1 Data Source and Collection

*The vulnerable dataset vs non-vulnerable dataset used in Section 3 is 10000:10000.

1.1 Vulnerable

The vulnerable dataset has two main resources:

- From previous research provided. (*The results in Section 3 are from this dataset)
- Newly fetched NVD data with 400+ more official data than the previous one. (*The CVE SEVERITY correlation analysis results in huge difference with new dataset than origin one.)

1.2 Non-vulnerable

The non-vulnerable dataset comes from latest Pytorch/Tensorflow GitHub repository. The latest version is considered to be non-vulnerable at this moment (*Need to be confirmed)

2 Metric Process

Transfer the raw dataset/source to a metric-processed one with details.

2.1 Metric Definition

Metrics	Description	Details				
Basic						
URL	The URL link source	GitHub commit/pull request				
Repo Name	The Repository	GitHub repository name				
	name					
Date	The time of the	The time that a vulnerability was solved				
	URL commit					
CVE ID	The CVE id	Official CVE ID from NVD				
CVE Severity	The CVE severity	The latest CVE severity metric's base score				
Name	Component name	The component is the base unit				
Component Type	Component type	File or Group. Group definition is based on the				
		locality within the same CVE issues/commit				
Code ownership						
Ownership	The ownership	The highest ownership of a component				
Num of Contributor	The sum of	The sum of the contributors to a component				
	contributors					
Num of Minor T%	The amount of the	The total amount of the minor contributors to a				
	minor contributors	component. Contributor with ownership under				
		T% is Minor contributor.				
		(* T%: 5%, 10%, 20%, 50%)				
Per of Minor T%	The proportion of	The proportion of the minor contributors over all				
	the minor	the contributor amount				
	contributors					

^{*}In the provided one, the origin research skipped/ignored some commits from security advisory or blob markdown.

Avg of Minor Contri T%	The average of	The average value of the minor contributors'
	minor contributor's	ownership
	ownership	
Time/Release (See 2.2 fo	r details)	
Days Difference	The project existing	The existing time of the project in GitHub
	time	repository till the Date
Age	The component	The component lifetime calculated based on Git
	lifetime	Log info
Time Stage Numeric	Five Time stages	The five Time stages' numerical value. Calculated
		by Days Difference
Time Stage Aged	Five Time stages	The five Time stages' numerical value. Calculated
Numeric	Aged	by Age
Oss Stage Numeric	Six Oss stages	The six Oss stages' numerical value. Calculated
		by Days Difference
Oss Stage Aged Numeric	Six Oss stages Aged	The six Oss stages' numerical value. Calculated
		by Age
Is Pre-release	Vulnerability found	The release tag where the vulnerability found is
	at pre-release	pre-release
Is Post-release	Vulnerability found	The release tag where the vulnerability found is
	at post-release	post-release
Release Amounts	Affected releases	The affected releases until the project latest time
Release Amounts Aged	Affected releases	The affected releases within the lifetime of the
	Aged	vulnerability existing
Classic metrics		
Code churn	NLOC	The number of lines changed = total added +
		total deleted
File Size	File Size	The number of lines for a component
Churn rate	Churn rate	= Code churn / File Size

2.2 Time Stage + Oss Stage Metric

Metric	Details	Numeric
		Value
Time Stage		
T1	The given time period is in 0 to 7 days	1
T2	The given time period is in 7 days to 3 months	2
T3	The given time period is in 3 months to 9 months	3
T4	The given time period is in 2 years to 3 years	4
T5	The given time period is beyond 3 years	5
Oss Stage		
SI	Success Initialisation.	1
	Has at least one successful release	
TI	Tragedy Initialisation.	2
	Within the given time period (>1year), no release	
SG	Success Growth.	5
	>= 3 releases AND >= 6 months between releases	
TG	Tragedy Growth.	6
	1 or 2 releases and >=1 year since the last release at the time of data	
	collection	
II	Indeterminate Initialisation.	3

	0 releases and < 1 year since project registration	
IG	Indeterminate Growth.	4
	1 or 2 releases and < 1 year since the last release OR 3 releases and <	
	6 months between releases	

2.3 Vulnerable process + App interface

An application interface (vulnerable process) is created for processing the raw dataset to distilled one under the metric defined. This application interface defaults to calculate the provided dataset from Section 1. And it enables user to calculate the metric info with given commit/pull request URLs (CVE ID optional).

2.4 Non-vulnerable process + App interface

An application interface (non-vulnerable process) is created for calculate the metric info of a Git repository. The application interface defaults to process the local Git repo specified in settings, while also allows user to examine the repo with an external GitHub repo URL.

3 Result Analysis

3.1 Exploring Nature of Data

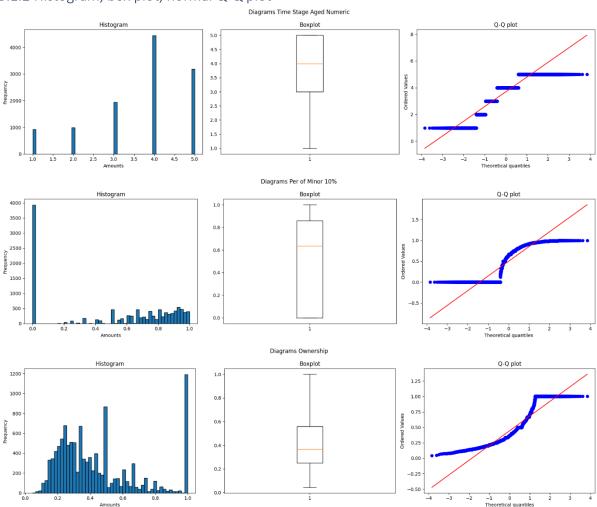
In summary, the dataset is not normally distributed.

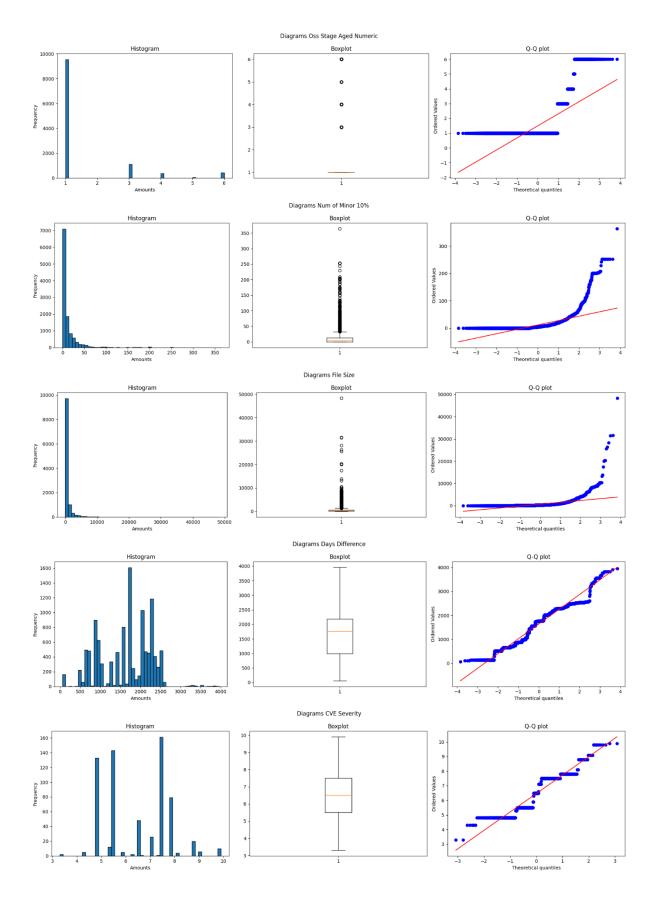
3.1.1 Descriptive statistics

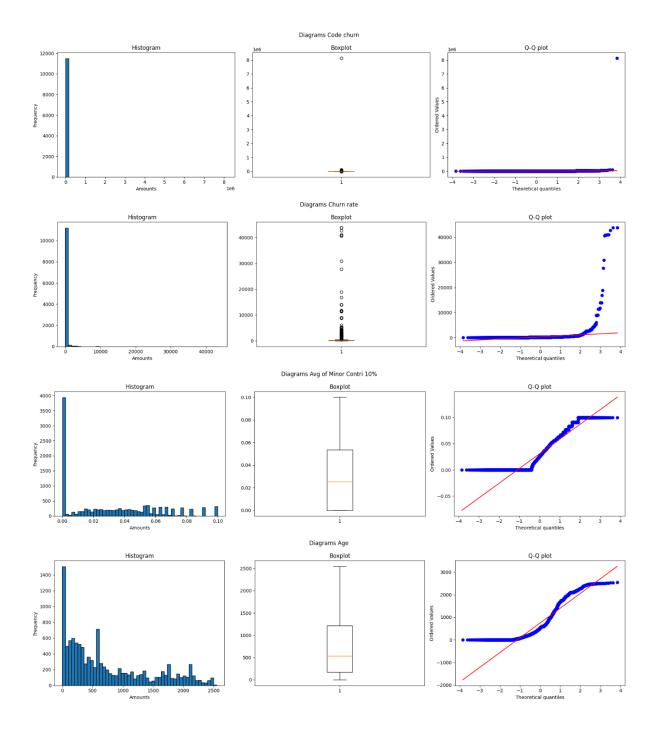
Metrics	N	Min	Max	Mean	Std. Dev.	
				Statistics	Std. Error	
CVE Severity	658	3.3	9.9	6.465502	0.05183	1.329519
Ownership	11491	0.044088	1.00000	0.443906	0.0023718	0.254253
Num of	11491	0.0	364	11.495779	0.212516	22.780893
Minor 10%						

Per of Minor	11491	0.000000	1.000000	0.499368	0.00364105	0.390307
10%					5	
Avg of Minor	11491	0.000000	0.100000	0.030714	0.00027895	0.029903
Contri 10%					49	
Days	11491	58	3962	1634.073275	5.91280283	633.829298
Difference						
Age	11491	0.000000	2548	745.52884	6.4916690	695.881488
Time Stage	11491	1.000000	5.000000	3.693325	0.011131	1.193224
Aged						
Numeric						
Oss Stage	11491	1.000000	6.000000	1.49195	0.01117426	1.197837
Aged						
Numeric						
File Size	11491	0.000000	48295	621.776956	12.4346814	1332.949133
					55	
Code churn	11491	0.00000	8.137754	2.659028e+03	709.30357	7.603456e+0
			e+06			4
Churn rate	11491	0.000000	43785.71	310.596780	12.2277954	1310.771761
			4286		2	

3.1.2 Histogram, box plot, normal Q-Q plot







3.1.3 Skewness and Kurtosis check

Metrics	Skewness			Kurtosis				
	Statistics	Std. Error	z-value	Statistics	Std. Error	z-value		
CVE Severity	0.1659	0.0955	1.7373	-1.0004	0.1910	-5.2382		
Ownership	0.9957	0.0229	43.5761	-0.0209	0.0457	-0.4567		
Num of Minor 10%	5.3191	0.0229	232.7788	40.1170	0.0457	877.8117		
Per of Minor 10%	-0.2888	0.0229	-12.6392	-1.6201	0.0457	-35.4499		
Avg of Minor Contri 10%	0.5876	0.0229	25.7146	-0.7820	0.0457	-17.1108		

Days Difference	-0.2815	0.0229	-12.3203	-0.6759	0.0457	-14.7900
Age	0.8587	0.0229	37.5774	-0.4880	0.0457	-10.6781
Time Stage Aged	-0.8369	0.0229	-36.6245	-0.1359	0.0457	-2.9747
Numeric						
Oss Stage Aged	2.5320	0.0229	110.8056	5.7030	0.0457	124.7880
Numeric						
File Size	11.2856	0.0229	493.8879	253.9972	0.0457	5557.7908
Code churn	106.6004	0.0229	4665.1112	11402.7559	0.0457	249507.2201
Churn rate	25.5640	0.0229	1118.7464	758.3895	0.0457	16594.5545

3.1.4 Shaprio-Wilk and Kologorow-Smirnow tests

Metrics	Shaprio-Wilk		Kolmogorov-Sm	irnov
	Statistics	Sig.	Statistics	Sig.
CVE Severity	0.8971	0.0000	0.9995	0.0000
Ownership	0.8746	0.0000	0.5380	0.0000
Num of Minor 10%	0.4976	0.0000	0.6089	0.0000
Per of Minor 10%	0.8113	0.0000	0.5000	0.0000
Avg of Minor Contri 10%	0.8807	0.0000	0.5000	0.0000
Days Difference	0.9497	0.0000	1.0000	0.0000
Age	0.8764	0.0000	0.9237	0.0000
Time Stage Aged Numeric	0.8495	0.0000	0.8966	0.0000
Oss Stage Aged Numeric	0.4640	0.0000	0.8413	0.0000
File Size	0.3893	0.0000	0.9453	0.0000
Code churn	0.0059	0.0000	0.9458	0.0000
Churn rate	0.0965	0.0000	0.8977	0.0000

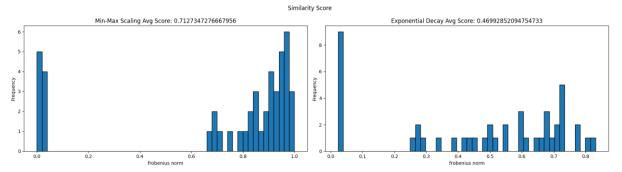
3.2 Possible Distortion Check

In summary, vulnerability proportion/threshold/locality does not have significant influence on correlation heatmap.

3.2.1 Proportion influence

Sample the vulnerable dataset as 10% - 100% to the non-vulnerable dataset.

	Frobenius Norm of Matrix Differences										
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
10%		0.4925	0.6945	4.018	0.9668	1.1172	1.1963	1.2680	1.2731	1.4400	
20%			0.4184	3.92628	0.5968	0.8514	0.8653	0.9323	0.9202	1.1338	
30%				3.9246	0.4991	0.5551	0.7219	0.6567	0.7922	0.8971	
40%					3.9006	3.9386	3.918	3.9287	3.9318	3.9357	
50%						0.5065	0.4408	0.5102	0.4017	0.7171	
60%							0.4028	0.3812	0.5271	0.4660	
70%								0.4939	0.3026	0.4382	
80%									0.4273	0.4149	
90%										0.4637	
100%											



	Mantel Test										
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
10%		0.9326	0.9943	0.9885	0.9858	0.9827	0.9790	0.9752	0.9759	0.9736	
20%			0.9326	0.9316	0.9300	0.9303	0.9294	0.9323	0.9297	0.9258	
30%				0.9968	0.9949	0.9931	0.9905	0.9871	0.9886	0.9864	
40%					0.9973	0.9974	0.9963	0.9927	0.9935	0.9919	
50%						0.9977	0.9965	0.9959	0.9953	0.9953	
60%							0.9971	0.9980	0.9968	0.9977	
70%								0.9986	0.9974	0.9978	
80%									0.9977	0.9988	
90%										0.9993	
100%											

3.2.2 Threshold influence

Check the similarity score between each threshold.

	Threshold vs Is_Defective												
		5%		10%			20%				50%		
	Cosine K-S Statistics Similarit		Cosine Similarit	K-S Statisti	cs	Cosine Similarit			Cosine Similarit				
	у	Statistic	P-value	y	Statistic	P-value	y	Statistic	P-value	у	Statistic	P-value	
5%				0.9196	0.0769	0.9765	0.7955	0.1282	0.5459	0.5715	0.1538	0.3159	
10%							0.9237	0.1025	0.8099	0.6818	0.1538	0.3159	
20%										0.8233	0.1282	0.5459	
50%													

	Threshold vs Vulnerable											
	5%			10%			20%			50%		
	Cosine Similarit	K-S Statistics				Cosine K-S Statistics Similarit		Cosine Similarit	K-S Statisti	cs		
	у	Statistic	P-value	у	Statistic	P-value	у	Statistic	P-value	у	Statistic	P-value
5%				0.9100	0.1282	0.5459	0.7651	0.1666	0.2296	0.5792	0.2179	0.0489
10%							0.9136	0.1025	0.8099	0.7248	0.1410	0.4221
20%										0.8851	0.1025	0.8099
50%												

	Threshold vs CVE Severity													
	5%			10%			20%				50%			
	Cosine K-S Statistics		Cosine K-S Statistics Similarit		Cosine K-S Statistics Similarit		cs	Cosine Similarit	K-S Statistics					
	у	Statistic	P-value	у	Statistic	P-value	у	Statistic	P-value	у	Statistic	P-value		
5%				0.8872	0.1282	0.5459	0.7028	0.1538	0.3159	0.5393	0.1794	0.1624		
10%							0.8906	0.1282	0.5459	0.7166	0.1666	0.2296		
20%										0.9090	0.1025	0.8099		
50%														

3.2.3 Locality Clustering influence

Check the similarity between heatmap generated by file component and group component.

Mantel Test									
	Group Component								
	Correlation	P-value							
File Component	0.8219009665635432	0.001							

3.3 Correlation Check

In summary:

• Check if is defective: Days Difference, Age

• Time Stage Aged Numeric: Num/Per of Minor 10%, Oss Stage Aged Numeric

• CVE Severity: Days Difference, Pre-/Post- release, Release Amounts

• **Pre-/Post- vs Code ownership**: No correlation in this case

Between each metric, it seems like they are independent/robust with each other.

3.3.1 Correlation

						C	orrelation	on							
	Is Defective		ive	Time Stage Aged			CVE Severity			Pre-release			Post-release		
					Numer	ic									
P: Pearson S: Spearman K: Kendall	Р	S	K	Р	S	K	P	S	K	P	S	K	P	S	К
Ownership	-0.12	-0.10	-0.08	-0.73	-0.62	-0.50	-0.03	-0.07	-0.05	0.05	0.03	0.03	-0.04	-0.03	-0.03
Num of contributor	0.16	0.22	0.19	0.38	0.70	0.58	0.11	0.08	0.06	-0.06	-0.06	-0.05	0.04	0.07	0.05
Num of Minor 10%	0.16	0.23	0.20	0.36	0.64	0.53	0.12	0.09	0.06	-0.06	-0.06	-0.05	0.04	0.07	0.06
Per of Minor 10%	0.23	0.23	0.20	0.64	0.63	0.52	0.06	0.07	0.05	-0.09	-0.06	-0.05	0.07	0.07	0.06
Avg of Minor Contri 10%	0.13	0.17	0.15	0.36	0.38	0.29	-0.01	0.01	0.01	0.04	0.03	0.03	-0.02	-0.01	-0.01
Days Difference	-0.81	-0.92	-0.82	0.22	0.32	0.25	0.45	0.44	0.35	-0.63	-0.69	-0.58	0.66	0.72	0.61
Age	-0.61	-0.60	-0.49	0.81	0.96	0.86	0.25	0.26	0.19	-0.34	-0.35	-0.28	0.36	0.37	0.30
Oss Stage Aged Numeric	0.25	0.28	0.27	-0.58	-0.59	-0.51	-0.02	-0.02	-0.02	0.01	-0.11	-0.11	-0.06	-0.11	-0.05
File Size	0.14	0.40	0.33	0.17	0.25	0.19	0.10	0.16	0.12	0.03	0.06	0.05	0.03	0.05	0.04
Code churn	0.01	0.36	0.30	0.02	0.49	0.38	0.12	0.14	0.10	0.03	-0.07	-0.06	0.04	0.07	0.05
Churn rate	-0.01	-0.00	-0.00	0.09	0.45	0.36	0.06	0.01	0.01	0.04	0.01	0.00	0.04	0.03	0.03
Is Pre-release	/	/	/	-0.18	-0.23	-0.22	-0.41	-0.32	-0.28	/	/	/	/	/	/
Is Post-release	/	/	/	0.19	0.24	0.23	0.40	0.31	0.28	/	/	/	/	/	/
Release Amounts	-0.84	-1.00	-0.98	0.28	0.45	0.37	0.43	0.43	0.34	/	/	/	/	/	/
Release Amounts Aged	-0.23	-0.23	-0.19	0.82	0.85	0.74	0.27	0.29	0.21	/	/	/	/	/	/

3.3.2 Robustness

Robustness (Multiple Linear Regression)										
	R-squared	Adj. R2	F-statistic	Coefficient	Std err	P-value				
Is Defective										
Days Difference	0.659	0.659	4.582e+04	-0.0005	2.51e-06	0.000				
Days Difference (Controlled by Classic)	0.665	0.665	1.572e+04	-0.0005	2.5e-06	0.000				
Age	0.367	0.367	1.373e+04	-0.0003	2.77e-06	0.000				
Age (Controlled by Classic)	0.388	0.388	5020.	-0.0003	2.72e-06	0.000				
Time Stage Aged Numeric										

Num of Minor 10%	0.131	0.131	1734.	0.0190	0.000	0.000
Per of Minor 10%	0.413	0.413	8099.	1.9658	0.022	0.000
Per of Minor 10%	0.416	0.415	2723.	2.0164	0.023	0.000
(Controlled by Classic)						
Oss Stage Aged Numeric	0.332	0.332	5715.	-0.5741	0.008	0.000
Oss Stage Aged Numeric	0.347	0.347	2033.	-0.5628	0.008	0.000
(Controlled by Classic)						
Per of Minor 10% +	0.553	0.552	7093.	1.5344 +	0.020 +	0.000
Oss Stage Aged Numeric				-0.3972	0.007	
CVE Severity						
Days Difference	0.202	0.201	222.7	0.0021	0.000	0.000
Days Difference (Controlled by Classic)	0.203	0.200	74.64	0.0020	0.000	0.000
Age	0.060	0.059	56.50	0.0005	6.06e-05	0.000
Days Difference (Controlled by Minor)	0.202	0.200	111.4	0.0020	0.000	0.002

4 Problems and Further

4.1 Problems

- With the newly fetch data from NVD, when I move to the correlation analysis between metrics and CVE Severity, it shows that there is no correlation between CVE Severity and any metric, which is significant different from the results from origin dataset. While, only 400+ entries are updated in the new dataset (1200+ in total).
- The definition of the OSS Stage metric. I just randomly assign the stage with numeric value, but it seems like there are some correlations.
- Non-vulnerable dataset source and definition.
- Is there any point of the metric needed to be re-defined/added? Like adding `Major` attribute.

4.2 Further

- Prediction?
- The reasons that affect or cause minor?
- Correlation between metrics? (Paper: Effects of measurements on correlations of software code metrics)