# 

**Identifying Cybersecurity Anomalies: Data Analytics**

Shelly Tingler

Colorado State University Global

MIS581: Capstone: Business Intelligence and Data Analytics

Dr. Orenthio Goodwin

May 8, 2022

**Abstract**

Cybersecurity plays a significant role in today’s information technology environment. Organizations rely on various techniques such as machine learning and data analytics to ensure data and information protections are in place to combat malicious actors. This study aims to determine if network features, specifically source and destination bytes, can determine an anomaly or normal indication. The objective sets a benchmark for organizations to utilize these results by creating policies and strategies to counter malicious activities and minimize risk to information systems. To test the hypothesis, multivariate logistic regression analysis was used to determine if a relationship exists between source and destination bytes with an outcome indication. The results showed that the predictor variables were statistically significant to the response variable. Results also showed that the analysis was accurate by use of a confusion matrix and offering more true predictions for classification. The results of this data analytic project suggest that network features such as source and destination bytes are related to the type of network indication. Organizations are more likely to use this data to create defensive strategies and policies to counteract suspicious activity associated with these results.

*Keywords:* data analytics, cybersecurity, logistic regression, network indication

**Contents**

[Identifying Cybersecurity Anomalies: Data Analytics 4](#_Toc102150234)

[Data and Information 4](#_Toc102150235)

[United States Cyber Command (US Cybercom) 5](#_Toc102150236)

[Organizational Opportunities and Objectives 5](#_Toc102150237)

[Cybersecurity 6](#_Toc102150238)

[Cybersecurity Polices and Strategic Planning 6](#_Toc102150239)

[Purpose/Problem 7](#_Toc102150240)

[Hypotheses Overview 7](#_Toc102150241)

[Literature Review 8](#_Toc102150242)

[NSL-KDD Dataset 9](#_Toc102150243)

[Research Methodologies and Tools 10](#_Toc102150244)

[Quantitative Methodologies 11](#_Toc102150245)

[Data Exploration, Data Visualization, and Findings 11](#_Toc102150246)

[*Descriptive Statistics* 12](#_Toc102150247)

[Regression Analyses 15](#_Toc102150248)

[Limitations 16](#_Toc102150249)

[Ethical considerations 17](#_Toc102150250)

[Findings 18](#_Toc102150251)

[Conclusion 19](#_Toc102150252)

[Recommendations 20](#_Toc102150253)

[References 22](#_Toc102150254)

# Identifying Cybersecurity Anomalies: Data Analytics

Information technology has grown drastically over the course of the last few years. Many organizations use information technology (IT) to conduct normal operations and will continue to do so well into the future. A large reason for IT playing a large role in the success of these organizations is data. Information technology uses computers to mainly process, create, retrieve, and use data. Organizations recognize the intrinsic value behind data. Organizations know that data is integral to remain competitive and grow within their industry. The opportunity to use data effectively is a challenge most organizations face today. Organizations, though, are willing to navigate through these challenges because the value in analysis creates tremendous opportunity.

# Data and Information

Data analytics is a critical tool used by many organizations. The generalizations produced by data analytics provide insights and trends significant in meeting an organization’s mission objectives. There are four types of analytics: descriptive, diagnostic, predictive, and prescriptive. Combined with a business setting, analytics “extract meaningful insights from data that an organization can use to inform its strategy and ultimately, reach its objectives” (Stobierski, 2021, para 2). Organizations know not to underestimate the value of data. Value can be derived from data in numerous ways. Sadowski (2019) notes that value can be found by the way data is used to profile and target people, optimize systems, manage, and control things, model probabilities, and grow the value of assets.

Information technology plays a large role in meeting an organization’s mission objectives. Big data’s influence in strategic planning changes the scope and direction for an organization. Data and information are significant parts of big data and create value in several ways. The possibilities to build on its potential are endless. The constant need to collect and protect data is a necessary task for organizations in this data-driven world. An organization’s success is dependent on how it manipulates and utilize data especially when this year alone in 2022, seventy percent (70%) of the world’s GDP underwent digitalization (Bulao, 2022).

# United States Cyber Command (US Cybercom)

The United States Cyber Command works as a combatant command for the nation’s cyberspace domain. The command consists of several entities comprising of military and intelligence communities, including information capabilities. Its mission is to “direct, synchronize, and coordinate cyberspace planning and operations to defend and advance nationals interests” (US Cybercom, n.d.). The United States Cyber Command mainly consists of the United States Army Cyber Command, the 24th Air Force, the United States Tenth Fleet /Fleet Cyber Command, and the Marine Corps Forces Cyberspace Command.

# Organizational Opportunities and Objectives

Data provided by network monitoring provides various opportunities for organizations to use for machine learning. Machine learning uses statistical analyses and computer algorithms with data. Network features such as source and destination bytes provide valuable information about network traffic. Source bytes are the specific number of bytes from the source host. Destination bytes are the specific number of bytes from the destination host. For this research study, the number of source and destination bytes can hypothetically determine if an anomaly exists. If there is a relationship between the number of bytes (predictor) and anomaly (response), the United States Cyber Command can strategically plan countermeasures against the associated internet protocol (IP) connections. These policies and plans then aid an organization in cybersecurity protections.

## **Cybersecurity**

Cybersecurity uses practices to help critical systems, networks, and programs against digital attacks. These practices utilize physical, operational, and technical functions for protections. It is one of the most significant components in an organization’s information security strategy. Digital attacks remain a substantial focus. These attacks come from various directions and sources and can be devastating to an organization. Global markets have resulted in estimated annual losses of $445 billion (Caporole et al., 2021). Not only do these attacks cause financial loss and revenue, but it can also be harmful. For governments and communities, the danger lies in national interests, especially infrastructure and communication.

## **Cybersecurity Polices and Strategic Planning**

Many organizations combat cybersecurity threats by creating and utilizing policies to form countermeasure and defensive strategies. Cybersecurity policies provide a culture within an organization that encourages shared values, guidelines, and practices. Policies are social, operational, and/or technical. All in all, these policies help organizations build an effective strategy against digital attacks. Without these policies in place, organizations are left vulnerable and can be easily exploited. The United States Cyber Command is no different from most organizations in their fight with malicious activities and cyber threats.

# Purpose/Problem

Cybersecurity threats are constantly evolving and the task to create a multi-faceted, effective policy is difficult. It is not, though, impossible. Data analytics provides the necessary tools to prevent major incidents. Continuous real-time collection and network monitoring offer a “suitable risk mitigation process before attacks can cause severe damages” (Rassam et al., 2017, p 124). Network monitoring consists of several different tactics. Its approach varies from organization to organization. It also provides system administrators and network analysts the needed information to determine if the organization’s network is behaving normally. Several factors influence the behavior of a network so recognizing anomalies is critical. If system administrators use data analytics to determine what factor influences an anomaly, organizations may have the ability to mitigate risks to an advantage.

* Do specific network features have a significant effect on a normal and/or anomaly network indication?

# Hypotheses Overview

Hypotheses are an integral component of the research methodology. Hypotheses are generated after a concise and adequate research question is formed. It provides the direction necessary for testing. Dubois (n.d) notes that hypothesis testing solidifies an observation based on statistics and compares results. These results determine if an occurrence arose from an event that was statistically significant or if merely coincidental. The following hypotheses focuses on two network features: source and destination bytes.

* H0: The number of source bytes is statistically significant to the outcome of network detection.
* HA: The number of source bytes is not statistically significant to the outcome of network detection.
* H0: The number of destination bytes is statistically significant to the outcome of network detection.
* HA: The number of destination bytes is not statistically significant to the outcome of network detection.

# Literature Review

Utilizing big data analytics and data mining in the ability to monitor network traffic have been a long assessed by experts. Results concluded that a correlation exists among features corresponding to network traffic logs. According to Nkiru and Fredrick (2021), big data analytics “significantly enhance the detection capabilities, enabling to detect advanced persistent threat (APT) activities that are passing under the radar of traditional security solutions” (p. 663). Organizations are implementing data analytics within research methods to combat cybersecurity threats. It is more common for organizations to use machine learning to classify and predict network traffic.

Verma and Ranga mentions machine learning playing a vital role in developing better network intrusion detection systems. Better network intrusion detection systems are crucial in instilling effective cyber defense tools against malicious activities. The methods utilized in Verma and Ranga’s study consist of large databases where applying data mining techniques may be difficult to conduct because of the amount of data points. The use of different data analytic techniques outside of regression analyses, especially in Verma and Ranga’s study, only solidifies that different machine learning techniques will vary but can be used in developing effective network intrusion detection systems.

Sai et al. (2021) also adds to more current research in the use of data analytics in developing network intrusion detection. Ironically, the authors also use the KDD dataset as part of its research study. For this study, the authors use a support vector machine request count within the features and create a model that identify and organize average traffic. The authors offer insight on the significance behind data analytics in today’s cyber environment and the use of big data is an inevitable part of the process.

The main idea behind the research consists of two elements: implementing big data analytics to cybersecurity practices and connecting these conclusions in building strategies and more effective defensive strategies and policies. Organizations often require specific guidelines to identify and defend against cyberattacks and/or malicious activities. Without a policy in place, devastating implications on organizational operations occur. The article written by Oyelami and Kassim reinstate the importance of strategic planning and policy writing in a cyber space environment. Conclusions derived from network monitoring help identify and “synchronise the security practices into one holistic framework for cybersecurity defense policies” (Oyelami & Kassim, 2020, p 131).

# NSL-KDD Dataset

The dataset used for this research project stems from the KDDCup in 1999. The KDDCup is an annual Data Mining and Knowledge Discovery Competition organized by ACM Special Interests Group on Knowledge Discovery and Data Mining. The dataset tasked competitors to formulate and use data mining techniques and tools to create a predictive model pertaining to a network intrusion detector. Network intrusion detectors distinguish illegitimate connections within a computer network. As part of the task, the data collected was from a simulated military network environment.

The dataset in use, NSL-KDD, holds the same forty-two (42) different attributes as the original KDDCup dataset and is divided into three sections: individual transmission control protocols (TCPs), connections suggested by domain knowledge, and traffic features. The NSL-KDD dataset, though, is thought to be an improvement of the KDDCup dataset. The NSL-KDD dataset has 22,544 rows vice the more than 1 million rows from the original dataset.

The specific features used for this research study are source bytes, destination bytes, and outcome. Source and destination bytes are continuous, numerical values; outcome is categorical. For the sake of this research, outcome will be replaced by either the values 1 or 0. Normal tendencies will be labeled as ‘1’ while anomalies will be labeled as ‘0’. The modification is necessary as part of the data preparation process. Data preparation is a significant part of data analyses and processing. By modifying the data, all values within the dataset remain numerical for ease of comparison and forming relationships. This aids the use of data analytic tools and methods.

# Research Methodologies and Tools

Hypothesis testing methods “offers validity of results through the use of statistics and probability” (O’Leary, 2021, p 143). Cybersecurity policies require necessary planning through statistical analyses. The purpose behind statistical analyses is to provide an organization, like that of the United States Cyber Command, an idea of how to manage the decision-making process associated with network security. It is not a means to end all malicious actors and/or activities but provides a chance for the organization to act proactively vice reactively. Planning allows for an organization to build a network defensive strategy that recognizes a threat and mitigate the risks associated with such threat.

## **Quantitative Methodologies**

There are several types of research methodologies: quantitative, qualitative, and mixed methods. Qualitative methods require the use of data to understand and interpret people and situations under investigation (O’Leary, 2021). It is less exact than a quantitative approach. Quantitative methodology deals more with numbers, statistical analyses, measurements, and their inferences (McLeod, 2019). Mixed methods use a combination of qualitative and quantitative methods to conduct analyses. The use of methodology depends on what the researchers’ objectives are. Data preparation and exploration have shown that majority of the data used for this research project contains data requiring a quantitative approach.

## **Data Exploration, Data Visualization, and Findings**

Data exploration is a critical part of data analytics. Data exploration allows researchers to uncover patterns and characteristics prior to data and statistical analysis. Data exploration create a general picture of the data and gives a chance for researchers to look for missing values and/or details that may need to be studied more in greater detail (Sisense, n.d.). It provides more insight and often leads to a better questioning attitude and possible hypotheses. Data exploration consists of different types of study, but a major advantage, especially when using analytic tools, is visualization. Data visualization adds a lot of critical details that may be missed from exploring the numerical data alone. Data visualization offers the “understanding of the underlying structure within the data, which can enable hypothesis generation and data interpretation” (Rhodes et al., 2020).

### ***Descriptive Statistics***

There are various types of descriptive statistics prevalent in research. Descriptive statistics is a beneficial way to determine what kind of data is available for research. Descriptive statistics helps researchers “understand and describe the aspects of a specific set of data by providing brief observations and summaries about the sample” (Conner & Johnson, 2017, p 52). Determining the means of tendency describes a dataset by identifying the central position. In this case, the central tendency is a significant feature.

The research question focuses on three variables: source bytes, destination bytes, and outcome. Central tendency consists of the mean, median, and mode. It allows research to focus on numerical values to form a working hypothesis relevant to the research question. Figure 1 shows a table of descriptive statistics for three variables: source bytes, destination bytes, and outcome. Data points between the 0-50% show similar number of bytes in terms of tendency. Toward the 75% quartile, numbers drastically change for source and destination bytes. The data for destination bytes increase in numbers are higher than 500 bytes as supposed to source bytes which stay under 500 bytes. The means are also significantly different between source and destination bytes. The number of bytes for network traffic flow is more evident for source bytes. The mean for source bytes is approximately 10,000 bytes while destination bytes remain closer to 2,000 bytes.

**Figure 1**

Descriptive Statistics- Source bytes, destination bytes, and outcome

Table

Description automatically generated

*Note.* Figure 1 depicts a screenshot of the descriptive statistics associated with source bytes, destination bytes, and outcome from NSL-KDD dataset.

Bar charts are also helpful in determining the ratio among values for a variable. In this case, a bar chart offers the researcher an insight between a normal interaction vice an anomaly within a network system. The value count between normal and anomalies are 9,711 to 12,833. Figure 2 offers a visualization in respect to the value counts of outcome. At first glance, both indications are close in value and do not offer a clear visualization between the two types of outcomes. Further investigation and research are required to make proper assumptions and propositions. For instance, the number of anomalies is much higher than normal indications. The number of source and destination bytes that show an anomaly, though, are much lower than what shows of a normal indication. The number of source bytes show very similar numbers for both an anomaly and normal indication as compared to destination bytes.

Descriptive statistics, alongside visualizations, also show different traits about the data and noticeable trends like outliers. Outliers call attention for the researcher to further review the data (Shmueli et al., 2018). Boxplots, for example, provide information about outliers, the minimum, first quartile, median, third quartile, and the maximum position of where data points lie. Figures 3 and 4 show the number of outliers outside the central tendency of both the source and destination bytes. The number of destination bytes for a normal indication have a larger range than that of an anomaly. There are several outliers present for both features and may require further investigation regarding the dataset. Although box plots offer a visual summary into destination and source bytes, grouped by outcome, the outliers may indicate different possibilities. These possibilities may range among insufficient data information, skewness, and/or bias.

**Figure 2**

Bar Graph- Anomalies versus Normal indication

Chart, bar chart

Description automatically generated

*Note.* Figure 2 depicts a bar graph denoting the comparison between an anomaly and normal indication. Anomalies are denoted by ‘0’; normal indications are denoted by ‘1’.

**Figure 3**

Box Plot- Destination Bytes Grouped by Outcome

Chart, box and whisker chart

Description automatically generated

*Note.* Figure 3 depicts a box plot comparing the network feature destination bytes grouped by outcome. Anomalies are denoted by ‘0’; normal indications are denoted by ‘1’.

**Figure 4**

Box Plot- Source Bytes Grouped by Outcome

Chart, box and whisker chart

Description automatically generated

*Note.* Figure 3 depicts a box plot comparing the network feature source bytes grouped by outcome. Anomalies are denoted by ‘0’; normal indications are denoted by ‘1’.

## **Regression Analyses**

Regression models are used to determine if there is a relationship between two variables. The focus of regression analysis is to see if source and destination bytes have an impact on the outcome of a network detection system. Regression analyses uses p-values and coefficients to define if the variables are statistically significant and whether a relationship exists between the predictor and response variables. If the p-value is greater than a given significant level, in this case 0.05, there is not enough evidence to conclude that a relationship exists.

A logistic regression model is well suited for the research question at hand. Logistic regression models predict if the response variable, outcome, has a relationship with a predictor variable. The predictor variables are source and destination bytes. The difference between a linear regression and logistic regression model is that logistic regression predicts a binary outcome. The binary outcome for this research project is 1 or 0, normal or anomaly respectively. Logistic regression model provides “a measure of how a predictor (coefficient size) is, but also its direction of association (positive or negative)” (Ranjan-Rout, 2020, para 2). Logistic regression models also work best with response variables that are categorical in nature. In this case, *outcome* is denoted by either a 0 or 1, anomaly or normal respectively.

# Limitations

One of the most significant limitations of this research project was that the data does not represent existing real networks. McHugh (2000) argues that the data contained within the dataset would not allow for specific intrusions to be executed properly and the distribution of intrusions unrealistic. It does, though, provide an “effective benchmark dataset to help researchers compare different intrusion detection methods”(University of New Brunswick, n.d., para 1). Although sharing data knowledge about advanced persistent threats (APTs) and malware can be helpful to organizations worldwide, it can be difficult for organizations to divulge information that may be exploited and used against their own networks.

The original dataset is out of date and has many redundant entries making the data imbalanced. Even with the more updated version of the dataset in NSL-KDD, data and information for network intrusion detection systems are far and few in between. Many organizations keep data and information regarding networks private for security reasons. Majority of the information available online are similar and hold the same objectives/purpose of the original dataset: to provide the support that machine learning is an effective tool for network intrusion detection systems.

# Ethical considerations

There are several ethical considerations to contemplate with the type of data associated with the NSL-KDD dataset. Cyberspace is a new landscape being embraced by organizations worldwide. More organizations, communities, and businesses are moving toward digital transformation. By 2023, seventy-five percent of organizations would have implemented a digital transformation roadmap to their business strategy (McLellan, 2021). The goals of hackers and malicious actors have drastically shifted their focus as organizations welcome and accept digitalization as part of their business strategy within the last two decades. Advanced technologies like machine learning, artificial intelligence, cloud computing, and the Internet of Things (IoT) are reshaping the cyber defense landscape and forces many organizations to look at what lays ahead within the industry (Pattison-Gordon, 2021).

The research behind cybersecurity, though, is not clearly represented in a realistic and standardized way. There is no such methodology in place and researchers scramble for ways to study the field without overstepping boundaries of ethics and privacy. Macnish and van der Ham (2020) note that cybersecurity practices “go largely ungoverned and unguided, despite the potential for significant harm” (p 1).

Data analytics and statistical analyses resulting from the NSL-KDD dataset can also be damaging to an organization. The results from the research compiled by this project can disclose system vulnerabilities and leave an organization open for exploitation. What risk, though, are organizations willing to expose themselves for the sake of research? The fight against cyberattacks should occur as a collective manner and sharing intelligence plays a critical part to worldwide cyber defense and success in network security. Sharing intelligence allows faster response time to cyberattacks while allowing businesses to prepare against said danger.

# Findings

Regression analysis results show that the two predictors, source and destination bytes, are statistically significant and have a relationship with the response variable, outcome. Source and destination bytes have p-values less than the significant level set at 0.05. These two variables have an impact on the type of outcome a network indication system would present. Therefore, both null hypotheses are accepted. There is enough evidence to suggest that the number of source and destination bytes do have an impact on the type of outcome for network indication. The accuracy of the logistic regression classifier on the test set was: 0.68. This high number represents the strength of the regression analyses as a classifier for an anomaly or normal indication. The confusion matrix results also suggest 3,067 correct predictions and 1,442 incorrect predictions.

**Figure 5**

Logistic Regression Analysis- Source and destination bytes versus Outcome

A screenshot of a computer

Description automatically generated with low confidence

*Note.* Figure 5 depicts the results of the logistic regression analysis between source and destination bytes versus outcome.

**Figure 6**

Precision, recall, f-measure, and support

Table

Description automatically generated

*Note.* Figure 6 shows a screenshot of the computed precision, recall, f-measure, and support of the logistic regression analysis between source/destination bytes and outcome.

# Conclusion

Data and information play a large role for organizations in today’s IT environment. As intangible assets, organizations recognize the need to protect and safeguard what they can against malicious actors. Utilizing the right analytic tools to explore data marks a significant part of a research project. Without a clean and adequate dataset, organizations can not properly commence statistical analyses to ensure the right research components and methodologies are employed.

Forming hypotheses and statistical testing are two major components of research project. It allows for analyses and establishes relationships between both predictor and response variables. The outcome of these research questions determines whether the concluding evidence justify the given hypotheses. Network features such as source and destination bytes are significant variables in network detection for anomalies and normal indications. These features must be used together with other factors such as types of protocols, internet protocol addresses, and interactions to and from different servers.

The NSL-KDD dataset provides a baseline approach of data information critical to what the United States Cyber Command may utilize for continued and future data analytic projects. The United States Cyber Command, like most organizations conducting network monitoring, benefit from these research methods. With statistical evidence supporting the analyses among variables, organizations use these results to create policies and/or strategies for an advantage against malicious actors and activities.

# Recommendations

The results from testing and analyses used to evaluate the NSL-KDD dataset present several key points for the United States Cyber Command. Testing and analyses show that key points must be addressed for the command to acknowledge and seek recommendations for an effective cyber defense strategy. This defense strategy must refocus and utilize the organization’s mission and provide ways to align past, present, and future sources to accommodate an evolving environment. These must include projects, policies, and activities (proactive and reactive) to successfully deliver those mission objectives.

The United States Cyber Command should continue to focus on network features, like source and destination bytes, and note when relationships exist with anomalies and normal indications. It must be stressed, though, that the NSL-KDD dataset was used as a benchmark and the United States Cyber Command must utilize more network features, combined with machine learning as a way to combat cybersecurity threats to build stronger policies. Data collection is an essential tool to continue a viable strategy for continued defensive countermeasures.

The range of data points that were present was large and incurred several outliers. The data visualizations, though, suggested that smaller number of source and destination bytes presented more anomalies than a higher number of bytes during network traffic. The average number of source and destination bytes that caused anomalies were less than five hundred (500) bytes- something that should not go unnoticed or ignored. In this case, rules must be set up on current network intrusion detection and prevention systems to counterattack these activities. It is important that guidelines and policies exist to protect data and information for the entire organization and its subsidiaries. Careful and diligent monitoring is a must. There were no significant indications to suggest what specific number of source and destination bytes would indicate a network indication, only that a relationship does exist between the two predictor variables and the response variable.

Guidelines and policies should denote that those numbers are what that United States Cyber Command acknowledge and are aware of but also suggest that malicious activities occur at all levels. Cyberthreats are not packaged under one set of factors or a set of numbers. Cyberthreats are a culmination of all risks and means of exploitation for any data and information system.

# References

Bulao, J. (2022 March 14). How much data is created every day in 2022? https://techjury.net/blog/how-much-data-is-created-every-day/

Caporale, G., Kang, W., Spagnolo, F. & Spagnolo, N. (2020 February). Cyberattacks and cryptocurrencies. *Cyber-attacks and cryptocurrencies* (Working Paper No. 2003). Brunel University London. https://www.brunel.ac.uk/economics-and-finance/research/pdf/2003-Feb-GMC-Cyber-attacks-and-Cryptocurrencies.pdf

Conner, B. & Johnson, E. (2017). Descriptive statistics. *American Nurse Today, 12*(11), 52-55. https://www.myamericannurse.com/wp-content/uploads/2017/11/ant11-Research-101-1017a.pdf

Dubois, S. (n.d.). The importance of hypothesis testing. https://sciencing.com/the-importance-of-hypothesis-testing-12750921.html

Harrington, D. (2021 November 18). How to monitor network traffic: Effective steps and tips https://www.varonis.com/blog/how-to-monitor-network-traffic

Horton, M. (2022 March 28). Simple random sample: Advantages and disadvantages. https://www.investopedia.com/ask/answers/042815/what-are-disadvantages-using-simple-random-sample-approximate-larger-population.asp

Macnish, K. and van der Ham, J. (2020). Ethics in cybersecurity research and practice. *Technology in Society, 63*(2020), 1-10. https://reader.elsevier.com/reader/sd/pii/S0160791X19306840?token=E6610A96975DC465EABB4BBD69D436F8A6260BD459DF165CCFF2B6EA87B0E4AE6B747986F16CFE73DED6B06E749E69EB&originRegion=us-east-1&originCreation=20220409210322

McHugh, J. (2000). Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and System Security, 3*(2000), 262-294.

McLeod, S. A. (2019). Qualitative vs. quantitative research. https://www.simplypsychology.org/qualitative-quantitative.html

McLellan, C. (2021 October 1). Digital transformation is changing. Here’s what comes next. https://www.zdnet.com/article/digital-transformation-is-changing-heres-what-comes-next/

Nadimmai, G. & Hemalatha, M. (2014). Effective approach toward intrusion detection system using data mining techniques. *Egyptian Informatics Journal, 15* (2015), 37-50. https://reader.elsevier.com/reader/sd/pii/S1110866513000418?token=B8B3D6046F5748B6659A69A8F47BCE77360A4E76F6D17A084A383DB32FCAED9EBB6BCDB1C01448938F4F078102723E15&originRegion=us-east-1&originCreation=20220417101758

Nkiru, E. and Fredrick, O. (2020). A data driven anomaly based on behavior detection method for advanced persistent threats (APT). *International Journal of Science and Research, 10*(8), 663-667. https://www.researchgate.net/profile/Monday-Ohemu/publication/356911879\_A\_Data\_Driven\_Anomaly\_Based\_Behavior\_Detection\_Method\_for\_Advanced\_Persistent\_Threats\_APT/links/61b27b8d19083169cb7f15d8/A-Data-Driven-Anomaly-Based-Behavior-Detection-Method-for-Advanced-Persistent-Threats-APT.pdf

O’Leary, Z. (2021). The Essential Guide to Doing Your Research Project (4th ed.). SAGE Publications Ltd.

Oyelami, J. and Kassim, A. (2020). Cyber security defence policies: A proposed guidelines for organisations cyber security practices. *International Journal of Advanced Computer Science and Applications, 11*(8), 131-138. https://pdfs.semanticscholar.org/10c4/20957afccbf85a2d99225931802d4aebe276.pdf

Pattison-Gordon. (2021 November). Through the years: A broad look at two decades in cybersecurity. https://www.govtech.com/security/through-the-years-a-broad-look-at-two-decades-in-cybersecurity

Ranjan-Rout, A. (2020 September 2). Advantages and disadvantages of logistic regression. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Rassam, M., Maarof, M. & Zainal, A. (2017). Big data analytics adoption for cybersecurity: A review of current solutions, requirements, challenges and trends. *Journal of Information Assurance and Security, 11*(2017), 124-145. http://mirlabs.org/jias/secured/Volume12-Issue4/Paper14.pdf

Rhodes, J., Cutler, A., Wolf, G. and Moon, K. (2020 Jun 15). Supervised visualization for data exploration, 1-21. https://arxiv.org/pdf/2006.08701.pdf

Sadowki, J. (2019 January). When data is capital: Datafication, accumulation, and extraction. *Big Data & Society,* 1-12. https://doi.org/10.1177/2053951718820549

Sai, R., Bhargav, J., Aneesh, M., Sahit, G., Nikhil, A. (2021 February). Discovery network intrusion using machine learning and data analytics approach. *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks*, 118-123. https://www.researchgate.net/profile/N-Raghavendra-Sai/publication/350531669\_Discovering\_Network\_Intrusion\_using\_Machine\_Learning\_and\_Data\_Analytics\_Approach/links/608d2a9e92851c490faac7e2/Discovering-Network-Intrusion-using-Machine-Learning-and-Data-Analytics-Approach.pdf

Shmueli, G., Bruce, P., Yahav, I., Patel, N. & Lichtendahl, K. (2018). Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. John Wiley & Sons Inc.

Sisense. (n.d.). What is data exploration? https://www.sisense.com/glossary/data-exploration/

Stobierski, T. (2021 January 5). What’s the difference between data analytics and data science? https://online.hbs.edu/blog/post/data-analytics-vs-data-science

University of New Brunswick. (n.d.). *NSL-KDD dataset.* https://www.unb.ca/cic/datasets/nsl.html

US Cybercom. (n.d.). U.S. Cyber Command: Our History. https://www.cybercom.mil/About/History/

Verma, A. and Ranga, V. (2018). Statistical analysis of CIDDS-001 dataset for network intrusion detection systems using distance-based machine learning. *6th International Conference on Smart Computing and Communications, 125*(2018), 709-716. https://reader.elsevier.com/reader/sd/pii/S1877050917328594?token=32CFDCBA2A96BC78B5F74E053B8636D0392E9BE3F5BCB23ED86523DB746867022D2837030E3040A916F779E3F6098ADC&originRegion=us-east-1&originCreation=20220417101001