EXPLORATORY DATA REPORT FOR VIRUS TOTAL DETECTION ENGINE LOGS

Introduction and Metadata definition

The raw data is identified to be in the JSON format and contains data columns identifying the following:

Variables	Definition
permalink	A permalink or permanent link is a URL that is intended to remain unchanged for many years into the future
submission	Contains information abut how the link was submitted for scan
url	The address of the webpage
response_code	The response codes gotten when trying to access the url
scan_date	The date the url was scanned
scan_id	The id assigned to the url
verbose_msg	Information regarding the status of the scan
last_seen	The last time the url was requested to be scanned
filescan_id	
positives	The number of engines which flagged this url
first_seen	The first time the url was scanned
total	The total number of engines which scanned the url
additional_info	The summary of the engine scan results for each url
scans	The Engines and results of the scans for each individual engine

Data Transformation

The data is transformed as required to extract insights following the steps below.

1. The data was loaded and manipulated using the Pandas library.
2. The 'url', 'total' and 'scan_date' columns were imported into a dataframe from the original data file, with the 'total' column being renamed as 'totalScannedEngines'.
3. The FQDN was extracted from the 'url' and imported into the dataframe.
4. The 'numberOfDetectionEngines' were extracted from information in the scan column of the original data.
5. The 'detectedURL' column was also parsed from the scan column.

Metadata

Column	Definition
FQDN	The fully qualified data name extracted from the url
url	The url being scanned
numberOfDetectionEngines	The number of detection engines used for scanning.
detectedURL	The total number of engines which flagged a url as malicious
totalScannedEngines'	The total number of Engines used to scan a url
scan_date	The date the scan was performed

RESULTS AND VISUALIZATION

1. Engine Detection:

An introductory exploration of the data reveals the Engines which detected malicious FQDNs for each individual FQDN. A merge and sort operation performed on the data allowed for transforming of the data to reveal all Engines used to scan the FQDNs and the number of malicious classifications by each Engine.

A sample of the results showing the top 10 Engines sorted in descending order of files malicious classifications is shown in the bar plot below.

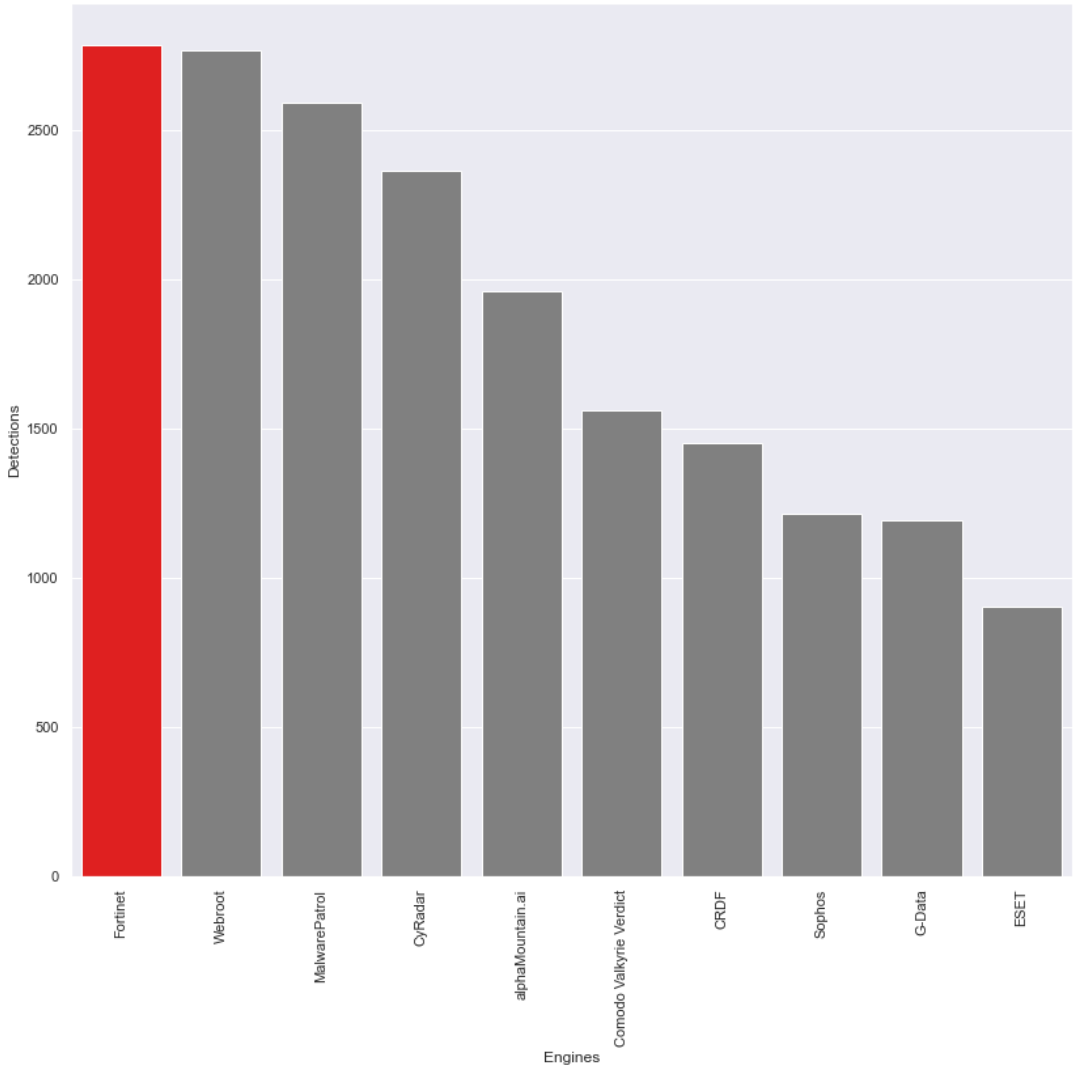

Figure 1: Detection engines by number of malicious sites identified

The above plot indicates that Fortinet identifies the highest number of malicious FQDNs, followed closely by Webroot, with ESET coming in last among the top 10 engines.

2. Similarity between Detection Engines:

Further analysis and data transformation enables us to discover what detection engines get the most similar results. We assess this using a heat map to visualize the similarity score between each of the top 10 detection engines sorted by how many malicious classifications were identified. We calculated the similarity between two engines as

$$\frac{Number\ of\ detected\ sites(A) * 100}{Number\ of\ detected\ sites(B)}$$

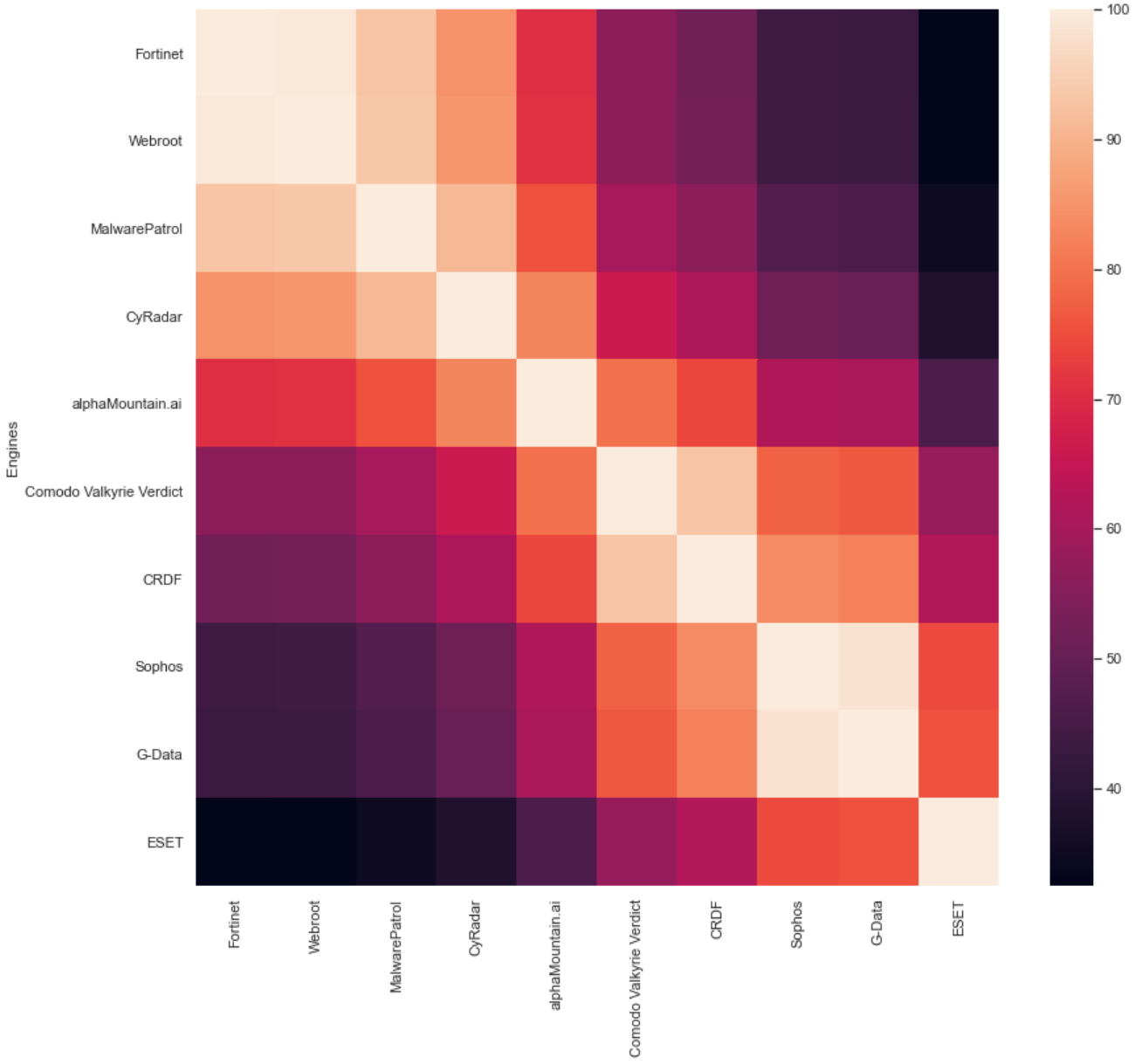

Figure 2: Similarity Heatmap of Top Ten Engines

The above plot shows the similarity score of each Detection Engine among the top 10. We immediately identify that the darker colors indicate low similarity, and the brighter colors indicate greater similarity

between the identified columns. The similarity range for each color can be seen in the Legend on the right of the image.

3. **Cumulative Distribution Frequency:**

From the CDF, we visualize the contribution of each individual engine to the total number of malicious FQDNs identified in the data, with the top 3 contributors being:

- Fortinet
- Webroot
- Malware Patrol

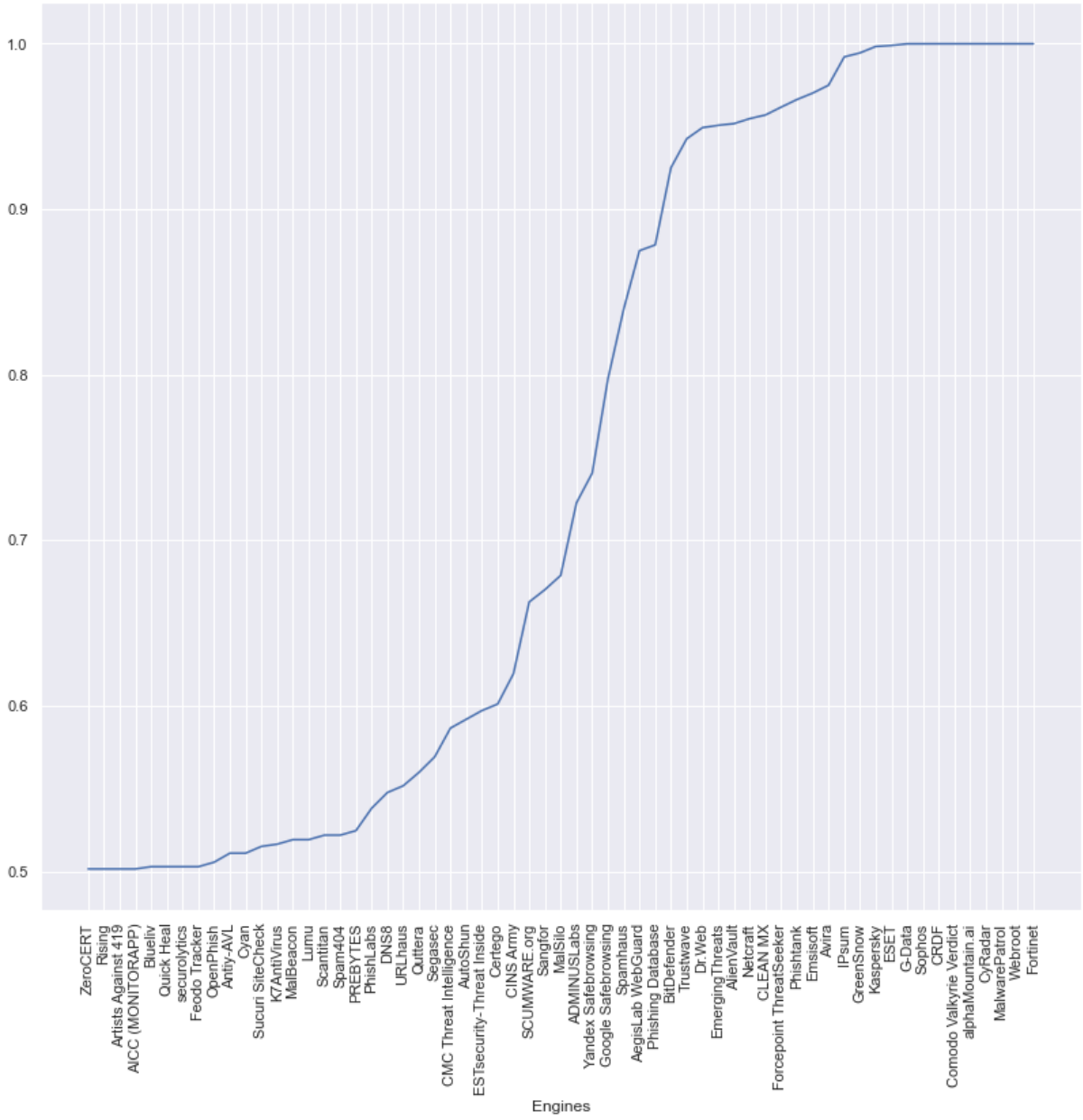

Figure 3: Cumulative Frequency Curve showing contribution of all engines

4. Generated CSV output snippet:

A sample of the output csv file generated after transforming the data is as shown below.

The CSV File created is as shown in the figure below

		permalink	submission	url	response_	scan_date	scan_id	verbose_r	last_seen	filescan_i	positives	first_seen	total	additional	scal
2	0	https://www.virustotal.com/	{'submitter_region': None,	http://120.48.19.10/	1	5/7/2021 8:53	cff8fb5f59	Scan finisl	5/7/2021 8:53		6	12/6/2020 3:24	87	{'resolutic	{'CN
3	1	https://www.virustotal.com/	{'submitter_region': None,	http://43.251.159.144	1	5/7/2021 8:53	4e53aa685	Scan finisl	5/7/2021 8:53		5	7/7/2019 10:25	87	{'Respons	{'CN
4	2	https://www.virustotal.com/	{'submitter_region': None,	http://180.167.168.2/	1	5/7/2021 8:53	2dedd8f24	Scan finisl	5/7/2021 8:53		5	4/14/2019 9:34	87	{'Respons	{'CN
5	3	https://www.virustotal.com/	{'submitter_region': 'be', 's	http://clas.sangor.es/	1	5/7/2021 8:58	2a87f7cc6	Scan finisl	5/7/2021 8:58		0	10/8/2019 12:28	88	{'Respons	{'CN
6	4	https://www.virustotal.com/	{'submitter_region': 'wa', 's	https://crewsalon.co	1	5/7/2021 8:58	757b9bc0a	Scan finisl	5/7/2021 8:58		0	8/12/2020 11:00	87	{'BitDefer	{'CN
7	5	https://www.virustotal.com/	{'submitter_region': None,	http://103.146.221.12	1	5/7/2021 8:53	8b3fa439f	Scan finisl	5/7/2021 8:53		6	2/1/2021 8:54	87	{'Respons	{'CN
8	6	https://www.virustotal.com/	{'submitter_region': 'wa', 's	https://whenrelation	1	5/7/2021 8:58	3c4b11c18	Scan finisl	5/7/2021 8:58		0	9/4/2020 9:53	87	{'Respons	{'CN
9	7	https://www.virustotal.com/	{'submitter_region': 'wa', 's	http://www.jndsgg.c	1	5/7/2021 8:58	23fb1c9f9	Scan finisl	5/7/2021 8:58		0	1/25/2017 10:20	87	{'Respons	{'CN
10	8	https://www.virustotal.com/	{'submitter_region': None,	http://91.203.145.116	1	5/7/2021 8:53	9cb7f7374	Scan finisl	5/7/2021 8:53		7	4/15/2016 17:56	87	{'Respons	{'CN
11	9	https://www.virustotal.com/	{'submitter_region': None,	http://81.70.223.143/	1	5/7/2021 8:53	45c759a22	Scan finisl	5/7/2021 8:53		0	4/18/2021 16:24	87	{'resolutic	{'CN
12	10	https://www.virustotal.com/	{'submitter_region': '?', 'sul	http://mipasillointeri	1	5/7/2021 8:58	9cfac1d85	Scan finisl	5/7/2021 8:58	8a21d5d7	2	7/27/2020 16:53	87	{'Respons	{'CN
13	11	https://www.virustotal.com/	{'submitter_region': 'wa', 's	https://foodforthethi	1	5/7/2021 8:58	71d094d6	Scan finisl	5/7/2021 8:58		0	5/7/2021 8:58	87	{'BitDefer	{'CN

Figure 4: Top ten rows of the transformed dataset

CONCLUSION

The malicious nature of a 'url' is independent of the number of Engines used to perform a scan. The reliability of an engine should also be considered to avoid false positives. As seen in insights gotten from this data, the most reliable engines are:

1. Fortinet
2. Webroot
3. Malware Patrol

This is shown with the CDF curve in Figure 3.

The heatmap in Figure 2 shows the similarity scores for each of the engines, allowing us to pick the most similar engines. This allows us to use a particular engine as a benchmark when comparing results.

Finally, the engine with the most detections is the Fortinet Engine as shown in Figure 1. Assuming the engine optimally minimizes false positives, this would be ideal as a benchmark to compare the other engines.