information

Article

A New Time Series Dataset for Cyber-Threat Correlation, Regression and Neural-Network-Based Forecasting

Fahim Sufi

School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC 3004, Australia;
fahim.sufi@monash.edu

Abstract: In the face of escalating cyber threats that have contributed significantly to global economic losses, this study presents a comprehensive dataset capturing the multifaceted nature of cyber-attacks across 225 countries over a 14-month period from October 2022 to December 2023. The dataset, comprising 77,623 rows and 18 fields, provides a detailed chronology of cyber-attacks, categorized into eight critical dimensions: spam, ransomware, local infection, exploit, malicious mail, network attack, on-demand scan, and web threat. The dataset also includes ranking data, offering a comparative view of countries' susceptibility to different cyber threats. The results reveal significant variations in the frequency and intensity of cyber-attacks across different countries and attack types. The data were meticulously compiled using modern AI-based data acquisition techniques, ensuring a high degree of accuracy and comprehensiveness. Correlation tests against the eight types of cyber-attacks resulted in the determination that on-demand scan and local infection are highly correlated, with a correlation coefficient of 0.93. Lastly, neural-network-based forecasting of these highly correlated factors (i.e., on-demand scan and local infection) reveals a similar pattern of prediction, with an MSE and an MAPE of 1.616 and 80.13, respectively. The study's conclusions provide critical insights into the global landscape of cyber threats, highlighting the urgent need for robust cybersecurity measures.

Keywords: global cyber-attack dataset; cyber intelligence; time-series data; low-code data acquisition; cyber prediction; forecasting; regression; neural network

check for updates

Citation: Sufi, F. A New Time Series Dataset for Cyber-Threat Correlation, Regression and Neural-Network-Based Forecasting. *Information* **2024**, *15*, 199. https://doi.org/10.3390/info15040199

Academic Editors: Francesco Camastra, Maanak Gupta, Leandros Maglaras and Jiguo Li

Received: 16 February 2024
Revised: 19 March 2024
Accepted: 2 April 2024
Published: 4 April 2024

1. Introduction

In the contemporary digital landscape, cyber-attacks have emerged as formidable threats, inflicting substantial economic and security damages globally. Recent estimates suggest that these attacks have contributed to an alarming USD 3.5 trillion in economic losses, underscoring the urgency for comprehensive research in this domain [1,2]. Other than the detrimental economic impact, it has become increasingly evident that cyber-attacks manifest profound influences within both the psychological and behavioral strata, precipitating far-reaching ramifications at the societal echelon [3]. Addressing this critical need, our research team has compiled an extensive dataset, encompassing a period from 11 October 2022 to December 2023, marking a significant stride in understanding and mitigating the impacts of cyber threats.

This novel dataset, unprecedented in its scope and detail, covers 399 dates over 14 months, meticulously gathered from open-source daily statistics (e.g., Kaspersky [4]). It encapsulates data on eight critical dimensions of cyber-attacks: spam [5], ransomware [6], local infection [7], exploit [8], malicious mail [9], network attack [10], on-demand scan [11], and web threat [12]. The dataset's global reach, spanning 225 countries, provides a comprehensive view of the cyber-threat landscape, making it an invaluable resource for longitudinal studies in this field. The dataset offers both percentage-wise attack statistics and world-ranking-wise attack statistics, offering researchers a multifaceted view of the cyber-threat landscape. Though some of our recent projects have already utilized a smaller subset of this dataset, demonstrating its versatility and applicability across various domains [13–16], the full dataset had yet to be publicly released (as shown in Table 1). The public release and description of this dataset offers numerous benefits. It not only fosters transparency and collaboration among researchers but also accelerates the development of effective countermeasures against cyber threats. By providing a rich source of empirical data, it enables the scientific community to conduct more nuanced and comprehensive analyses (as well as predictions), ultimately contributing to the enhancement of global cyber resilience. The data acquisition process employed modern AI-based techniques, ensuring high accuracy and reliability. These techniques included advanced data manipulation and integration methods, tailored to capture the nuanced dynamics of cyber-attacks.

In the last six years there have studies that have focused on the regression [17–19], correlation [20,21], and prediction [22–29] of cyber-attacks in order to boost cyber preparedness. However, because of the lack of a robust cyber-attack dataset, these existing studies faced limitations in terms of their prediction accuracy. After generating the novel and robust dataset in this study, several attempts at statistical modelling (e.g., correlation, regression etc.) were performed, along with predictions related to cyber threats against randomly selected countries like Australia, New Zealand, Qatar, Saudi Arabia, and UAE. A thorough performance evaluation revealed a higher accuracy in predicting cyber threats, in comparison with existing studies [22–29].

The core contributions of this study can be succinctly summarized as follows:

- Innovative methodology: We developed an automated methodology for the systematic collection of global cyber-attack data, offering a novel and efficient approach to dataset creation.
- Comprehensive dataset: We generated a comprehensive dataset spanning 225 countries over a 14-month period, comprising 77,623 rows and 18 fields, capturing the multifaceted nature of cyber-attacks across diverse dimensions (publicly available at https://github.com/DrSufi/CyberData, accessed on 14 February 2024).
- Correlation and interdependency insights: We unveiled significant correlations between specific cyber-attack dimensions, particularly highlighting a notable correlation (coefficient of 0.93) between on-demand scan and local infection. Subsequent linear regression provided insights into their interdependency.
- Neural-network-based forecasting: We applied neural-network-based forecasting to predict temporal patterns of highly correlated cyber threat dimensions (on-demand scan and local infection), achieving a high level of predictive accuracy with mean squared error (MSE) and mean absolute percentage error (MAPE) values of 1.616 and 80.13, respectively.
- Practical applicability: We positioned the research within the realm of practical cybersecurity applications, offering valuable tools for risk assessment, strategic planning, and informed decision-making in the face of dynamic and evolving cyber threats.

In summary, this dataset represents a pivotal resource for the ongoing battle against cyber threats by releasing the most robust global cyber-attack dataset and demonstrating its use in cyber analytics and prediction (as highlighted clearly in Table 1). Its comprehensive nature, coupled with the advanced methodologies employed in its compilation and analysis, positions it as a cornerstone for future cybersecurity research. Due to the significant contributions of both releasing a robust global cyber dataset and enhancing analytical capabilities for the prediction of cyber threats, they will be emphasized in the subsequent sections.

Table 1. Summary of the contributions of this study.

Reference	Start Date	End Date	Day	Public Release with Data Descriptors	Correlation Analysis	Regression Analysis	Neural-Network-Based Forecasting
[13]	11 October 2022	31 October 2022	21	No	No	No	No
[16]	11 October 2022	25 December 2022	76	No	No	No	No
[15]	11 October 2022	6 April 2023	178	No	No	No	No
[14]	11 October 2022	12 July 2023	275	No	No	No	No
This	11 October 2022	11 December 2023	427	Yes	Yes	Yes	Yes

2. Methods

The methodology employed in the aggregation and collation of this comprehensive cyber dataset leverages a combination of modern low-code tools and AI-based robotic process automation (RPA), ensuring both efficiency and accuracy in data collection and processing [16]. This section details the methods applied for data collection, treatment, validation, and curation, along with notes on data quality and noise.

2.1. Data Collection and Treatment

2.1.1. Automated Data Collection

The data collection process was initiated through Microsoft Power Automate [30], which scheduled a cloud flow to trigger an unattended Microsoft Power Automate Desktop flow. This setup allowed for the autonomous operation of a pre-configured Azure virtual machine, even when signed out, thereby facilitating continuous data collection. The Microsoft Power Automate Cloud Flow (highlighted as step 1 in Figure 1), in a completely autonomous manner, logs in to a Microsoft Azure remote machine at 6:00 a.m. every morning. As seen from Figure 2, the cloud flow autonomously operates for 20 to 30 min and during this time requests, queues, assigns, connects, and executes, before finally logging out (i.e., finalizes) of the remote machine. The cloud flow seen in Figure 2, calls for a sequence of steps (i.e., RPA) within the Microsoft Azure virtual machine (i.e., remote machine). This machine has Microsoft Power Automate Desktop installed, which carries out all the sequential RPA steps.

2.1.2. Robotic Process Automation

As Microsoft Power Automate Cloud Flow initiates the remote machine (as seen in Figure 2), it also logs in to this remote machine as an admin user. After logging in as an admin user, a series of RPA tasks are executed. As seen from Figure 3, these tasks involve reaching eight different sites, clicking the appropriate button to initiate a download, renaming the files (e.g., NetworkAttack.xls, Spam.xls, etc.), and saving them in Microsoft One Drive. Within the RPA tasks of Microsoft Power Automate Desktop, as shown in Figure 3, the task called "Launch New Chrome" launches the Chrome browser in an automated manner and points to the appropriate URLs from the eight different web links [4–12]. The task "Wait for web page content" allows the process to stall until the page content of the cyber statistics loads completely. Then, the download button is clicked with "Click link on web page" task. Next, "Focus window" focuses on the "Save As" window for renaming the files. The RPA task "Populate text field in window" enters the new file name in order to save them inside a Microsoft OneDrive for Business folder location. The modern RPA routines demonstrated in this method (i.e., Figure 3) use several AI-based techniques [31,32] for locating items (such as download buttons, or items or links) through computer vision. These RPA's can perform optical character recognition (OCR) in order to locate download links from images, mimicking the manual actions of humans [33]. In the context delineated by step 2 of Figure 1, OCR, denoting image processing, serves the purpose of identifying a download button's whereabouts and effectuating its activation automatically to commence the download process.

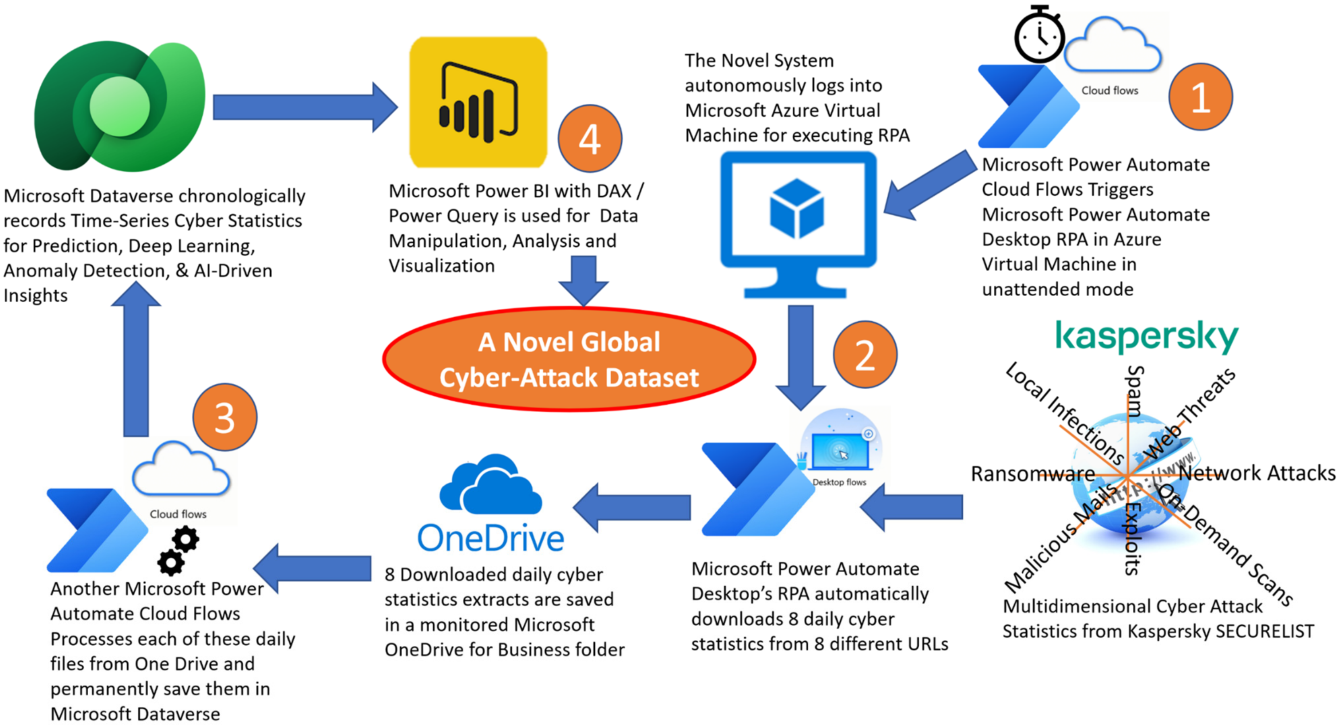
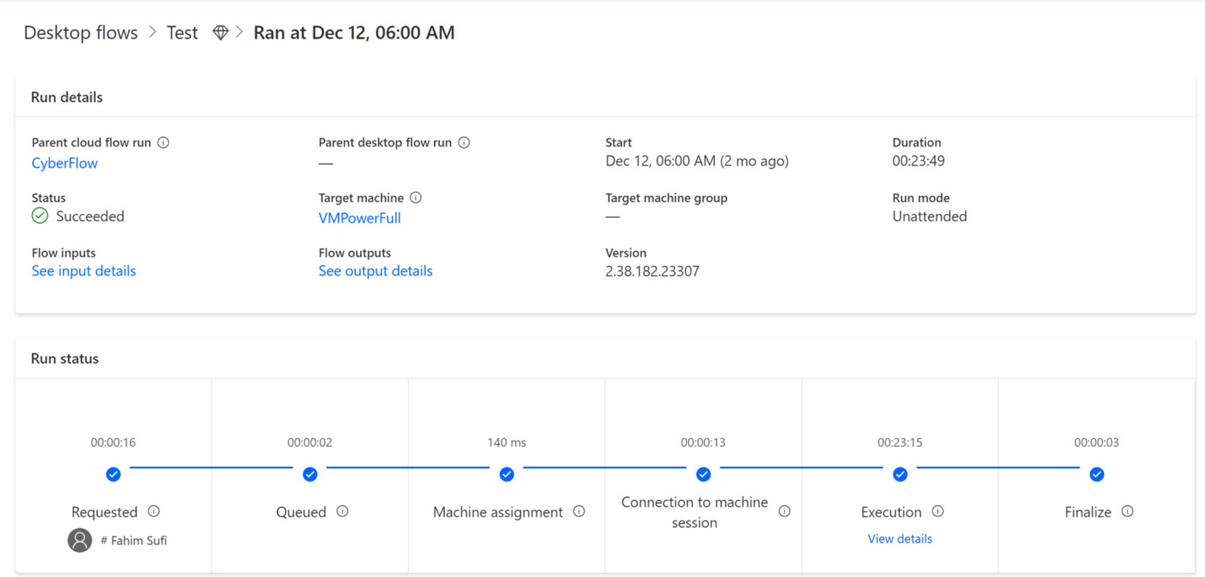

Microsoft Dataverse chronologically records Time-Series Cyber Statistics for Prediction, Deep Learning, Anomaly Detection, & AI-Driven Insights
4
Microsoft Power BI with DAX / Power Query is used for Data Manipulation, Analysis and Visualization
The Novel System autonomously logs into Microsoft Azure Virtual Machine for executing RPA
Cloud flows
1
Microsoft Power Automate Cloud Flows Triggers Microsoft Power Automate Desktop RPA in Azure Virtual Machine in unattended mode
A Novel Global Cyber-Attack Dataset
2
kaspersky
Local Infections
Spam
Web Threats
Ransomware
Network Attacks
Malicious Mails
Exploits
On-Demand Scans
Multidimensional Cyber Attack Statistics from Kaspersky SECURELIST
3
Cloud flows
Another Microsoft Power Automate Cloud Flows Processes each of these daily files from One Drive and permanently save them in Microsoft Dataverse
OneDrive
8 Downloaded daily cyber statistics extracts are saved in a monitored Microsoft OneDrive for Business folder
Desktop flows
Microsoft Power Automate Desktop's RPA automatically downloads 8 daily cyber statistics from 8 different URLs

Figure 1. Method of generating the novel global cyber-attack dataset.


Desktop flows > Test > Ran at Dec 12, 06:00 AM

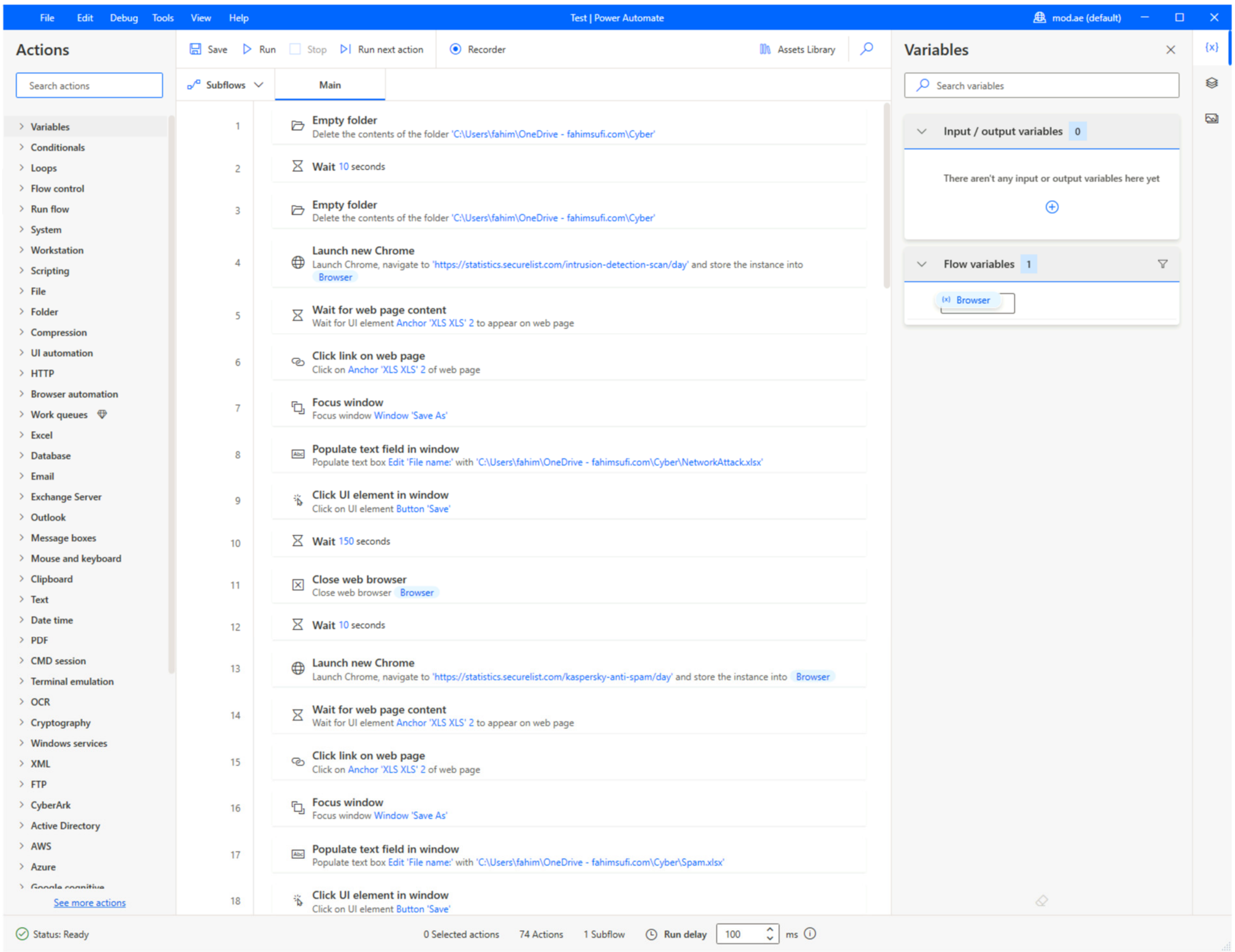
Run details

Parent cloud flow run ⓘ CyberFlow	Parent desktop flow run ⓘ —	Start Dec 12, 06:00 AM (2 mo ago)	Duration 00:23:49
Status ✓ Succeeded	Target machine ⓘ VMPowerFull	Target machine group —	Run mode Unattended
Flow inputs See input details	Flow outputs See output details	Version 2.38.182.23307	

Run status

00:00:16	00:00:02	140 ms	00:00:13	00:23:15	00:00:03
✓	✓	✓	✓	✓	✓
Requested ⓘ # Fahim Sufi	Queued ⓘ	Machine assignment ⓘ	Connection to machine session ⓘ	Execution ⓘ View details	Finalize ⓘ

Figure 2. Microsoft Power Automate Cloud Flow calling the Azure remote machine (called VMPowerFull) in a completely automated manner.



1 Empty folder
Delete the contents of the folder 'C:\Users\fahim\OneDrive - fahimsufi.com\Cyber'

2 Wait 10 seconds

3 Empty folder
Delete the contents of the folder 'C:\Users\fahim\OneDrive - fahimsufi.com\Cyber'

4 Launch new Chrome
Launch Chrome, navigate to 'https://statistics.securelist.com/intrusion-detection-scan/day' and store the instance into Browser

5 Wait for web page content
Wait for UI element Anchor 'XLS XLS' 2 to appear on web page

6 Click link on web page
Click on Anchor 'XLS XLS' 2 of web page

7 Focus window
Focus window Window 'Save As'

8 Populate text field in window
Populate text box Edit 'File name:' with 'C:\Users\fahim\OneDrive - fahimsufi.com\Cyber\NetworkAttack.xlsx'

9 Click UI element in window
Click on UI element Button 'Save'

10 Wait 150 seconds

11 Close web browser
Close web browser Browser

12 Wait 10 seconds

13 Launch new Chrome
Launch Chrome, navigate to 'https://statistics.securelist.com/kaspersky-anti-spam/day' and store the instance into Browser

14 Wait for web page content
Wait for UI element Anchor 'XLS XLS' 2 to appear on web page

15 Click link on web page
Click on Anchor 'XLS XLS' 2 of web page

16 Focus window
Focus window Window 'Save As'

17 Populate text field in window
Populate text box Edit 'File name:' with 'C:\Users\fahim\OneDrive - fahimsufi.com\Cyber\Spam.xlsx'

18 Click UI element in window
Click on UI element Button 'Save'

0 Selected actions 74 Actions 1 Subflow Run delay 100 ms

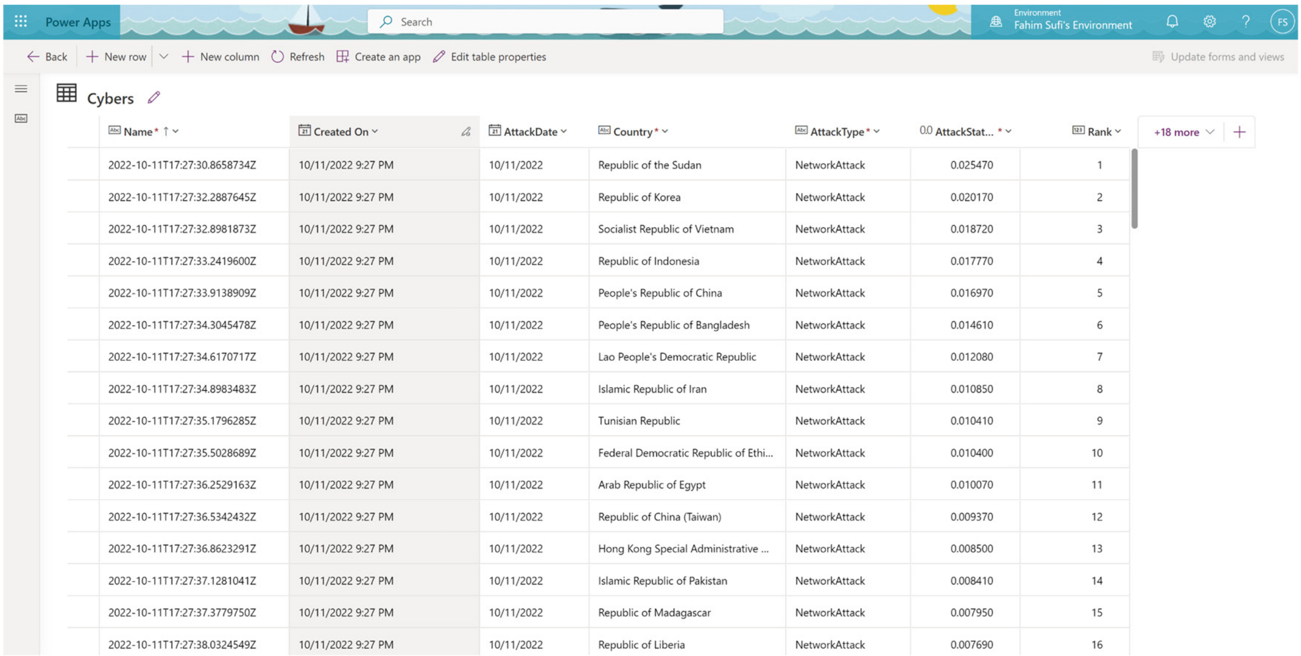
Figure 3. Microsoft Power Automate Desktop running in the remote machine and executing a series of RPA tasks.

2.1.3. Data Storage and Integration

Upon the deposition of files in the OneDrive folder, another Microsoft Power Automate Cloud Flow becomes activated. This flow processes the files and stores the data permanently in Microsoft Dataverse [34], ensuring secure and organized data management. Figure 4 shows several of these cyber-attack statistics by date, country, attack type, percentage of attack, and ranking of cyber-attack. It should be noted that the time-series compilation of cyber-attack statistics was possible because of the methodological and well-orchestrated employment of automations tools like Microsoft Power Automate Cloud Flow, Microsoft Power Automate Desktop Flow, Microsoft One Drive, and Microsoft Dataverse. While Figure 3 represents the disposition of files containing single attack type data for a single day, Figure 4 presents aggregated time-series data across multiple days, and multiple attack types. As clearly seen from Figure 4, the rank column provides rankings based on the AttackStatistics (i.e., percentage of attack) for a particular day on a particular attack type.

2.1.4. Data Manipulation and Analysis

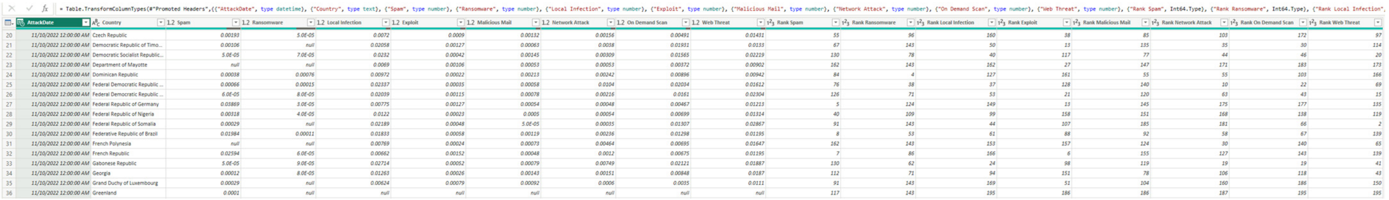
The final step involved using Microsoft Power BI [35], equipped with Data Analysis Expressions (DAX) and Power Query. This tool was instrumental in manipulating the data, performing AI-based analysis, and visualizing the time-series cyber data. The final dataset comprised 77,623 rows and 18 columns.



Name* ↑	Created On	AttackDate	Country*	AttackType*	AttackStat...*	Rank
2022-10-11T17:27:30.8658734Z	10/11/2022 9:27 PM	10/11/2022	Republic of the Sudan	NetworkAttack	0.025470	1
2022-10-11T17:27:32.2887645Z	10/11/2022 9:27 PM	10/11/2022	Republic of Korea	NetworkAttack	0.020170	2
2022-10-11T17:27:32.8981873Z	10/11/2022 9:27 PM	10/11/2022	Socialist Republic of Vietnam	NetworkAttack	0.018720	3
2022-10-11T17:27:33.2419600Z	10/11/2022 9:27 PM	10/11/2022	Republic of Indonesia	NetworkAttack	0.017770	4
2022-10-11T17:27:33.9138909Z	10/11/2022 9:27 PM	10/11/2022	People's Republic of China	NetworkAttack	0.016970	5
2022-10-11T17:27:34.3045478Z	10/11/2022 9:27 PM	10/11/2022	People's Republic of Bangladesh	NetworkAttack	0.014610	6
2022-10-11T17:27:34.6170717Z	10/11/2022 9:27 PM	10/11/2022	Lao People's Democratic Republic	NetworkAttack	0.012080	7
2022-10-11T17:27:34.8983483Z	10/11/2022 9:27 PM	10/11/2022	Islamic Republic of Iran	NetworkAttack	0.010850	8
2022-10-11T17:27:35.1796285Z	10/11/2022 9:27 PM	10/11/2022	Tunisian Republic	NetworkAttack	0.010410	9
2022-10-11T17:27:35.5028689Z	10/11/2022 9:27 PM	10/11/2022	Federal Democratic Republic of Ethi...	NetworkAttack	0.010400	10
2022-10-11T17:27:36.2529163Z	10/11/2022 9:27 PM	10/11/2022	Arab Republic of Egypt	NetworkAttack	0.010070	11
2022-10-11T17:27:36.5342432Z	10/11/2022 9:27 PM	10/11/2022	Republic of China (Taiwan)	NetworkAttack	0.009370	12
2022-10-11T17:27:36.8623291Z	10/11/2022 9:27 PM	10/11/2022	Hong Kong Special Administrative ...	NetworkAttack	0.008500	13
2022-10-11T17:27:37.1281041Z	10/11/2022 9:27 PM	10/11/2022	Islamic Republic of Pakistan	NetworkAttack	0.008410	14
2022-10-11T17:27:37.3779750Z	10/11/2022 9:27 PM	10/11/2022	Republic of Madagascar	NetworkAttack	0.007950	15
2022-10-11T17:27:38.0324549Z	10/11/2022 9:27 PM	10/11/2022	Republic of Liberia	NetworkAttack	0.007690	16

Figure 4. Microsoft Dataverse managing cyber statistics by date, country, attack type and attack statistics (i.e., percentage of attack), and rank.

Within this step, DAX and Power Query transposes the attack-type (i.e., percentage values of the eight different attack types) and rank (ranking of the eight different attack types) columns. Thus, Figure 4 is turned into Figure 5 using Microsoft Power BI. As seen in Figure 5, the final data columns compose of the attack date and country of Figure 4 with the addition of eight attack percentages columns (i.e., spam, ransomware, local infection, exploit, malicious mail, network attack, on-demand scan, web threat etc.) and eight attack ranking columns (i.e., eighteen columns in total). The corresponding DAX code to perform this data transformation is provided in Appendix A.



```
= Table.TransformColumnTypes(#"Promoted Headers",{{"AttackDate", type datetime}, {"Country", type text}, {"Spam", type number}, {"Ransomware", type number}, {"Local Infection", type number}, {"Exploit", type number}, {"Malicious Mail", type number}, {"Network Attack", type number}, {"On Demand Scan", type number}, {"Web Threat", type number}, {"Rank Spam", Int64.Type}, {"Rank Ransomware", Int64.Type}, {"Rank Local Infection",
```

	AttackDate	Country	Spam	Ransomware	Local Infection	Exploit	Malicious Mail	Network Attack	On Demand Scan	Web Threat	Rank Spam	Rank Ransomware	Rank Local Infection	Rank Exploit	Rank Malicious Mail	Rank Network Attack	Rank On Demand Scan	Rank Web Threat
20	11/10/2022 12:00:00 AM	Czech Republic	0.00193	5.0E-05	0.0072	0.0009	0.00132	0.00156	0.00491	0.01431	55	96	160	38	85	103	172	97
21	11/10/2022 12:00:00 AM	Democratic Republic of Timo...	0.00106	null	0.02058	0.00127	0.00063	0.0038	0.01931	0.0133	67	143	50	13	135	35	30	114
22	11/10/2022 12:00:00 AM	Democratic Socialist Republic...	5.0E-05	7.0E-05	0.0232	0.00042	0.00145	0.00309	0.01565	0.02219	130	78	40	117	77	44	46	20
23	11/10/2022 12:00:00 AM	Department of Mayotte	null	null	0.0069	0.00106	0.00053	0.00053	0.00372	0.00902	162	143	162	27	147	171	183	173
24	11/10/2022 12:00:00 AM	Dominican Republic	0.00038	0.00076	0.00972	0.00022	0.00213	0.00242	0.00896	0.00942	84	4	127	161	55	55	103	166
25	11/10/2022 12:00:00 AM	Federal Democratic Republic ...	0.00066	0.00015	0.02337	0.00035	0.00058	0.0104	0.02034	0.01612	76	38	37	128	140	10	22	69
26	11/10/2022 12:00:00 AM	Federal Democratic Republic ...	6.0E-05	8.0E-05	0.02039	0.00115	0.00078	0.00216	0.0161	0.02304	126	71	53	21	120	63	43	15
27	11/10/2022 12:00:00 AM	Federal Republic of Germany	0.03869	3.0E-05	0.00775	0.00127	0.00054	0.00048	0.00467	0.01213	5	124	149	13	145	175	177	135
28	11/10/2022 12:00:00 AM	Federal Republic of Nigeria	0.00318	4.0E-05	0.0122	0.00023	0.0005	0.00054	0.00699	0.01314	40	109	99	158	151	168	138	119
29	11/10/2022 12:00:00 AM	Federal Republic of Somalia	0.00029	null	0.02189	0.00048	5.0E-05	0.00035	0.01307	0.02867	91	143	44	107	185	181	66	2
30	11/10/2022 12:00:00 AM	Federative Republic of Brazil	0.01984	0.00011	0.01833	0.00058	0.00119	0.00236	0.01298	0.01195	8	53	61	88	92	58	67	139
31	11/10/2022 12:00:00 AM	French Polynesia	null	null	0.00769	0.00024	0.00073	0.00464	0.00695	0.01647	162	143	153	157	124	30	140	65
32	11/10/2022 12:00:00 AM	French Republic	0.02594	6.0E-05	0.00662	0.00152	0.00048	0.0012	0.00675	0.01195	7	86	166	6	155	127	143	139
33	11/10/2022 12:00:00 AM	Gabonese Republic	5.0E-05	9.0E-05	0.02714	0.00052	0.00079	0.00749	0.02121	0.01887	130	62	24	98	119	19	19	41
34	11/10/2022 12:00:00 AM	Georgia	0.00012	8.0E-05	0.01263	0.00026	0.00143	0.00151	0.00848	0.0187	112	71	94	151	78	106	118	43
35	11/10/2022 12:00:00 AM	Grand Duchy of Luxembourg	0.00029	null	0.00624	0.00079	0.00092	0.0006	0.0035	0.0111	91	143	169	51	104	160	186	150
36	11/10/2022 12:00:00 AM	Greenland	0.0001	null	null	null	null	null	null	null	117	143	195	186	186	187	195	195

Figure 5. Microsoft Power BI transposes eight values of attack-type percentages and eight values of attack-type ranking to generate an eighteen-column global cyber-attack dataset (along with date and country columns).

2.2. Validation and Curation

- Automated process monitoring: The automated data collection process was closely monitored. It was noted that, for 17 days of the 14-month period, the process was interrupted due to essential updates required by the Azure Virtual Machine or Power Automate Desktop. These interruptions were promptly addressed to minimize data loss.
- Data integrity checks: Regular integrity checks were performed to ensure the accuracy and completeness of the data collected. This involved verifying the data against secondary sources and conducting random sample checks. It should be mentioned that the same methodology was implemented on another virtual machine for ensuring business continuity. In case of the unavailability of one virtual machine (due to maintenance or any other reasons), the other virtual machine would be automatically utilized.

2.3. Data Quality and Noise

- Noise reduction: Efforts were made to reduce noise and irrelevant information during the data collection phase. This was achieved through the careful configuration of the RPA scripts to selectively download relevant data.
- Quality assurance and data cleaning: The quality of the data was maintained through rigorous validation protocols. In the data manipulation phase, Power BI was used to clean the dataset. This involved removing duplicates, correcting inconsistencies, and handling missing values, thereby enhancing the quality and usability of the dataset. For example, an automated script replaced the null values (as seen in Figure 5) with 0.

2.4. Data Analysis and Forecasting

After validating the data quality, the data are fed through a three-step process as shown in Figure 6, beginning with correlation.

The correlation matrix, being the output of the correlation process, shows all of the cyber-parameters that are highly correlated. The values of the correlation coefficients range from −1 to 1, with the following insights:

- 1: A perfect positive correlation, meaning that, as one variable increases, the other variable also increases proportionally.
- 0: No correlation, indicating that there is no linear relationship between the variables.
- −1: A perfect negative correlation, signifying that, as one variable increases, the other variable decreases proportionally.

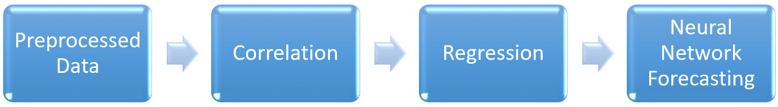


Figure 6. Method of data analysis and forecasting.

The correlation matrix is a symmetric matrix, as the correlation between Variable A and Variable B is the same as the correlation between Variable B and Variable A. For a set of variables $\{X_1, X_2, \ldots, X_n\}$, the correlation matrix might look like Equation (1).

$$Correlation\ Matrix = \begin{vmatrix} 1 & \cdots & r_{1n} \\ \ldots & \ddots & \ldots \\ r_{n1} & \cdots & 1 \end{vmatrix} \tag{1}$$

where r_{ij} represents the correlation coefficient between X_i and X_j. The diagonal elements (e.g., $r_{11}, r_{22}, \ldots, r_{nn}$) are always 1 because a variable has a perfect correlation with itself. Correlation matrices are commonly used in statistics, finance, and other fields to explore relationships between variables, identify patterns, and understand the interdependencies within a dataset.

As seen from Figure 6, once the correlation step is complete, regression is performed on the cyber-threat data. A linear regression identifies the linear relationship between the dependent and independent variables in the form of Equation (2).

$$y = b_0 + b_1 x_1 + \varepsilon \tag{2}$$

The process ends with the forecasting of related parameters using a neural network. Equation (3) shows an autoregressive model in the order of p, where $h(.)$ is a nonlinear function. Here, ε_t is a sequence of random independent parameters with the distribution of zero mean and finite variance σ^2. An autoregressive neural network is shown in Equations (4) and (5) as a feed-forward network. The $f(.)$ function is an activation function with a parameter vector of $\theta = (\beta_0, \beta_1, \ldots, \beta_q, \alpha_1, \ldots, \alpha_q, \omega_{11}, \ldots, \omega_{qn})$.

$$y_t = h(y_{t-1}, \ldots, y_{t-p}) + \varepsilon_t \tag{3}$$

$$\hat{y}_t = \hat{h}(y_{t-1}, \ldots, y_{t-p}) \tag{4}$$

$$\hat{y}_t = \beta_0 + \sum_{i=1}^{I} \beta_j f(\alpha_i + \sum_{j=1}^{p} w_{ij} y_{t-j}) \tag{5}$$

Equation (6), on the other hand, shows a generalized linear model in a nonlinear case, with $h(.)$ being a nonlinear function and ε_t representing the same contextual information portrayed in Equation (5). As shown in Equations (7) and (8), $\hat{y}_t$ is calculated with recursive estimation, because of the non-observational nature of $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$.

$$y_t = h(y_{t-1}, \ldots, y_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}) + \varepsilon_t \tag{6}$$

$$\hat{y}_t = h(y_{t-1}, \ldots, y_{t-p}, \hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-q}) \tag{7}$$

$$\hat{\varepsilon}_t = y_{t-1} - \hat{y}_j,\ j = t-1, \ldots t-q \tag{8}$$

The recurrent neural network model can now be expressed with a recurrent network, as defined in Equations (9) and (10).

$$\hat{y}_t = a_0 + \sum\nolimits_{j=1}^{h} a_j g(\beta_{0j}) + \sum\nolimits_{i=1}^{p} \beta_{ij} y_{t-i} + \sum\nolimits_{i=p+1}^{p+q} \beta_{ij} \hat{\varepsilon}_{t+p-i} \tag{9}$$

$$\hat{\varepsilon}_{t+p-i} = y_{t+p-i} - \hat{Y}_{t+p-i} \tag{10}$$

In conclusion, the methodological framework applied in this study harnesses the power of modern low-code tools, AI-driven automation, and AI-driven analytics. As summarized in Table 2 and shown in Figure 1, Microsoft Power Automate Desktop, Microsoft Power Automate Cloud Flow, Microsoft One Drive, Microsoft Dataverse, and Microsoft Power BI were used to automate the cyber-threat data acquisition process. Correlation, regression and a neural network were implemented with Python (shown in Appendix A). Finally, the entire solution was deployed and tested in multiple platforms and devices covering Windows, iOS, and Android. This approach not only streamlined the data collection process but also ensured the high quality and reliability of the dataset, making it a valuable resource for cybersecurity research.

Table 2. The use of various technology components for fulfilling functional requirements (● = Supported).

Technology Component	**Automate Data Acquisition**	**Correlation**	**Regression**	**Neural-Network-Based Forecasting**	**Deployment in iOS, Android, and Windows Devices**
Microsoft Power Automate Desktop	●				
Microsoft Power Automate Cloud Flow	●				
Microsoft One Drive	●				
Microsoft Dataverse	●				
Microsoft Power BI Desktop		●	●	●	●
Python on Power BI		●	●	●	●
Microsoft Power BI Service					●

3. Results

The dataset, encompassing a broad spectrum of cyber-attack dimensions, consists of 77,623 rows and 18 fields, as shown in Table 3. Table 3 also shows sample data (row 37,349 of the Supplementary Materials). Each field's statistical distribution, along with 25th, 50th, and 75th percentiles are detailed in Tables 4 and 5, providing a comprehensive overview of the global cyber threat landscape.

- Attack date (date): Records the date of each cyber-attack. All 77,623 entries are valid, indicating a complete dataset with no missing dates (as shown in Table 3).
- Country (text): Specifies the country of the cyber-attack occurrence. All entries are valid, covering 225 countries, and provide a global perspective on cyber threats (as shown in Table 3).
- Spam (decimal): Records the percentage of spam-related attacks for a specific country on a particular day (percentage value in decimal). The sample data, as shown in Table 3, for spam is 0.01358, meaning that Bangladesh suffered 1.358% of spam attacks on that day. With 62,982 valid entries, the mean is 0.006094, and the standard deviation is 0.024337. The range is from a minimum of 0.000010 to a maximum of 0.302490. The 25th, 50th, and 75th percentiles are 0.000090, 0.000590, and 0.003530, respectively (as shown in Tables 3 and 4).
- Ransomware (decimal): Records the percentage of ransomware related attacks for a specific country on a particular day. This consists of 52,144 valid entries (Table 3). As per Table 4, the average value is 0.000130, with a standard deviation of 0.000186. The range spans from 0.000010 to 0.009180. The percentiles are 0.000040 (25th), 0.000070 (50th), and 0.000140 (75th).
- Local infection (decimal): Records the percentage of local infections for a specific country on a particular day. With 74,469 entries, the mean is 0.013350, and the standard deviation is 0.008415. The values range from 0.000240 to 0.049370. The percentiles are 0.007150 (25th), 0.010790 (50th), and 0.017660 (75th) (detailed in Tables 3 and 4).
- Exploit (decimal): Records the percentage of exploit-related attacks for a specific country on a particular day. Contains 64,264 valid entries. The mean is 0.000469, with a standard deviation of 0.000368. The range is from 0.000010 to 0.004660. The 25th, 50th, and 75th percentiles are 0.000210, 0.000390, and 0.000620, respectively (shown in Tables 3 and 4).
- Malicious mail (decimal): Records the percentage of malicious-mail-related attacks for a specific country on a particular day. Comprises 69,184 entries. The mean is 0.001292, with a standard deviation of 0.001606. The range spans from 0.000010 to 0.043220. The percentiles are 0.000300 (25th), 0.000730 (50th), and 0.001690 (75th) (Tables 3 and 4).
- Network attack (decimal): Records the percentage of network-attack-related incidents for a specific country on a particular day. With 71,532 entries, the average value is 0.002222, and the standard deviation is 0.003034. The range is from 0.000020 to 0.058260. The 25th, 50th, and 75th percentiles are 0.000700, 0.001290, and 0.002350, respectively (shown in Tables 3 and 4).
- On-demand scan (decimal): Records the percentage of on-demand scans (by the antivirus program) for a specific country on a particular day. This contains 74,231 valid entries. The mean is 0.009756, with a standard deviation of 0.006080. The minimum and maximum values are 0.000240 and 0.124880, respectively. The percentiles are 0.005210 (25th), 0.008140 (50th), and 0.012870 (75th) (detailed in Tables 3 and 4).
- Web threat (decimal): Records the percentage of web-threat-related attacks for a specific country on a particular day. This comprises 73,892 entries. The mean value is 0.013006, with a standard deviation of 0.004943. The range spans from 0.000240 to 0.048630. The 25th, 50th, and 75th percentiles are 0.009700, 0.012570, and 0.015973, respectively (shown in Tables 3 and 4).
- 11–18. Rank fields (text): These fields (rank spam, rank ransomware, rank local infection, rank exploit, rank malicious mail, rank network attack, rank on-demand scan, and rank web threat) provide the world ranking of each country in the respective cyber-attack dimension. The rankings range from 1 to a maximum of 196, with the mean rankings varying slightly across different attack types. The standard deviation

in these rankings indicates a moderate spread, reflecting the varying impact of cyber threats across different countries (as detailed in Tables 3 and 4).

In summary, the dataset offers a detailed and comprehensive view of cyber threats across multiple dimensions. The statistical analysis reveals the varying intensity and frequency of different types of cyber-attacks, providing valuable insights for researchers and policymakers. The complete and valid entries for the attack date and country fields ensure a robust temporal and geographical representation of the data. Figure 7 shows the new dataset driving a cyber-intelligence solution within an Android deployment on the latest Samsung Galaxy S23 Mobile. As seen from Figure 7, a particular time period (i.e., from 1 November 2022 to 8 March 2023) was selected for Australia. As seen from the spider chart on the mobile environment, Australia suffers most critically from web threats (1st), local infections (2nd), on-demand scan (3rd), spam (4th) and then the other cyber-attack dimensions. Any other countries within the globe can be selected by the user, and the mobile-based solution (available in iOS, Android, and Windows) could empower the cyber analyst with interactive insights on threat dimensions, threat ranking, threat patterns, and event forecasting. Unlike existing research works that have used network traffic data (that is applicable to a single institute or organization) [36–38], the global cyber-attack data generated in this research facilitates comparative analytics among multiple countries. Therefore, national leaders, cyber strategists, and policymakers can use an intelligent system like that shown in Figure 7 on their mobile devices, facilitating just-in-time decisions on national cyber postures.

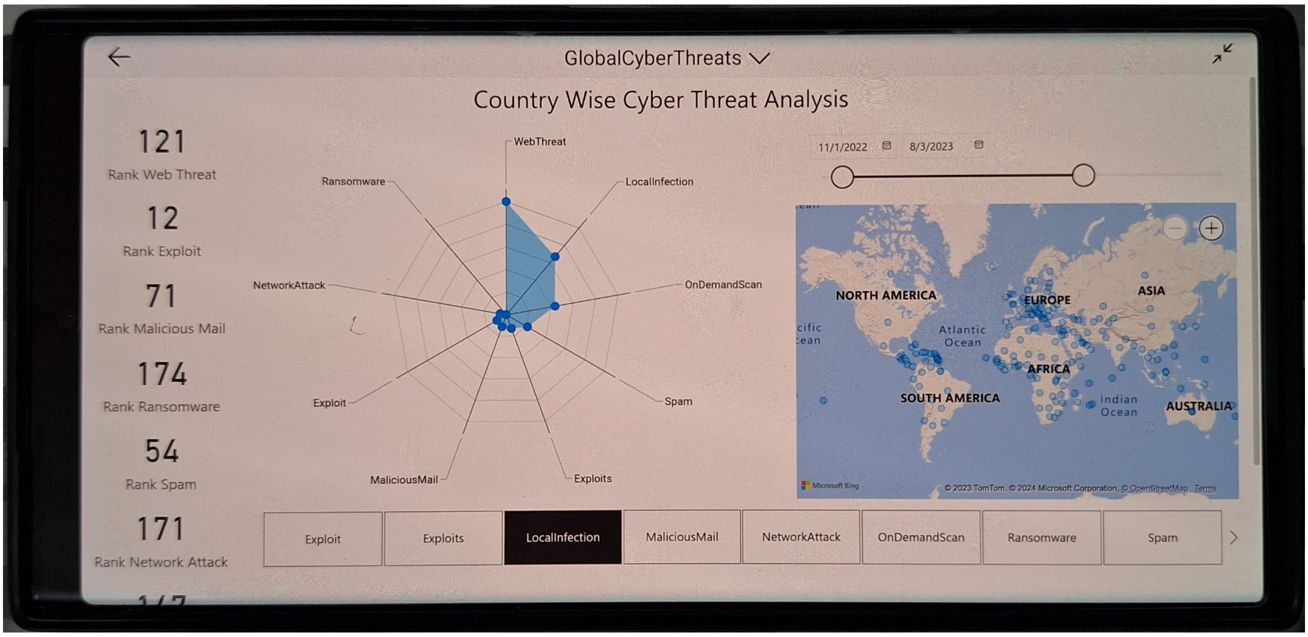


Figure 7. Cyber intelligence deployed in Android Version 14 within a Samsung Galaxy S23 Ultra Mobile (Australia is selected).

Table 3. Data types and descriptions for all of the 18 fields, with sample data (# means number).

Field Name	Data Type	Valid #	Valid %	Error #	Error %	Empty #	Empty %	Sample Data
Attack date	Date	77,623	100%	0	0%	0	0%	25 April 2023
Country	Text	77,623	100%	0	0%	0	0%	Bangladesh
Spam	Decimal	62,982	81%	0	0%	14,641	19%	0.01358
ransomware	Decimal	52,144	67%	0	0%	25,479	33%	0.00025
Local infection	Decimal	74,469	96%	0	0%	3154	4%	0.00726
Exploit	Decimal	64,264	83%	0	0%	13,359	17%	0.00021

Table 3. *Cont.*

Field Name	Data Type	Valid #	Valid %	Error #	Error %	Empty #	Empty %	Sample Data
Malicious mail	Decimal	69,184	89%	0	0%	8439	11%	0.00019
Network attack	Decimal	71,532	92%	0	0%	6091	8%	0.00076
On-demand scan	Decimal	74,231	96%	0	0%	3392	4%	0.00445
Web threat	Decimal	73,892	95%	0	0%	3731	5%	0.01116
Rank spam	Text	77,623	100%	0	0%	0	0%	34
Rank ransomware	Text	77,623	100%	0	0%	0	0%	6
Rank local infection	Text	77,623	100%	0	0%	0	0%	100
Rank exploit	Text	77,623	100%	0	0%	0	0%	108
Rank malicious mail	Text	77,623	100%	0	0%	0	0%	56
Rank network attack	Text	77,623	100%	0	0%	0	0%	94
Rank on-demand scan	Text	77,623	100%	0	0%	0	0%	109
Rank web threat	Text	77,623	100%	0	0%	0	0%	109

Table 4. Data distributions and statistics of for the eight cyber-attack percentage (i.e., decimal) fields.

Index	Spam	Ransomware	Local Infection	Exploit	Malicious Mail	Network Attack	On-Demand Scan	Web Threat
min	0.00001	0.00001	0.00024	0.00001	0.00001	0.00002	0.00024	0.00024
25%	0.00009	0.00004	0.00715	0.00021	0.0003	0.0007	0.00521	0.0097
50%	0.00059	0.00007	0.01079	0.00039	0.00073	0.00129	0.00814	0.01257
75%	0.00353	0.00014	0.01766	0.00062	0.00169	0.00235	0.01287	0.015973
max	0.30249	0.00918	0.04937	0.00466	0.04322	0.05826	0.12488	0.04863
mean	0.006094	0.00013	0.01335	0.000469	0.001292	0.002222	0.009756	0.013006
std	0.024337	0.000186	0.008415	0.000368	0.001606	0.003034	0.00608	0.004943
count	62,982	52,144	74,469	64,264	69,184	71,532	74,231	73,892

Table 5. Data distributions and statistics of for the eight cyber-attack rankings (i.e., Text) fields.

Index	Rank Spam	Rank Ransomware	Rank Local Infection	Rank Exploit	Rank Malicious Mail	Rank Network Attack	Rank On-Demand Scan	Rank Web Threat
min	1	1	1	1	1	1	1	1
25%	48	41	47	38	48	48	47	46
50%	98	88	98	90	98	98	97	97
75%	144	131	148	144	147	148	148	148
max	186	158	196	188	194	192	196	196
mean	94.53299	84.00165	97.73	90.46	96.80494	97.76224	97.42858	96.97435
std	52.67272	48.97654	58.05	58.98	55.90629	56.90058	58.17745	58.45481
count	77,623	77,623	77,623	77,623	77,623	77,623	77,623	77,623

Figure 8 depicts the correlations among the eight different types of cyber-attack dimensions (i.e., exploit, local infection, malicious mail, network attack, on-demand scan, ransomware, spam and web threat). As previously shown in Equation (1), the correlation coefficient, denoted as "*r*", quantifies the strength and direction of a linear relationship. As seen in Figure 8, there is a correlation coefficient value of 0.93 between on-demand scan and local infection, which indicates a very strong positive linear relationship between these variables. As these variables (on-demand scan and local infection) are linearly related, linear regression (i.e., as shown previously in Equation (2)) can be used to quantify the relationship. As shown previously in Equation (2), given a value of X, the model predicts the corresponding value of Y based on the regression equation (as seen from Figure 9). The linear regression model can be used for prediction.

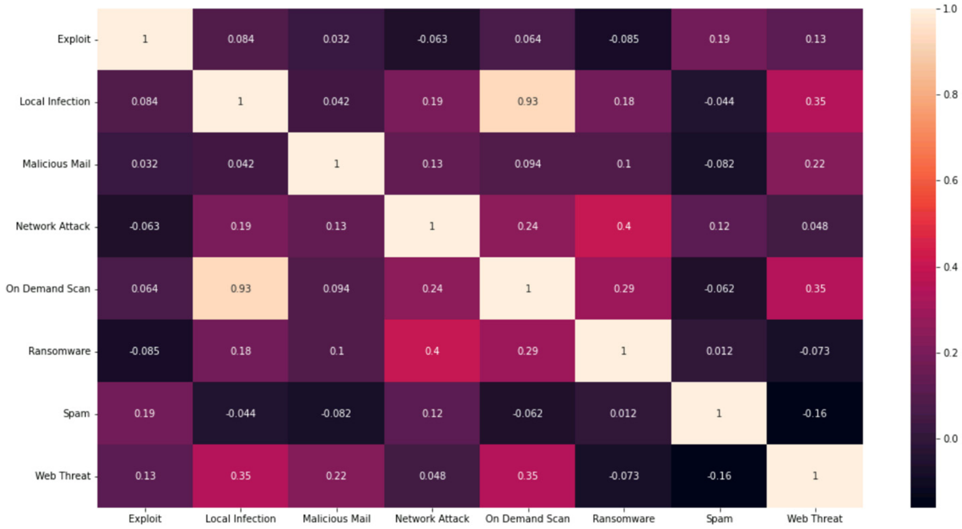


Figure 8. Correlation of cyber parameters, showing on-demand scan highly correlated with local infection (with a correlation coefficient of 0.93).

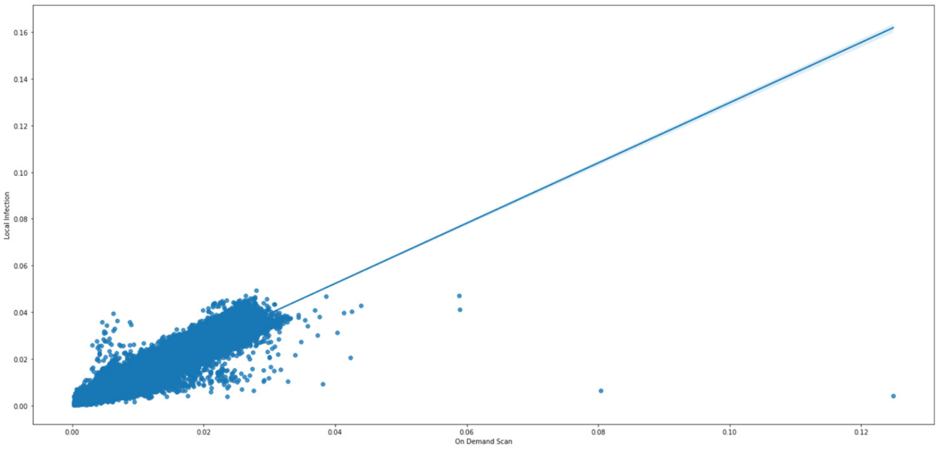


Figure 9. Linear regression between on-demand scan and local infection.

Finally, Figure 10 shows the prediction of both on-demand scan and local infection for Australia. As seen from Figure 10, the trends, patterns, seasonality for both these cases appear to be similar, since earlier experimentation found that both these factors (i.e., on-demand scan and local infection) are highly correlated, with a correlation coefficient of 0.93. Though Figure 10 only showcases predictions for on-demand scan and local infection (as these were found to be correlated), predictions could be generated seamlessly for all of the eight cyber-attack dimensions, for a more comprehensive understanding of future threats. For predicting time-series data there are some popular methods, like exponential smoothing [39], autoregressive integrated moving average (ARIMA) [40], and neural networks [40,41]. In our previous research, an exponential smoothing algorithm was used to predict cyber-attacks from time-series data [13–15] using Microsoft Power BI implementation. However, in this study neural-network-based forecasting was implemented with Python code within a Microsoft Power BI environment. This could also be executed using the visuals produced by MAQ Software, Redmond, WA, USA, following the implementation guidelines of [42]. Appendix A showcases the python code that was executed within a Microsoft Power BI environment. This code predicted cyber-attacks for 10 days with an 80% confidence level.

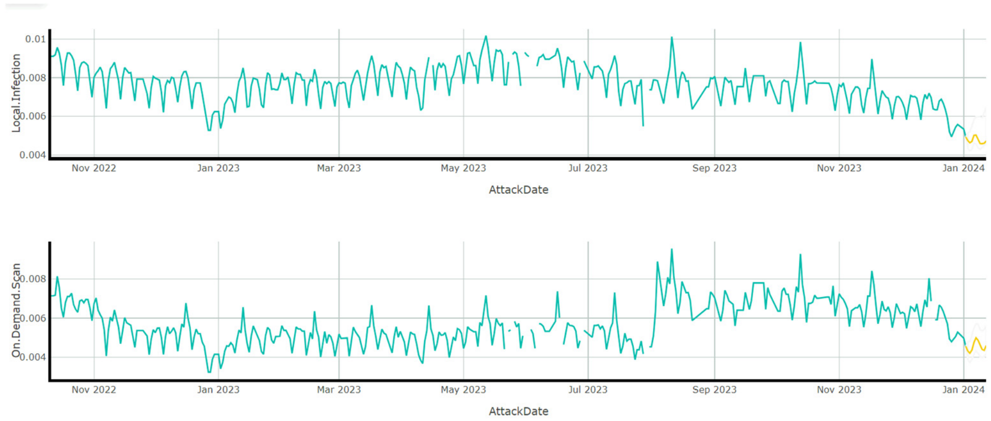


Figure 10. Neural-network-based cyber-threat prediction for local infection and on-demand scan (The vertical grey line denote current date and subsequently the yellow lines refer to the predicted values).

Although the aggregation process for this dataset is intricate, utilizing various features of modern AI-driven tools, the study's resulting data are stored in a simple and easy-to-use CSV file format (publicly accessible at https://github.com/DrSufi/CyberData, accessed on 14 February 2024).

4. Discussion

While there are several cyber intelligence solutions [43–47] that portray global cyber threat statistics, none of these solutions provide access to time-series cyber statistics data. This research uses an innovative AI-driven RPA methodology to acquire, manage, and store global cyber threat statistics. The presented methodology of data aggregation is fully automated and much more robust in comparison with existing systems [48–52]. The comprehensive cyber data deriving from this study have been made available to researchers and academics for the first time in order to facilitate open-source cyber research. This dataset could be used to identify correlations [20,21], for regression [17–19], and for prediction. The dataset, meticulously compiled from various open-source statistics, and employing modern AI techniques for data treatment and analysis, includes sensitive dimensions of cyber-attacks but does not contain individual or personally identifiable information. This is because Kaspersky [4] tracks the cyber-attack data at country level (i.e., not a personal level).

4.1. Use of This Dataset

This section provides essential guidance for researchers and analysts intending to utilize the dataset. The section aims to facilitate efficient and effective use of the data, ensuring that users can quickly familiarize themselves with its structure and nuances.

- Date format and chronological analysis: The 'Attack-Date' field is crucial for any time-series analysis. Users should ensure that their analytical tools correctly interpret the date format for accurate temporal analysis.
- Handling missing values: Fields such as spam, ransomware, and others contain empty entries, representing days with no recorded activities. Users should decide on their approach to handling these missing values based on their research objectives, i.e., whether to treat them as zeros, ignore them, or use imputation techniques.
- Statistical analysis: The dataset provides rich ground for statistical analysis, with fields showing varying degrees of standard deviation and range. Users should consider appropriate statistical methods to analyze these fields, keeping in mind the distribution characteristics like mean, median, percentiles, and range.
- Comparative and ranking analysis: The rank fields offer a unique opportunity for comparative analysis across countries and cyber-attack dimensions. Researchers interested in global or regional comparisons will find these fields particularly useful.

- Data visualization: Given the dataset's complexity and breadth, effective visualization tools (like Microsoft Power BI used in its creation) are recommended for a more intuitive understanding of trends and patterns.
- Cross-referencing with external data: Users are encouraged to augment this dataset with external data sources for more comprehensive analysis (e.g., [15]). This could include socio-economic data, digital infrastructure metrics, or other relevant datasets.
- Ethical considerations: While the dataset does not contain personal data, users should still adhere to ethical guidelines in their analysis, particularly when publishing results or drawing conclusions that might influence public policy or perception.
- Updates and version control: Users should note that the dataset covers up to December 2023. Any developments in the cyber threat landscape post this period will not be reflected. Keeping track of dataset versions and updates is crucial for longitudinal studies.
- Collaboration and sharing findings: Users are encouraged to share their findings with the broader research community. Collaboration can lead to more robust insights and a deeper understanding of the global cyber threat landscape.

4.2. Performance Evaluation

As seen from Table 6, the performance of the forecasting algorithm using a neural network was computed for several countries (e.g., Australia, New Zealand, Saudi Arabia, UAE, Qatar). It should be mentioned that the presented system empowered with the presented time-series would support cyber-threat forecasting for any countries on any of the threat dimensions. However, for the performance evaluations, few countries were randomly selected. The justification for choosing local infection and on-demand scan is that, in our earlier experimentation, we observed that these two dimensions were correlated. For performance metrics, we used mean square error (MSE) and mean absolute percent error (MAPE), as shown in Table 6. With A_t being actual threat and F_t being the forecasted threat for time t, MSE and MAPE could be represented with Equations (11) and (12).

$$MSE = \frac{1}{n}\sum_{t}^{n}(A_t - F_t)^2 \tag{11}$$

$$MAPE = \frac{100}{n}\sum\nolimits_{t=1}^{n}\frac{|A_t - F_t|}{\overline{X}} \tag{12}$$

Table 6. Performance evaluation of the neural-network-based cyber-threat prediction.

Country	Attack Type	MSE	MAPE (%)
Australia	Local infection	1.45	79.2%
Australia	On-demand scan	1.85	88.4%
New Zealand	Local infection	1.72	83.1%
New Zealand	On-demand scan	1.55	81.3%
Saudi Arabia	Local infection	1.32	71.7%
Saudi Arabia	On-demand scan	1.39	72%
UAE	Local infection	1.68	80.3%
UAE	On-demand scan	1.71	82.9%
Qatar	Local infection	1.66	78.5%
Qatar	On-demand scan	1.73	83.9%

As seen from Table 6, the mean of MSE was found to be 1.616 ± 0.2446 (where, 1.616 was the mean, μ and 0.2446 was the standard deviation, σ of the MSE). Similarly, MAPE was found to be 80.13 ± 4.9742%. The highest value of MSE/MAPE was recorded for Australia on on-demand scan and the lowest value of the MSE/MAPE was identified for Saudi Arabia's local infection. Hence, the accuracy of predictions showcased with this considerably extended dataset markedly surpasses that of existing studies [22–29].

4.3. Limitations of This Study

In order to obtain evidence-based policy and strategic decision-making on cyber related issues, the decision-makers use global cyber threat maps like those in [43–47]. Ideally, in order to take evidence-based policy decisions, decision makers should cross reference with other data sources. As of the time of writing this paper, none of these cyber intelligence solutions provide access to historical cyber-attack data (i.e., time-series) at a national level. Hence, this study used an innovative methodology to capture historical cyber data using daily cyber-threat statistics from Kaspersky [45]. Due to data inaccessibility and unavailability, this study could not cross reference the time-series dataset generated in this study with other datasets, such as those in [43,44,46,47]. That is why it is imperative that researchers and policymakers should cross-reference this dataset with other available sources (as previously suggested in Section 4.1).

Another limitation of this study is that it does not use taxonomies for the generation of a comprehensive understanding of the data. When cyber-attacks are targeted at individual agencies or institutions, researchers need to comprehend the attack through the lenses of taxonomies or frameworks [53]. Taxonomies, like those of [53], would allow a researcher to comprehend multiple dimensions like attack target, operational impact, informational impact etc. However, when comparing cyber-attacks at the country or national level, operational impacts and informational impacts, along with other dimensions of the taxonomy, become aggregated. As a result, a taxonomy, like that of [53], has limited value for comparing country level cyber-threats. Hence, almost all of the modern country-level threat intelligence providers (e.g., [43–47]) portray country-level threats with respect to cyber-attack types (e.g., spam, network attack, web-threat, ransomware etc.).

5. Conclusions

This study introduces an innovative automated methodology for the systematic collection of global cyber-attack data, resulting in the creation of a comprehensive dataset spanning 225 countries over a 14-month period from October 2022 to December 2023. The dataset, encompassing 77,623 rows and 18 fields, intricately delineates cyber-attacks across eight critical dimensions: spam, ransomware, local infection, exploit, malicious mail, network attack, on-demand scan, and web threat. Additionally, the dataset incorporates ranking data, affording a nuanced comparative analysis of countries' susceptibilities to distinct cyber threats. The findings of the analysis illuminate discernible disparities in the frequency and intensity of cyber-attacks, both across different countries and within varying attack types. A correlation analysis underscores a noteworthy correlation (coefficient of 0.93) between on-demand scan and local infection. Subsequent linear regression between these correlated dimensions serves to elucidate their interdependency. An avant-garde application of neural-network-based forecasting is employed to predict the temporal patterns of on-demand scan and local infection. The predictive model demonstrates a notable precision, substantiated by mean squared error (MSE) and mean absolute percentage error (MAPE) values of 1.616 and 80.13, respectively.

Beyond the releasing of comprehensive cyber data for the first time (facilitating open-source cyber research), the practical ramifications of this research are manifold. This study offers actionable intelligence for cybersecurity strategies and strategic planning through the discernment of correlation and predictive patterns pertaining to specific cyber threat dimensions. As such, this study constitutes a substantial contribution to the evolving field of cybersecurity, providing valuable tools for risk assessment and informed decision-making in the face of the ever-evolving cyber landscape.

Future endeavors of this research direction would involve using more robust techniques to capture live cyber threat statistics from a diverse range of sources, like [43,44,46,47] the use of AI, computer vision, and RPA. This would create a more robust global cyber-attack dataset. After compiling a more robust global cyber-attack dataset, we envision the creation of a generic fit-for-purpose taxonomy in order to address the myriad dimensions of global cyber threats and use innovative transformer technology [54].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info15040199/s1, Dataset comprising of 77,623 rows and 18 fields in .csv format.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

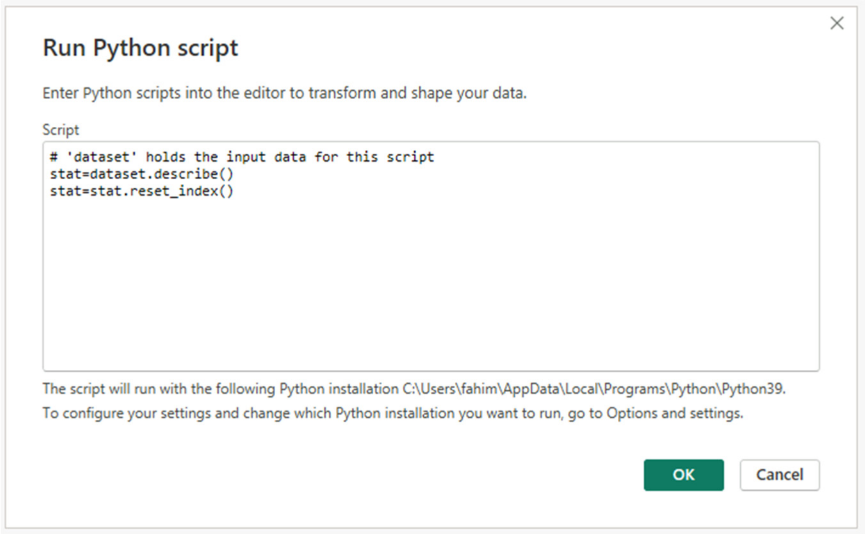
Informed Consent Statement: Not applicable.

Data Availability Statement: Complete dataset is made available.

Acknowledgments: The author would like to thank Edris Alam of Disaster and Emergency Crisis Management, Rabdan Academy, Abu Dhabi, UAE for evaluating the study and providing his valuable feedback on the deployed solution.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A



Run Python script

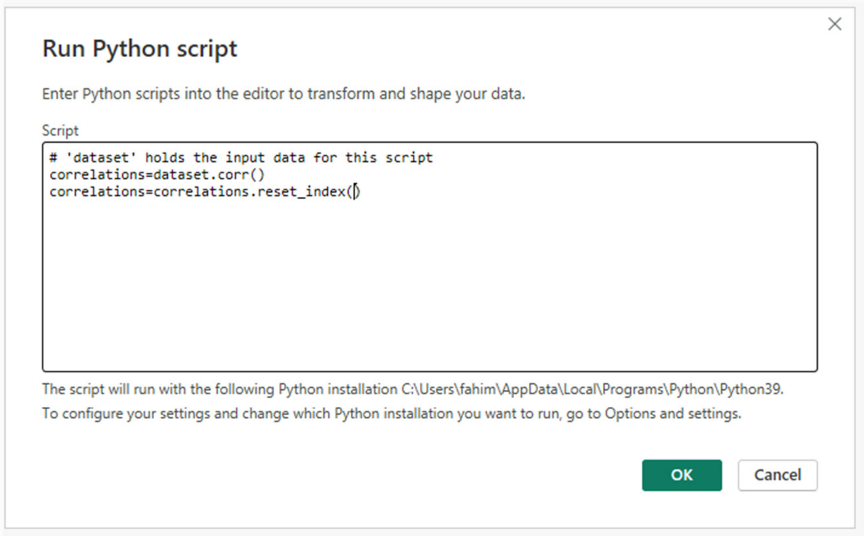
Enter Python scripts into the editor to transform and shape your data.

Script

```
# 'dataset' holds the input data for this script
stat=dataset.describe()
stat=stat.reset_index()
```

The script will run with the following Python installation C:\Users\fahim\AppData\Local\Programs\Python\Python39.
To configure your settings and change which Python installation you want to run, go to Options and settings.

Figure A1. Obtaining data statistics using Python in Power BI.



Run Python script

Enter Python scripts into the editor to transform and shape your data.

Script

```
# 'dataset' holds the input data for this script
correlations=dataset.corr()
correlations=correlations.reset_index()
```

The script will run with the following Python installation C:\Users\fahim\AppData\Local\Programs\Python\Python39.
To configure your settings and change which Python installation you want to run, go to Options and settings.

Figure A2. Running correlation using Python within Microsoft Power BI.

Algorithm A1 DAX Code for transposing the AttackType Column in Power BI

```
TransposedTable =
ADDCOLUMNS (
    SUMMARIZE ( DataverseCyberTable, DataverseCyberTable[Country], DataverseCyberTable[AttackDate] ),
     “Spam Percentage”, CALCULATE ( MAX ( DataverseCyberTable[AttackPercentage] ), DataverseCyberTable[AttackType]
     = “Spam” ),
     “Ransomware Percentage”, CALCULATE ( MAX ( DataverseCyberTable[AttackPercentage] ), DataverseCyberTable[Attack
     Type] = “Ransomware” ),
     “Local Infection Percentage”, CALCULATE ( MAX ( DataverseCyberTable[AttackPercentage] ), DataverseCyberTable[Attack
     Type] = “Local Infection” ),
     “Exploit Percentage”, CALCULATE ( MAX ( DataverseCyberTable[AttackPercentage] ), DataverseCyberTable[AttackType] =
     “Exploit” ),
     “Malicious Mail Percentage”, CALCULATE ( MAX ( DataverseCyberTable[AttackPercentage] ), DataverseCyberTable[Attack
     Type] = “Malicious Mail” ),
     “Network Attack Percentage”, CALCULATE ( MAX ( DataverseCyberTable[AttackPercentage] ),
     DataverseCyberTable[AttackType] = “Network Attack” ),
     “On Demand Scan Percentage”, CALCULATE ( MAX ( DataverseCyberTable[AttackPercentage] ),
     DataverseCyberTable[AttackType] = “On Demand Scan” ),
     “Web Threat Percentage”, CALCULATE ( MAX ( DataverseCyberTable[AttackPercentage] ),
     DataverseCyberTable[AttackType] = “Web Threat” ),
     “Rank Spam”, CALCULATE ( MAX ( DataverseCyberTable[Rank] ), DataverseCyberTable[AttackType] = “Spam” ),
     “Rank Ransomware”, CALCULATE ( MAX ( DataverseCyberTable[Rank] ), DataverseCyberTable[AttackType] =
     “Ransomware” ),
     “Rank Local Infection”, CALCULATE ( MAX ( DataverseCyberTable[Rank] ), DataverseCyberTable[AttackType] =
     “Local Infection” ),
     “Rank Exploit”, CALCULATE ( MAX ( DataverseCyberTable[Rank] ), DataverseCyberTable[AttackType] = “Exploit” ),
     “Rank Malicious Mail”, CALCULATE ( MAX ( DataverseCyberTable[Rank] ), DataverseCyberTable[AttackType] =
     “Malicious Mail” ),
     “Rank Network Attack”, CALCULATE ( MAX ( DataverseCyberTable[Rank] ), DataverseCyberTable[AttackType] =
     “Network Attack” ),
     “Rank On Demand Scan”, CALCULATE ( MAX ( DataverseCyberTable[Rank] ), DataverseCyberTable[AttackType] = “On
     Demand Scan” ),
     “Rank Web Threat”, CALCULATE ( MAX ( DataverseCyberTable[Rank] ), DataverseCyberTable[AttackType] = “Web
     Threat” )
)
```

Algorithm A2 Python code for linear regression

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
# Assuming your dataset is stored in a variable named ‘data’
# Extracting relevant features for linear regression
features = data[[‘local_infection_percentage’, ‘on_demand_scan_percentage’]]
# Assuming ‘target_variable’ is the dependent variable you want to predict
target_variable = data[‘target_variable’]
# Splitting the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(features, target_variable, test_size = 0.2, random_state = 42)
# Initializing the Linear Regression model
linear_reg_model = LinearRegression()
# Training the model
linear_reg_model.fit(X_train, y_train)
# Making predictions on the test set
predictions = linear_reg_model.predict(X_test)
```

Algorithm A2 *Cont.*

```
# Calculating Mean Squared Error (MSE) for evaluation
mse = mean_squared_error(y_test, predictions)
print(f’Mean Squared Error (MSE): {mse}’)
# Displaying the coefficients of the linear regression model
coefficients = linear_reg_model.coef_
intercept = linear_reg_model.intercept_
print(f’Coefficients: {coefficients}’)
print(f’Intercept: {intercept}’)
```

Algorithm A3 Python Code for forecasting using a neural network (predicting cyber-attacks for 10 days with a 0.8 confidence level)

```
# Importing the requisite libraries
import pandas as pd
from sklearn.neural_network import MLPRegressor
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
# Reading the dataset into a Pandas DataFrame
data = pd.read_csv(“your_dataset.csv”) # Replace “your_dataset.csv” with the actual file path
# Filtering data for the illustrious country of Australia
australia_data = data[data[‘country’] == ‘Australia’]
# Extracting relevant features for forecasting
features = australia_data[[‘local_infection_percentage’, ‘on_demand_scan_percentage’]]
# Extracting the target variable
target = australia_data[‘date’] # Assuming ‘date’ is the target variable, please replace it accordingly
# Splitting the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size = 0.2, random_state = 42)
# Standardizing the features for optimal neural network performance
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
# Initializing the Neural Network model
model = MLPRegressor(hidden_layer_sizes = (100, 50), max_iter = 1000, random_state = 42)
# Training the model
model.fit(X_train_scaled, y_train)
# Forecasting for the next 10 days with a confidence level of 0.8
forecasted_dates = pd.date_range(start = data[‘date’].max(), periods = 10, freq = ‘D’)
forecasted_features = scaler.transform(your_forecasting_data) # Replace with your own forecasting data
forecasted_results = model.predict(forecasted_features)
# Displaying the comprehensive results
forecasted_data = pd.DataFrame({‘date’: forecasted_dates, ‘forecasted_result’: forecasted_results})
print(forecasted_data)
```

Algorithm A4 Python code for the evaluation of performance with MSE and MAPE

```
from sklearn.metrics import mean_squared_error
import numpy as np
# Assuming you have the actual values for the next 10 days in a variable named ‘actual_values’
actual_values = np.array([your_actual_values]) # Replace with your actual values
# Predicting the values for the next 10 days using the trained model
predicted_values = model.predict(forecasted_features)
# Calculating Mean Squared Error (MSE)
mse = mean_squared_error(actual_values, predicted_values)
print(f’Mean Squared Error (MSE): {mse}’)
# Calculating Mean Absolute Percentage Error (MAPE)
mape = np.mean(np.abs((actual_values—predicted_values)/actual_values)) * 100
print(f’Mean Absolute Percentage Error (MAPE): {mape}%’)
```

References

1. Cremer, F.; Sheehan, B.; Fortmann, M.; Kia, A.N.; Mullins, M.; Murphy, F.; Materne, S. Cyber risk and cybersecurity: A systematic review of data availability. *Geneva Pap. Risk Insur.-Issues Pract.* **2022**, *47*, 698–736. [CrossRef] [PubMed]
2. Cybercrime Magazine. Cybercrime to Cost the World $10.5 Trillion Annually by 2025. 2020. Available online: https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/ (accessed on 15 October 2022).
3. Bada, J.R.N.M. Chapter 4—The social and psychological impact of cyberattacks. In *Emerging Cyber Threats and Cognitive Vulnerabilities*; Academic Press: Cambridge, MA, USA, 2020; pp. 73–92.
4. Kaspersky. Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/ (accessed on 3 August 2023).
5. Kaspersky. Daily Spam Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/kaspersky-anti-spam/day (accessed on 11 November 2023).
6. Kaspersky. Daily Ransomware Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/ransomware/day (accessed on 11 November 2023).
7. Kaspersky. Daily Local Infections Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/on-access-scan/day (accessed on 3 August 2023).
8. Kaspersky. Daily Exploit Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/vulnerability-scan/day (accessed on 11 November 2023).
9. Kaspersky. Daily Mailicious Mail Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/mail-anti-virus/day (accessed on 3 August 2023).
10. Kaspersky. Daily Network Attack Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/intrusion-detection-scan/day (accessed on 3 August 2023).
11. Kaspersky. Daily On-Demand Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/on-demand-scan/day (accessed on 3 August 2023).
12. Kaspersky. Day Web Threat Cyber Threat Statistics. 2023. Available online: https://statistics.securelist.com/web-anti-virus/day (accessed on 11 November 2023).
13. Sufi, F. A global cyber-threat intelligence system with artificial intelligence and convolutional neural network. *Decis. Anal. J.* **2023**, *9*, 100364. [CrossRef]
14. Sufi, F. Novel Application of Open-Source Cyber Intelligence. *Electronics* **2023**, *12*, 3610. [CrossRef]
15. Sufi, F. A New AI-Based Semantic Cyber Intelligence Agent. *Future Internet* **2023**, *15*, 231. [CrossRef]
16. Sufi, F. Algorithms in Low-Code-No-Code for Research Applications: A Practical Review. *Algorithms* **2023**, *16*, 108. [CrossRef]
17. Lalou, M.; Kheddouci, H.; Hariri, S. Identifying the Cyber Attack Origin with Partial Observation: A Linear Regression Based Approach. In Proceedings of the IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), Tucson, AZ, USA, 18–22 September 2017.
18. Cai, F.; Ozdagli, A.; Koutsoukos, X. Variational Autoencoder for Classification and Regression for Out-of-Distribution Detection in Learning-Enabled Cyber-Physical Systems. *Appl. Artif. Intell.* **2022**, *36*, 2131056. [CrossRef]
19. Ghafouri, A.; Vorobeychik, Y.; Koutsoukos, X. Adversarial Regression for Detecting Attacks in Cyber-Physical Systems. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018.
20. Albasheer, H.; Siraj, M.M.; Mubarakali, A. Cyber-Attack Prediction Based on Network Intrusion Detection Systems for Alert Correlation Techniques: A Survey. *Sensors* **2022**, *22*, 1494. [CrossRef]
21. Pires, S.; Mascarenhas, C. Cyber Threat Analysis Using Pearson and Spearman Correlation via Exploratory Data Analysis. In Proceedings of the Third International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 5 May 2023.
22. Werner, G.; Yang, S.; McConky, K. Time Series Forecasting of Cyber A,ack Intensity. In Proceedings of the Cyber and Information Security Research (CISR) Conference, Oak Ridge, TN, USA, 4–6 April 2017.
23. Bakdash, J.Z.; Hutchinson, S.; Zaroukian, E.G.; Marusich, L.R.; Thirumuruganathan, S.; Sample, C.; Hoffman, B.; Das, G. Malware in the future? Forecasting of analyst detection of cyber events. *J. Cybersecur.* **2018**, *4*, tyy007. [CrossRef]
24. Husák, M.; Komárková, J.; Bou-Harb, E.; Celeda, P. Survey of Attack Projection, Prediction, and Forecasting in Cyber Security. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 640–660. [CrossRef]
25. Aflaki, A.; Gitizadeh, M.; Kantarci, B. Accuracy improvement of electrical load forecasting against new cyber-attack architectures. *Sustain. Cities Soc.* **2022**, *77*, 103523. [CrossRef]
26. Alrahmani, Z.A.; Elleithy, K. DDoS Attack Forecasting Using Transformers. In Proceedings of the IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Abu Dhabi, United Arab Emirates, 14 November 2023.
27. Choi, C.; Shin, S.; Shin, C. Performance evaluation method of cyber attack behaviour forecasting based on mitigation. In Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 20 October 2021.
28. Alrawi, O.; Ike, M.; Pruett, M.; Kasturi, R.P.; Barua, S.; Hirani, T.; Hill, B.; Saltaformaggio, B. Forecasting Malware Capabilities From Cyber Attack Memory Images. In Proceedings of the 30th Usenix Security Symposium, Vancouver, BC, Canada, 11–13 August 2021.

29. Qasaimeh, M.; Hammour, R.A.; Yassein, M.B.; Al-Qassas, R.S.; Torralbo, J.A.L.; Lizcano, D. Advanced security testing using a cyber-attack forecastingmodel: A case study of financial institutions. *J. Softw. Evol. Process* **2022**, *34*, e2489. [CrossRef]
30. Microsoft Power Automate Documentation. 2021. Available online: https://docs.microsoft.com/en-us/power-automate/ (accessed on 29 August 2021).
31. Serverless Notes. Use Wait for Image Action When Trying to Locate Objects. Available online: https://www.serverlessnotes.com/docs/wait-for-image-action-power-automate-desktop (accessed on 12 February 2024).
32. Microsoft Learn. Use AI Builder in Power Automate. 2023. Available online: https://learn.microsoft.com/en-us/power-automate/use-ai-builder (accessed on 12 February 2024).
33. Microsoft Learn. OCR Actions. 2022. Available online: https://learn.microsoft.com/en-us/power-automate/desktop-flows/actions-reference/ocr (accessed on 12 February 2024).
34. Microsoft Dataverse. 2022. Available online: https://powerplatform.microsoft.com/en-us/dataverse/ (accessed on 25 October 2022).
35. Microsoft. Microsoft Power BI Documentation. 2022. Available online: https://docs.microsoft.com/en-us/power-bi/ (accessed on 21 March 2022).
36. Xu, S.; Qian, Y.; Hu, R.Q. Data-Driven Network Intelligence for Anomaly Detection. *IEEE Netw.* **2019**, *33*, 88–95. [CrossRef]
37. Keshk, M.; Sitnikova, E.; Moustafa, N.; Hu, J.; Khalil, I. An Integrated Framework for Privacy-Preserving Based Anomaly Detection for Cyber-Physical Systems. *IEEE Trans. Sustain. Comput.* **2021**, *6*, 66–79. [CrossRef]
38. Shi, D.; Guo, Z.; Johansson, K.H.; Shi, L. Causality Countermeasures for Anomaly Detection in Cyber-Physical Systems. *IEEE Trans. Autom. Control* **2018**, *63*, 386–401. [CrossRef]
39. Bartolomei, S.M.; Sweet, A.L. A note on a comparison of exponential smoothing methods for forecasting seasonal series. *Int. J. Forecast.* **1989**, *5*, 111–116. [CrossRef]
40. Zhang, G. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **2003**, *50*, 159–175. [CrossRef]
41. Zhang, L.; Wang, R.; Li, Z.; Li, J.; Ge, Y.; Wa, S.; Huang, S.; Lv, C. Time-Series Neural Network: A High-Accuracy Time-Series Forecasting Method Based on Kernel Filter and Time Attention. *Information* **2023**, *14*, 500. [CrossRef]
42. Interactive Chaos. Forecast Using Neural Network by MAQ Software. Available online: https://interactivechaos.com/es/powerbi/visual/forecast-using-neural-network-maq-software (accessed on 11 November 2023).
43. Bitdefender. Bitdefencer Cyberthreat Real-Time Map. Available online: https://threatmap.bitdefender.com/ (accessed on 12 February 2024).
44. Fortinet. Fortinet Live Threatmap. Available online: https://threatmap.fortiguard.com/ (accessed on 12 February 2024).
45. Kaspersky. Cyber Threat Real-Time Map. Available online: https://cybermap.kaspersky.com/ (accessed on 12 February 2024).
46. Radware. Radware Live Threat Map. Available online: https://livethreatmap.radware.com/ (accessed on 12 February 2024).
47. Check Point. Check Point Live Cyber Threat Map. Available online: https://threatmap.checkpoint.com/ (accessed on 12 February 2024).
48. Kim, N.; Lee, S.; Cho, H.; Kim, B.-I.; Jun, M. Design of a Cyber Threat Information Collection System for Cyber Attack Correlation. In Proceedings of the International Conference on Platform Technology and Service (PlatCon), Jeju, Republic of Korea, 29–31 January 2018.
49. Maosa, H.; Ouazzane, K.; Sowinski-Mydlarz, V. Real-Time Cyber Analytics Data Collection Framework. *Int. J. Inf. Secur. Priv. (IJISP)* **2022**, *16*, 1–10. [CrossRef]
50. Milenkovic, D.Z. Cyber Security and Data Collection. *Secur. Sci. J.* **2023**, *4*, 102–118. [CrossRef]
51. Doenhoff, J.; Tamura, Y.; Tsuji, D.; Shigemoto, T. Data collection method for security digital twin on cyber physical systems. *IEICE Commun. Express* **2022**, *11*, 829–834. [CrossRef]
52. Koloveas, P.; Chantzios, T.; Alevizopoulou, S.; Skiadopoulos, S.; Tryfonopoulos, C. INTIME: A Machine Learning-Based Framework for Gathering and Leveraging Web Data to Cyber-Threat Intelligence. *Electronics* **2021**, *10*, 818. [CrossRef]
53. Simmons, C.; Ellis, C.; Shiva, S.; Dasgupta, D.; Wu, Q. *AVOIDIT: A Cyber Attack Taxonomy*; CTIT Technical Reports Series; University of Twente: Enschede, The Netherlands, 2009.
54. Ranaldi, L.; Pucci, G. Knowing Knowledge: Epistemological Study of Knowledge in Transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.