Assignment 1 – Data Cleaning and Preparation

MAT3120

Jordan Farrow (10653054)

Edith Cowan University, Joondalup

Dr. Johnny Lo

Table of Contents

1. Categorial and Continuous Summary Tables

Table 1

Categorical and Binary Variable Frequencies and Proportions

Categorical	Category	N (%)
Port	25	130 (32.5%)
	80	66 (16.5%)
	443	144 (36.0%)
	993	53 (13.25%)
	Missing	7 (1.75%)
Protocol	TCP	362 (90.5%)
	UDP	38 (9.5%)
	Missing	0 (0.0%)
Target Honeypot Server OS	Linux	3 (0.8%)
	MacOS (All)	172 (43.0%)
	Windows (Desktops)	224 (56.0%)
	Windows (Servers)	1 (0.3%)
	Missing	0 (0.0%)
Source OS (Detected)	Linux (>4.0)	50 (12.5%)
	Linux (2.6.3)	51 (12.8%)
	MacOS	154 (38.5%)
	Windows 10	135 (33.8%)
	Windows Server 2008	6 (1.5%)
	???	4 (1.0%)
	Missing	0 (0.0%)
Source Port Range	1024 - 4096	10 (2.5%)
	4096 - 8192	50 (12.5%)
	>8192	208 (52.0%)
	Blank	132 (33.0%)
	Missing	0 (0.0%)
Source IP Type (Detected)	IPv6 (unkown)	31 (7.8%)
	Standard ISP	305 (76.3%)
	Suspected Proxy/VPN	62 (15.5%)
	TOR Exit Node	2 (0.5%)
	Missing	0 (0.0%)
Advanced Persistent Threat (APT)	YES	183 (45.8%)
	NO	217 (54.3%)
	Missing	0 (0.0%)

Table 2

Continuous Variable Summaries

Continuous Feature	N (%) Missing	Min	Max	Mean	Median	Skewness
Hits	0 (0.0%)	223971.0	57600141.0	13305073.0	10439694.0	1.2
Average Request Size (bytes)	0 (0.0%)	1120.0	5836.0	3067.4	3076.5	0.2
Attack Window (seconds)	0 (0.0%)	3381.0	27555.0	13029.1	12700.5	0.4
Average Attacker Payload Entropy (bits)	0 (0.0%)	3.8	17.0	9.8	9.8	0.2
Attack Source IP Address Count	0 (0.0%)	-1.0	350.0	75.9	66.0	1.1
Average Ping to Attacking IP (milliseconds)	0 (0.0%)	0.0	99999.0	4406.6	136.2	4.5
Average Ping Variability (st.dev)	0 (0.0%)	1.4	345.6	23.9	15.0	4.4
IP Range Trust Score	328 (82.0%)	-3.4	4.0	0.2	0.3	0.0

Note. 82.0% of 'IP Range Trust Score' were missing values, resulting in the table row being formed by the 18.0% of complete cases.

2. Invalid Categorical Variable Categories and Values

Seen in table 1, 'source OS' has an invalid category '???' with 4 values accounting for 0.3% of the total entries. With the variable name as context, it is assumed that the category contains entries where an OS was not detected therefore, will be considered as missing values. Furthermore, 'source

port range' contains a category without a name, containing 132 values accounting for 33.0% of the variable. Without the ability to investigate further, it is assumed the values are outside the defined port ranges and will be treated as missing values. In addition to the variables noted, 'port' in table 1 has 7 missing values for 1.8% of entries. With the inability to investigate the causes of these missing values in each variable these will be omitted, viewing only complete cases and noting the percentage removed when conducting an analysis.

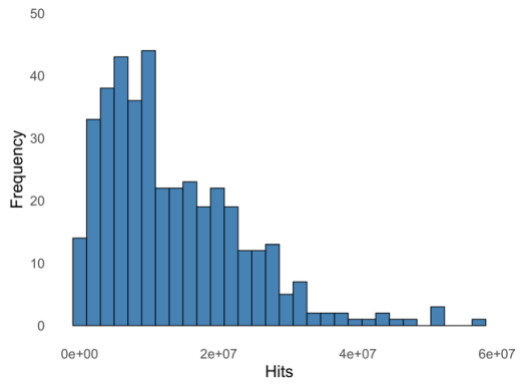
3. Continuous Variable Outliers

3.1 Hits

The variable's positive skew value indicates that outliers could exist closer to the maximum value, with most values being positioned closer to the minimum value as seen in figure 1. The need for investigation is supported by the large range between the minimum and maximum having a value of 57,376,170, indicating that the maximum could be influencing the accuracy of values in table 2. To determine the existence of outliers, the 1.5*IQR rule was utilised, where values below $Q_1 - 1.5*IQR$ and values above $Q_1 + 1.5*IQR$ are considered outliers (Kahn Academy, n.d.; Eberly College of Science, n.d.). As a result, 9 outliers were identified above the upper threshold, accounting for 2.3% of values. Due to the strong positive skewness seen in the table and figure, a square-root non-linear data transformation is recommended; this will reduce the skewness seen in figure 1, whilst reducing the impact the outliers have on the summary in Table 2 (Carrascosa, 2025; Lee, 2020; Osborne, 2002).

Figure 1

Frequency of hits made on networks



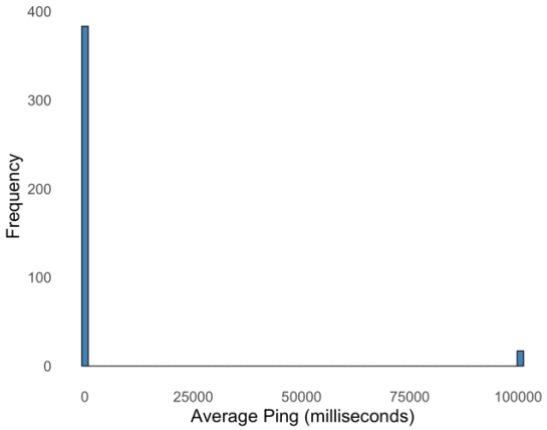
3.2 Attack Source IP Address Count

With the variable's positive skew value of 1.1 demonstrated in table 2, it shows that most values are positioned closer to the minimum of -1, with the possibility of outliers existing closer to the maximum value 350, supported visually by figure 1. Due to its positive skew, the 1.5*IQR method was used to determine the existence of outliers (Kahn Academy, n.d.; Eberly College of Science, n.d.). As a result, 7 outliers in the upper threshold were discovered, accounting for 1.75% of values. The square-root non-linear data transformation will be utilised for these values, having the same effect explored for the variable 'hits' (Carrascosa, 2025; Lee, 2020; Osborne, 2002).

3.3 Average Ping to Attacking IP

Figure 3

Frequency of the Average Ping to Attacking IP's



In table 2, the 4.5 skewness value implies a large cluster of values sit closer to the minimum 0, with a smaller quantity being closer to the maximum 99999. This is visually seen in figure 3, where the isolated values are shown skewing the skewness value. The isolated cluster includes 17 outliers accounting for 4.3% of the values. With the variable name as context, it is assumed that the outliers were caused by pings not reaching their target, causing a max ping of 99999 milliseconds before it was stopped automatically and timed out (Kentik, 2024). It is recommended to remove the outliers as they do not depict an average ping in order to populate table 2 with accurate values.

3.4 Average Ping Variability

Figure 4

Frequency of Average Ping Variabilities

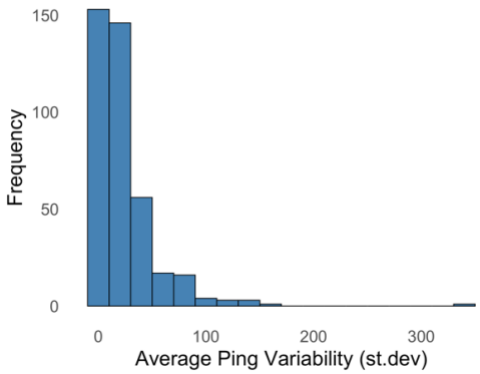


Table 2 notes the variable's strong positive skew with a value of 4.4, seen visually in figure 4 to be caused by a distant isolated outlier, accounting for 0.25% of values. It is assumed the outlier was caused by unstable network conditions, causing poor ping variability (Integrated Research, n.d.; Superloop, n.d.). Due to being far from an average value, it is recommended to remove the outlier to reduce skewing and obtain accurate results.

4. References

Carrascosa, I. P. (2025, March 11). *Dealing with Outliers: A Complete Guide*. KDnuggets. https://www.kdnuggets.com/dealing-with-outliers-a-complete-guide

Eberly College of Science. (n.d.). *Identifying Outliers: IQR Method*. Pennsylvania State University. https://online.stat.psu.edu/stat200/lesson/3/3.2

Integrated Research. (n.d.). *Network Jitter – Common Causes and Best Solutions*. https://www.ir.com/guides/what-is-network-jitter

Kahn Academy. (n.d.). *Identifying outliers with the 1.5xIQR rule*. https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule

Kentik. (2024, June 8). *Understanding the Ping Command in Network Troubleshooting and Monitoring*. https://www.kentik.com/kentipedia/ping-command-in-network-troubleshooting-and-monitoring/

Lee D. K. (2020). Data transformation: a focus on the interpretation. *Korean journal of anesthesiology*, 73(6), 503–508. https://doi.org/10.4097/kja.20137

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, 16(2). https://openpublishing.library.umass.edu/pare/article/1455/galley/1406/view/

Superloop. (n.d.). *Ping matter: Why low latency internet gives you the advantage*. https://www.superloop.com/blog/how-important-is-your-latency-and-ping-rate-really/