

## 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.**)

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

### 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
<b>Basic</b>		
URL	The URL link source	GitHub commit/pull request
Repo Name	The Repository name	GitHub repository name
Date	The time of the URL commit	The time that a vulnerability was solved
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
<b>Code ownership</b>		
Ownership	The ownership	The highest ownership of a component
Num of Contributor	The sum of contributors	The sum of the contributors to a component
Num of Minor T%	The amount of the minor contributors	The total amount of the minor contributors to a component. Contributor with ownership under T% is Minor contributor. (* T%: 5%, 10%, 20%, 50%)
Per of Minor T%	The proportion of the minor contributors	The proportion of the minor contributors over all the contributor amount

Avg of Minor Contri T%	The average of minor contributor's ownership	The average value of the minor contributors' ownership
<b>Time/Release (See 2.2 for details)</b>		
Days Difference	The project existing time	The existing time of the project in GitHub repository till the <b>Date</b>
Age	The component lifetime	The component lifetime calculated based on Git Log info
Time Stage Numeric	Five Time stages	The five Time stages' numerical value. Calculated by <b>Days Difference</b>
Time Stage Aged Numeric	Five Time stages Aged	The five Time stages' numerical value. Calculated by <b>Age</b>
Oss Stage Numeric	Six Oss stages	The six Oss stages' numerical value. Calculated by <b>Days Difference</b>
Oss Stage Aged Numeric	Six Oss stages Aged	The six Oss stages' numerical value. Calculated by <b>Age</b>
Is Pre-release	Vulnerability found at pre-release	The release tag where the vulnerability found is pre-release
Is Post-release	Vulnerability found at post-release	The release tag where the vulnerability found is post-release
Release Amounts	Affected releases	The affected releases until the project latest time
Release Amounts Aged	Affected releases Aged	The affected releases within the lifetime of the vulnerability existing
<b>Classic metrics</b>		
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
<b>Time Stage</b>		
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
<b>Oss Stage</b>		
SI	Success Initialisation. Has at least one successful release	1
TI	Tragedy Initialisation. Within the given time period (>1year), no release	2
SG	Success Growth. >= 3 releases AND >= 6 months between releases	5
TG	Tragedy Growth. 1 or 2 releases and >=1 year since the last release at the time of data collection	6
II	Indeterminate Initialisation.	3

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

### 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).

```
PS C:\Users\jiawe\Desktop\thesis\code\data_process\vulnerable_process> python app.py -h
usage: app.py [-h] [-c] [-p] [--cve] [--torchflow] [--files Src Dst]

Vulnerable_process Interface

options:
  -h, --help            show this help message and exit
  -c, --collect          Collect data from the vulnerability file
  -p, --process          Process the dataset
  --cve                 Process CVE dataset only
  --torchflow           Process Pytorch/Tensorflow dataset only
  --files Src Dst       Input two custom files
```

### 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.

```
PS C:\Users\jiawe\Desktop\thesis\code\data_process\non_vulnerable_process> python app.py -h
usage: app.py [-h] [-p] [--url URL] [--dst Dst]

Non_Vulnerable_process Interface

options:
  -h, --help            show this help message and exit
  -p, --process          Process the default URLs
  --url URL             Specify the REPO URL
  --dst Dst             Specify the result destination
```

## 3 Result Analysis

### 3.1 Exploring Nature of Data

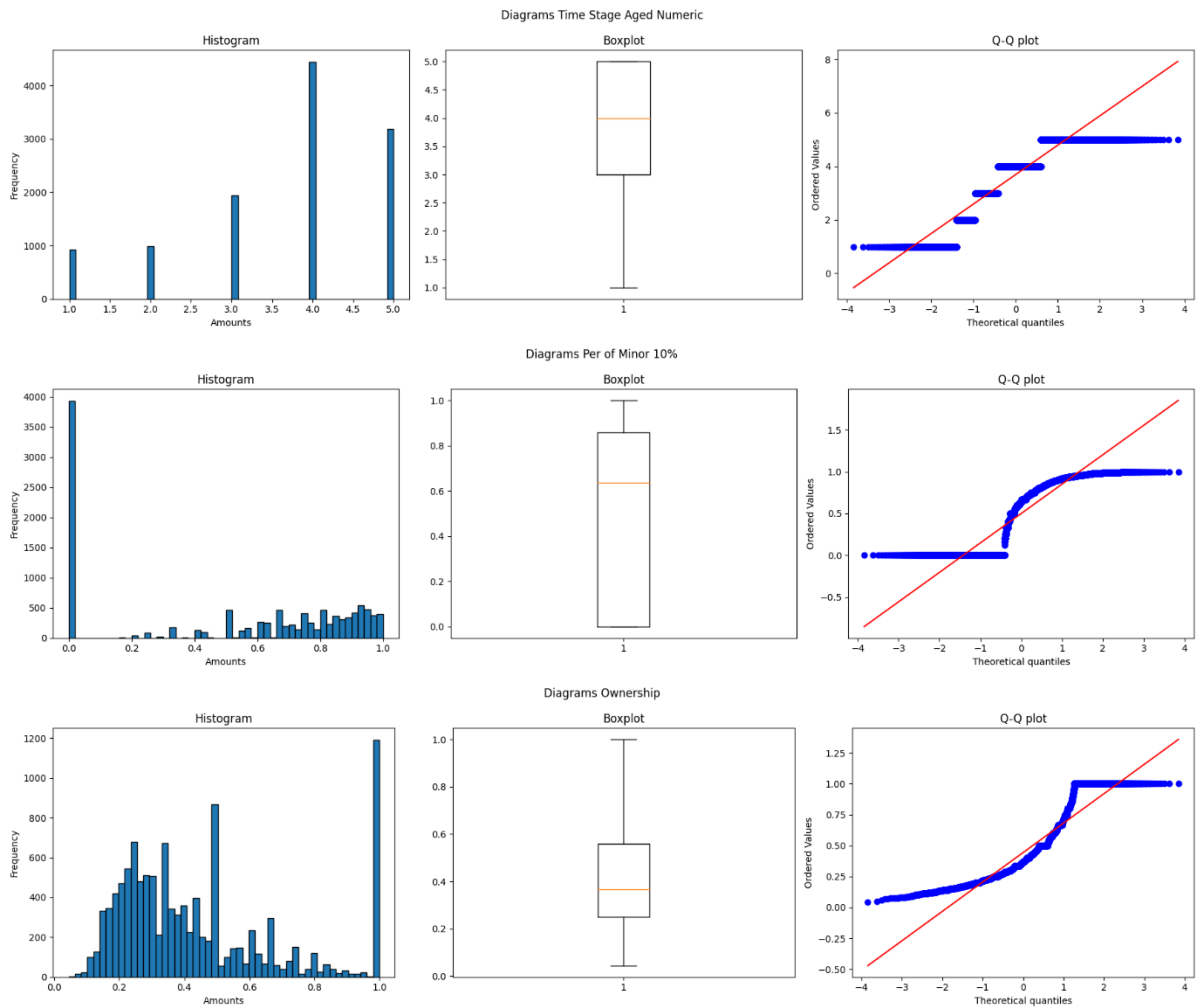
In summary, the dataset is not normally distributed.

#### 3.1.1 Descriptive statistics

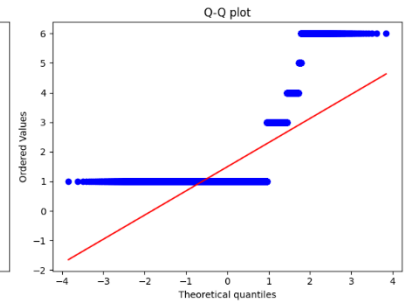
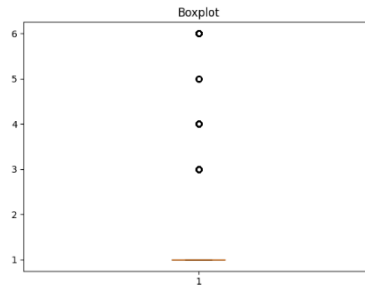
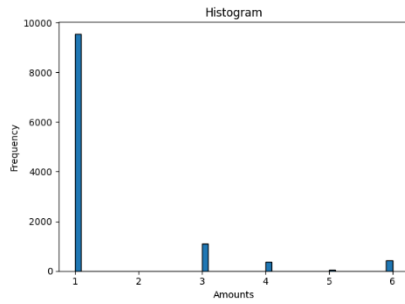
Metrics	N	Min	Max	Mean		Std. Dev.
				Statistics	Std. Error	
CVE Severity	658	3.3	9.9	6.465502	0.05183	1.329519
Ownership	11491	0.044088	1.00000	0.443906	0.0023718	0.254253
Num of Minor 10%	11491	0.0	364	11.495779	0.212516	22.780893

Per of Minor 10%	11491	0.000000	1.000000	0.499368	0.003641055	0.390307
Avg of Minor Contri 10%	11491	0.000000	0.100000	0.030714	0.0002789549	0.029903
Days Difference	11491	58	3962	1634.073275	5.91280283	633.829298
Age	11491	0.000000	2548	745.52884	6.4916690	695.881488
Time Stage Aged Numeric	11491	1.000000	5.000000	3.693325	0.011131	1.193224
Oss Stage Aged Numeric	11491	1.000000	6.000000	1.49195	0.01117426	1.197837
File Size	11491	0.000000	48295	621.776956	12.434681455	1332.949133
Code churn	11491	0.000000	8.137754e+06	2.659028e+03	709.30357	7.603456e+04
Churn rate	11491	0.000000	43785.714286	310.596780	12.22779542	1310.771761

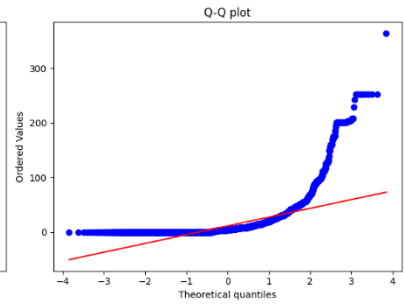
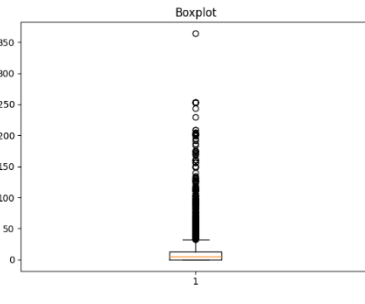
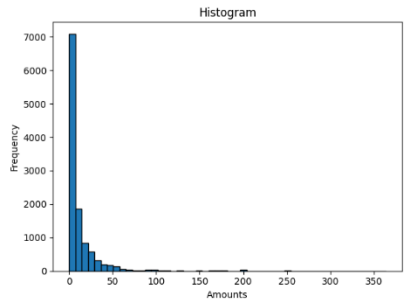
### 3.1.2 Histogram, box plot, normal Q-Q plot



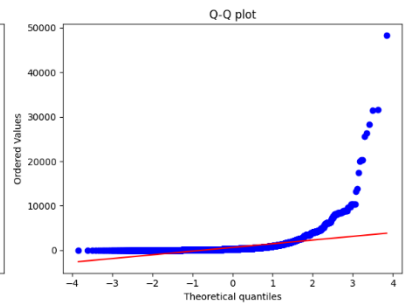
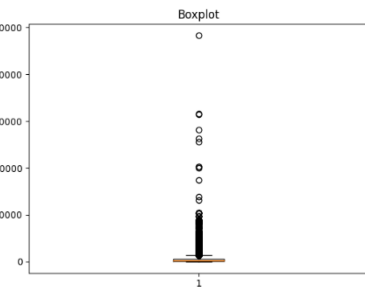
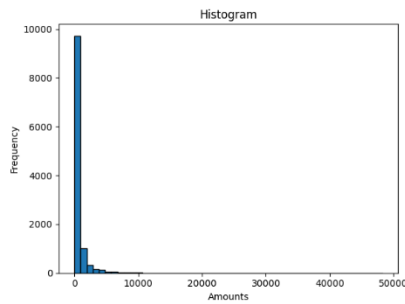
Diagrams Oss Stage Aged Numeric



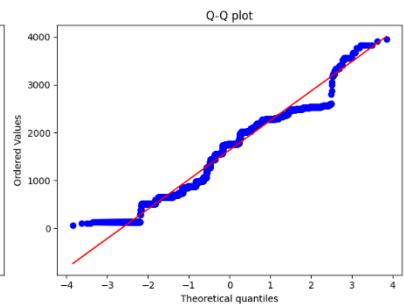
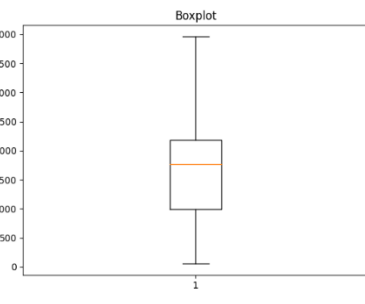
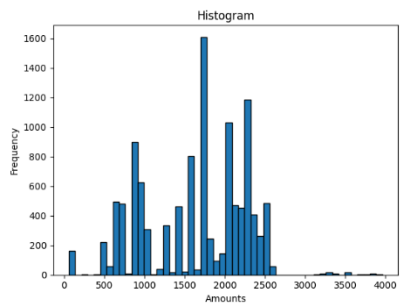
Diagrams Num of Minor 10%



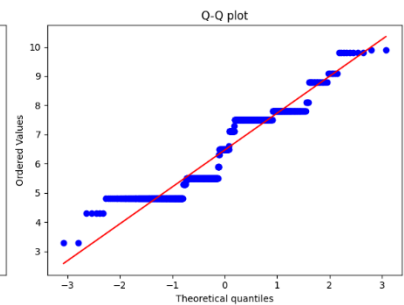
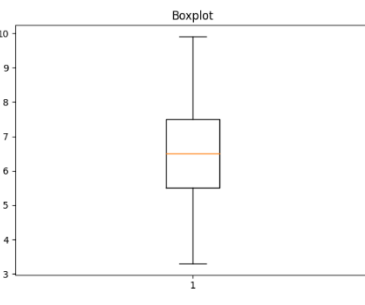
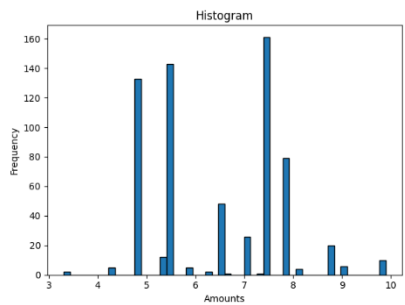
Diagrams File Size

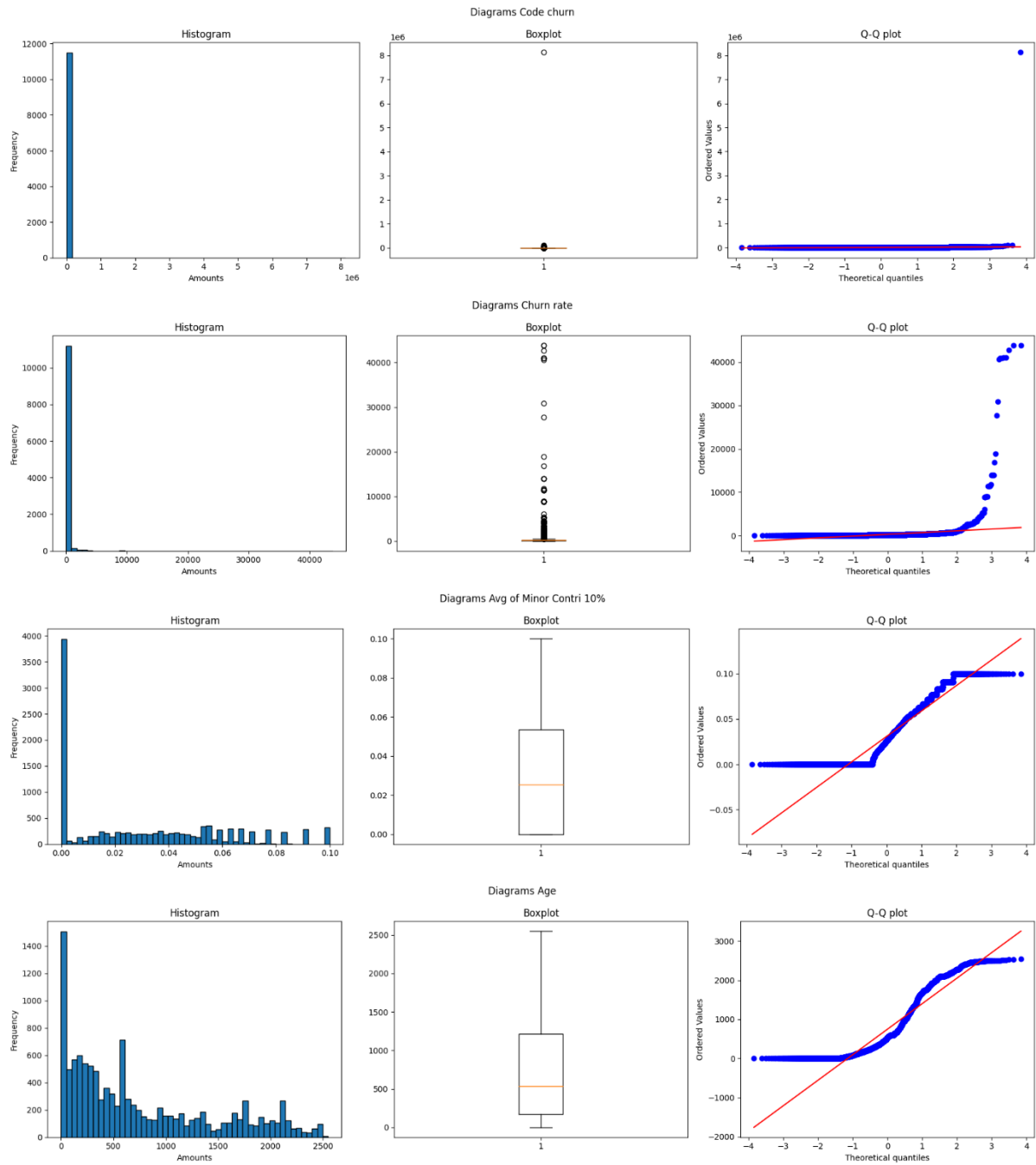


Diagrams Days Difference



Diagrams CVE Severity





### 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







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

### 3.3 Correlation Check

- **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

### 3.3.1 Correlation

### 3.3.2 Robustness

Robustness (Multiple Linear Regression)						
	R-squared	Adj. R2	F-statistic	Coefficient	Std err	P-value
<i>Is Defective</i>						
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
<i>Time Stage Aged Numeric</i>						

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% (Controlled by Classic)	0.416	0.415	2723.	2.0164	0.023	0.000
Oss Stage Aged Numeric	0.332	0.332	5715.	-0.5741	0.008	0.000
Oss Stage Aged Numeric (Controlled by Classic)	0.347	0.347	2033.	-0.5628	0.008	0.000
Per of Minor 10% + Oss Stage Aged Numeric	0.553	0.552	7093.	1.5344 + -0.3972	0.020 + 0.007	0.000
<b>CVE Severity</b>						
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)