0530
21  Adam Anderson <31078699+WhiteCat22@users.norep…  1580336341    -0600
22          Adam Englander <adamenglander@yahoo.com>  1567986655    -0700
23          Adam Englander <adamenglander@yahoo.com>  1567986942    -0700
24                      Adam Johnson <me@adamj.eu>  1580920151    +0000
25                      Adam Johnson <me@adamj.eu>  1577961272    +0000
26                     Adam Litke <agl@us.ibm.com>  1192813510    +0200
27      Adam Ludkiewicz <adam.ludkiewicz@intel.com>  1549494502    -0800
28       Adam Miller <maxamillion@fedoraproject.org>  1492739707    -0500
29       Adam Miller <maxamillion@fedoraproject.org>  1457115941    -0600
30       Adam Miller <maxamillion@fedoraproject.org>  1455735103    -0600
31       Adam Miller <maxamillion@fedoraproject.org>  1455807217    -0600
32       Adam Miller <maxamillion@fedoraproject.org>  1455807060    -0600
33                    Adam Shaw <shawa@us.ibm.com>  1541717034    -0600
34                Adam Uhlir <hello@adam-uhlir.me>  1543108210    -0800
35            Adam Williamson <awilliam@redhat.com>  1442861059    -0700
36            Adam Williamson <awilliam@redhat.com>  1502914196    -0700
37    Adam Ziaja <adamziaja@users.noreply.github.com>  1404309116    +0200
39       Adithya R <gh0strider.2k18.reborn@gmail.com>  1587302347    +0530
42       Adithya R <gh0strider.2k18.reborn@gmail.com>  1587313711    +0530
45       Adithya R <gh0strider.2k18.reborn@gmail.com>  1587301666    +0530
46       Adithya R <gh0strider.2k18.reborn@gmail.com>  1587313755    +0530
47       Adithya R <gh0strider.2k18.reborn@gmail.com>  1587302347    +0530
48       Adithya R <gh0strider.2k18.reborn@gmail.com>  1587302347    +0530
49       Adithya R <gh0strider.2k18.reborn@gmail.com>  1587313755    +0530
51       Adithya R <gh0strider.2k18.reborn@gmail.com>  1587313755    +0530
52             Adrian Portelli <adrianp@NetBSD.org>  1133626414    +0000
53             Adrian Portelli <adrianp@NetBSD.org>  1192989173    +0000
54             Adrian Portelli <adrianp@NetBSD.org>  1223022186    +0000
55             Adrian Portelli <adrianp@NetBSD.org>  1157200703    +0000
56             Adrian Portelli <adrianp@NetBSD.org>  1181856462    +0000
57             Adrian Portelli <adrianp@NetBSD.org>  1137938762    +0000

                                        commit_message
```



```
55      Fix for CVE-2006-3125 via Debian. Bump to nb8
56                                  Fix for CVE-2007-2691
57  Update to HylaFAX 4.2.5 From the CHANGES: > Ch…
```

2 Parse CVE IDs

```
[ ]: #Extract the CVE number from the commit_message column and add it to a new␣
     ↪column called 'CVE'
     df_parsed['CVE'] = df_parsed['commit_message'].str.
     ↪extract(r'(CVE-\d{4}-\d{4,7})')
     #drop rows where the CVE column is NaN or null
     df_parsed = df_parsed.dropna(subset=['CVE'])
```

```
/var/folders/6_/q2017nsd6q93lk54vxqvb3x00000gn/T/ipykernel_94153/396824212.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_parsed['CVE'] =
df_parsed['commit_message'].str.extract(r'(CVE-\d{4}-\d{4,7})')
```

```
[ ]: df_parsed.head(50)
```

```
[ ]:                                    commit_hash  \
0   72ff8767fbf4bd189192cd8240887324c5e4600b
1   01c809193e4c9f11986460b213b1a9f7c5a7d38a
4   a993b06ba8991ea9369ac252c261798e6af8d138
5   b5f2960231864ce7c8298d1a6b8e498e92d09676
9   213322cbc3ca9863a522f08b0430ea2488cccd26
10  347a103a28832ec6d5c4181f9fbd41d9617ef72d
11  646d100343e1f301ed44bedc5a19d3d1d9bfeed6
12  bb388c136740530f63784082c093405efd11c649
13  9c38377618130ffe957357e8ed15a45d809c2ce4
14  d4069b761d1151ac19cec9f7a43f509a1f374a65
15  2509d33f9a66f143ebf5f38ec8f35b80cc5b11a8
16  1fc6bff4727412ba32bf43d0109f2e7ecf7e4a55
17  7fa423abd6cad64a7829f9bd54a2e2930f731c9e
18  328fc733561a7f1e67367d407528ed1a9e15ae5f
19  f2253d87832493f44a2831d42828d618d952210c
20  fbaf9890c792a8e5f479289adae68e6ba9cdf7c1
21  331bb982df2c919c7b591433d643f0a9ecf93b30
22  2d700b116c0e05c4ba63b33d8815084d424ad844
23  da4f795d5074e52b088428228aaea2ea8a405cde
26  2c70c535915f621fe51508076804b4575171a23d
27  969f3c39a2d4e1a7fb5ffaf18b8c939921375206
```

```
28  03e77b3063e3ae503e9def221cd71bc53c4f499e
30  6f2caf7cea5047782fbefb3baf0f0cc98a247487
31  ab9fd4ce7524da4d91ed6e8f68d334ebbd62ba9b
32  bccd42bf4e9102ef10eff372488167948fdd0430
33  f58e8227040f2d3f95a12a3127096a7bc07d4a32
34  ada56fbbb922bd25406e85e5f34d7bacd3beb2e0
36  ceebe0743fc30fc45dd676fe4c7755d83cdfe89d
37  ce549a8dbd0130491aadb5a645f88084510e6f92
52  0177bd3f46b8b5f8f32d930553da8fed68fd627d
53  06b8d9f5dd9ba708f4729cee6d052210fa583da7
54  0789540462d717c8119232edcd79e248a679797d
55  0a4f9befd953235c61ecf60a67a1c2c73e1b905c
56  0c93b76a56072233ad8a3480afe74ceaff5f72b5
57  11e968aefa545a13a25d2940d184b94346156914
58  11f438f2db4a14f189833c87ff8c3af6fb9b8dc7
59  1211aafa145542f4cd3228a60d35a2d46bf343d1
60  135e64acbac05b2695780b7f267a35bc17990236
61  2356f1484cec5ab3493b52af887f903d92683217
62  2d4db5f3d636599aadc1f4ffaf9ff02ecc64aaf9
63  33f086d5607c8396f8fadb48dc611c32e66bd5fc
64  369ab45e424c55fe9e6f8bccadb7053bd873f7d4
65  386142659762142ac0dccec42bc0e31ef5666631
66  39d2cb4459ec87e08ea1e9dd45bbd94a2b1fb56e
67  47745d0fbdf3381a660a631b737a7ac9d3cbc3e0
68  4acecf1367c1d221735869bd8f702ed2392a3835
69  4d0d3cdce98d01c28dfb0ca6989fc0764c44a17a
70  54428e6215ca9156e02c13ff91867de11c912af6
71  68573fbd3c2f949f9470f205689303614 27a1ed9
72  68618b0859771c2c753b2664c1090e2083c59027
```

```
                                    commit_author   timestamp timezone  \
0                ./snoww0lf <codescyber@yahoo.co.id>  1414425713    +0700
1                     0pc0deFR <0pc0deFR@gmail.com>  1406361195    +0200
4                     0pc0deFR <0pc0deFR@gmail.com>  1406360941    +0200
5                     0pc0deFR <0pc0deFR@gmail.com>  1406960535    +0200
9                        0x023 <lagetseb@gmail.com>  1523268248    +0200
10  7hang <48209216+SmithEcon@users.noreply.github…  1593501160    +0800
11  7hang <48209216+SmithEcon@users.noreply.github…  1593497595    +0800
12    Aaron Gray <aarongray@users.noreply.github.com>  1568401551    -0500
13      Aaron J. Lang <aaronjameslang@googlemail.com>  1515625314    +1300
14                 Aaron M. Bond <ambond@gmail.com>  1519769725    -0600
15             Aaron Michal <amichal@greenriver.org>  1529446153    -0400
16                    Aaron Rose <github@acdr.co>  1552004239    +1300
17               Aaron Stone <aaron@brightroll.com>  1338495882    -0700
18            Abhijith Desai <desaia@codeaurora.org>  1554391923    +0530
19            Abhijith Desai <desaia@codeaurora.org>  1554391923    +0530
20            Abhijith Desai <desaia@codeaurora.org>  1554391923    +0530
```



```
16  request package.lock version, resolve CVE-2018…  CVE-2018-1000620
17  Update activerecord due to CVE-2012-2661 (bund…     CVE-2012-2661
18  Merge commit '681f798f4138e0ebc91896ea8d036752…     CVE-2019-7221
19  Merge commit '681f798f4138e0ebc91896ea8d036752…     CVE-2019-7221
20  Merge commit '681f798f4138e0ebc91896ea8d036752…     CVE-2019-7221
21  Merge pull request #10626 from WhiteCat22/CVE-…    CVE-2019-12406
22      Update Pipfile.lock to resolve CVE-2019-14806     CVE-2019-14806
23      Update Pipfile.lock to resolve CVE-2019-14806     CVE-2019-14806
26  Don't allow the stack to grow into hugetlb res…     CVE-2007-3739
27  i40e: Changed maximum supported FW API version…     CVE-2019-0140
28  update fedora latest for CVE-2017-5461\n\nSign…     CVE-2017-5461
30  update fedora 23 and rawhide for glibc: CVE-20…     CVE-2015-7547
31    update fedora 22 and 23 for glibc: CVE-2015-7547      CVE-2015-7547
32          update 22 - 20160218 - glibc: CVE-2015-7547     CVE-2015-7547
33  Updated version of paramiko for security patch…  CVE-2018-1000805
34  Bumping version of requests because of CVE-201…     CVE-2018-18074
36  Backport fixes for openQA input issues after C…     CVE-2017-8379
37                                    CVE-2014-1695 PoC     CVE-2014-1695
52  Update to webmin-1.170nb2 to address: \t http:…     CVE-2005-3912
53  Full and proper fix for CVE-2007-5135 PKGREVIS…     CVE-2007-5135
54  Fix for CVE-2008-3530 from matt@ Implement imp…     CVE-2008-3530
55      Fix for CVE-2006-3125 via Debian. Bump to nb8      CVE-2006-3125
56                                Fix for CVE-2007-2691      CVE-2007-2691
57  Update to HylaFAX 4.2.5 From the CHANGES: > Ch…     CVE-2005-3539
58  Fix two vulnerabilities in OpenSSH: - X11 forw…     CVE-2008-1483
59  Update 5.2.2 * Fixed CVE-2007-1001, GD wbmp us…     CVE-2007-1001
60              Fix for CVE-2007-4995 from OpenSSL CVS     CVE-2007-4995
61  Fix for CVE-2006-1990 Fix for CVE-2006-3011 In…     CVE-2006-1990
62                      Fix for CVE-2006-0300 via RedHat     CVE-2006-0300
63  Update to 2.1.3 This should fix PR#39952 MAKE_…     CVE-2008-4474
64          Add a fix for CVE-2006-4003 PKGREVISION++     CVE-2006-4003
65  Security Enhancements and Fixes in PHP 5.2.6:\…     CVE-2008-0599
66  Update to 1.2.0-beta2\n\nNot much info other t…     CVE-2007-5226
67  Fixes (via RedHat) for:\tCVE-2006-4334, CVE-20…     CVE-2006-4334
68  Fix for CVE-2006-5465 from PHP CVS http://www…     CVE-2006-5465
69  Add a patch for http://www.cve.mitre.org/cgi-b…     CVE-2007-6015
70              Add a patch to address CVE-2006-4339     CVE-2006-4339
71  Add a patch via Debain to address: \thttp://cv…     CVE-2005-3340
72  Update to 5.2.7.\n\nSecurity Enhancements and …     CVE-2008-2371
```

```
[ ]: #Convert 'timestamp' column to datetime format in format 'YYYY-MM-DD HH:MM:SS'
df_parsed['event_time'] = pd.to_datetime(df_parsed['timestamp'], unit='s')
#convert event time


df_parsed.head(10)
```



```
21  Adam Anderson <31078699+WhiteCat22@users.norep…  1580336341    -0600
22          Adam Englander <adamenglander@yahoo.com>  1567986655    -0700
23          Adam Englander <adamenglander@yahoo.com>  1567986942    -0700
26                       Adam Litke <agl@us.ibm.com>  1192813510    +0200
27      Adam Ludkiewicz <adam.ludkiewicz@intel.com>  1549494502    -0800
28      Adam Miller <maxamillion@fedoraproject.org>  1492739707    -0500
30      Adam Miller <maxamillion@fedoraproject.org>  1455735103    -0600
31      Adam Miller <maxamillion@fedoraproject.org>  1455807217    -0600
32      Adam Miller <maxamillion@fedoraproject.org>  1455807060    -0600
33                  Adam Shaw <shawa@us.ibm.com>  1541717034    -0600
34               Adam Uhlir <hello@adam-uhlir.me>  1543108210    -0800
36          Adam Williamson <awilliam@redhat.com>  1502914196    -0700
37  Adam Ziaja <adamziaja@users.noreply.github.com>  1404309116    +0200
52             Adrian Portelli <adrianp@NetBSD.org>  1133626414    +0000
53             Adrian Portelli <adrianp@NetBSD.org>  1192989173    +0000
54             Adrian Portelli <adrianp@NetBSD.org>  1223022186    +0000
55             Adrian Portelli <adrianp@NetBSD.org>  1157200703    +0000
56             Adrian Portelli <adrianp@NetBSD.org>  1181856462    +0000
57             Adrian Portelli <adrianp@NetBSD.org>  1137938762    +0000
58             Adrian Portelli <adrianp@NetBSD.org>  1207228154    +0000
59             Adrian Portelli <adrianp@NetBSD.org>  1178482057    +0000
60             Adrian Portelli <adrianp@NetBSD.org>  1192998854    +0000
61             Adrian Portelli <adrianp@NetBSD.org>  1153257679    +0000
62             Adrian Portelli <adrianp@NetBSD.org>  1157304256    +0000
63             Adrian Portelli <adrianp@NetBSD.org>  1229726958    +0000
64             Adrian Portelli <adrianp@NetBSD.org>  1193578847    +0000
65             Adrian Portelli <adrianp@NetBSD.org>  1209919844    +0000
66             Adrian Portelli <adrianp@NetBSD.org>  1192304483    +0000
67             Adrian Portelli <adrianp@NetBSD.org>  1159019537    +0000
68             Adrian Portelli <adrianp@NetBSD.org>  1162639675    +0000
69             Adrian Portelli <adrianp@NetBSD.org>  1197323088    +0000
70             Adrian Portelli <adrianp@NetBSD.org>  1157622271    +0000
71             Adrian Portelli <adrianp@NetBSD.org>  1137538137    +0000
72             Adrian Portelli <adrianp@NetBSD.org>  1228482457    +0000
```

```
                                    commit_message             CVE
0   Update README.md\n\nShellshock ( Bash CVE-2014…    CVE-2014-6271
1                                 Add CVE-2014-5034     CVE-2014-5034
4                                 Add CVE-2014-5034     CVE-2014-5034
5                                 Add CVE-2014-5072     CVE-2014-5072
9   CVE-2018-0171\n\nEmbedi PoC for Cisco vuln CVE…    CVE-2018-0171
10                             Create CVE-2018-12290    CVE-2018-12290
11                             Create CVE-2019-16920    CVE-2019-16920
12  Apply patch for CVE-2019-11358.\n\nThis is a p…    CVE-2019-11358
13  Workaround dep bug, jsdoc->marked@0.3.7\n\nhht…    CVE-2017-17461
14             Updated yard to address CVE-2017-17042    CVE-2017-17042
15                          Patch for CVE-2018-3760     CVE-2018-3760
```



```
[ ]:                                    commit_hash  \
0   72ff8767fbf4bd189192cd8240887324c5e4600b
1   01c809193e4c9f11986460b213b1a9f7c5a7d38a
4   a993b06ba8991ea9369ac252c261798e6af8d138
5   b5f2960231864ce7c8298d1a6b8e498e92d09676
9   213322cbc3ca9863a522f08b0430ea2488cccd26
10  347a103a28832ec6d5c4181f9fbd41d9617ef72d
11  646d100343e1f301ed44bedc5a19d3d1d9bfeed6
12  bb388c136740530f63784082c093405efd11c649
13  9c38377618130ffe957357e8ed15a45d809c2ce4
14  d4069b761d1151ac19cec9f7a43f509a1f374a65

                                    commit_author   timestamp timezone  \
0                ./snoww0lf <codescyber@yahoo.co.id>  1414425713    +0700
1                     0pc0deFR <0pc0deFR@gmail.com>  1406361195    +0200
4                     0pc0deFR <0pc0deFR@gmail.com>  1406360941    +0200
5                     0pc0deFR <0pc0deFR@gmail.com>  1406960535    +0200
9                        0x023 <lagetseb@gmail.com>  1523268248    +0200
10  7hang <48209216+SmithEcon@users.noreply.github…  1593501160    +0800
11  7hang <48209216+SmithEcon@users.noreply.github…  1593497595    +0800
12    Aaron Gray <aarongray@users.noreply.github.com>  1568401551    -0500
13      Aaron J. Lang <aaronjameslang@googlemail.com>  1515625314    +1300
14                 Aaron M. Bond <ambond@gmail.com>  1519769725    -0600

                                    commit_message             CVE  \
0   Update README.md\n\nShellshock ( Bash CVE-2014…    CVE-2014-6271
1                                 Add CVE-2014-5034     CVE-2014-5034
4                                 Add CVE-2014-5034     CVE-2014-5034
5                                 Add CVE-2014-5072     CVE-2014-5072
9   CVE-2018-0171\n\nEmbedi PoC for Cisco vuln CVE…    CVE-2018-0171
10                             Create CVE-2018-12290    CVE-2018-12290
11                             Create CVE-2019-16920    CVE-2019-16920
12  Apply patch for CVE-2019-11358.\n\nThis is a p…    CVE-2019-11358
13  Workaround dep bug, jsdoc->marked@0.3.7\n\nhht…    CVE-2017-17461
14             Updated yard to address CVE-2017-17042    CVE-2017-17042

          event_time
0  2014-10-27 16:01:53
1  2014-07-26 07:53:15
4  2014-07-26 07:49:01
5  2014-08-02 06:22:15
9  2018-04-09 10:04:08
10 2020-06-30 07:12:40
11 2020-06-30 06:13:15
12 2019-09-13 19:05:51
13 2018-01-10 23:01:54
14 2018-02-27 22:15:25
```

```
[ ]: # save the dataframe to a new csv file
     df_parsed.to_csv('data/CVECommitsParsed.csv', index=False)
```


B Project Links

Here are the links to several artifacts from our project.
GitHub Repository
PowerBI File
Presentation Slides

C Team Contributions

Clay Shubert. Lead in the creation of the WoC data collection script. Pre-processed data using dataparser.ipynb script to filter out invalid commit messages for our analysis.

Robert Grady Williams. Lead PowerBI analyst. Imported the data after pre-processing from Python into PowerBI and did transformations and merges with a dataset of CVE publish dates to create visuals showing project results.

Zach Williams. Assisted in RegEx pull query and Batch scripts. Lead PowerBI Aesthetics Coordinator. Lead in presentation and report creation.

Received 10 April 2024; revised 3 May 2024; accepted 14 May 2024