Final\_Project: Introduction To Analytics - Arturo Bohuchot

Arturo Bohuchot

2024-12

## Introduction

Cybersecurity is a critical area of focus. The area has grown rapidly due to the increasing number of cyber-threats from various sources. Sources includes individuals, criminal organizations, state actors, and many more.

This project aims to analyze cybersecurity data from multiple datasets and provide valuable insights into attack types, patterns, and vectors. The main goal of this project is to identify relevant and interesting trends that will contribute to the development of improved defensive strategies.

## The Data: Cyber Attacks and Data Analysis

This project utilizes four distinct datasets and each will uniquely contribute on building a clear perspective on cybersecurity incidents and patterns:

Cybersecurity Attacks (cybersecurity\_attacks.csv): This dataset provides details about various cyberattacks, including attack types, affected systems, and protocols.

AWS Honeypot Logs (AWS\_Honeypot\_marx-geo.csv): A honeypot is a decoy system designed to attract cyber attackers. This dataset contains logs of malicious activities targeting an AWS-based honeypot. This data offers insights into an attackers behaviors and techniques.

Cybersecurity Dataset (Cybersecurity\_Dataset.csv): This dataset contains a broader range of cybersecurity-related data, including network traffic information, attack vectors, and defense mechanisms.

Hacker News Articles (HN\_posts\_year\_to\_Sep\_26\_2016.csv): A collection of posts from Hacker News, a platform where cybersecurity incidents, techniques, and news are discussed. This dataset provides a look into real-world relevant cyber attacks occurring globally.

# Initial Look at the Datasets

We first begin by looking at the variables within the datasets to get a better understanding on which variables can be used for analysis. Immediately, we can see variables that are of interest including:

From Cybersecurity Attacks Dataset:  
 Source Port, Destination Port, Protocol, Attack Type, Packet Length  
  
From AWS Honeypot Logs:  
 Protocol (proto), Destination Port (dpt), Source Port (spt), Coutry  
  
From Cybersecurity Dataset:  
 Threat Category, Attack Vector, Suggested Defense Mechanism  
  
From Hacker News Articles Dataset:  
 title

These key variables will be explored and analyzed to identify relevant information. Below is a quick display of variables belonging to each dataset.

names(cyber\_attacks) %>%  
 head(5)

## [1] "Timestamp" "Source IP Address" "Destination IP Address"  
## [4] "Source Port" "Destination Port"

names(aws\_attacks) %>%  
 head(5)

## [1] "datetime" "host" "src" "proto" "type"

names(cyber\_dataset) %>%  
 head(5)

## [1] "Threat Category" "IOCs (Indicators of Compromise)"  
## [3] "Threat Actor" "Attack Vector"   
## [5] "Geographical Location"

names(hn\_titles) %>%  
 head(5)

## [1] "id" "title" "url" "num\_points" "num\_comments"

# Clean and Combine Data for Analysis

The datasets were cleaned and combined to create unified data frames for analysis. This process involved:

1. Merging AWS Honeypot Logs and Cybersecurity Attacks  
2. Integrating Geographical Information  
3. Combining Attack Type Data

This combined dataset serves as a foundation for deeper exploration of attack patterns, network behavior, and geographical trends. The variables of the combined datasets can be seen below

df\_aws\_cyberattacks <- bind\_rows(  
 aws\_attacks %>%  
 select(  
 Protocol = proto,  
 `Source Port` = spt,  
 `Destination Port` = dpt  
 ) %>%  
 mutate(  
 Protocol = as.character(Protocol),  
 `Source Port` = as.numeric(`Source Port`),  
 `Destination Port` = as.numeric(`Destination Port`)  
 ),  
   
 cyber\_attacks %>%  
 select(  
 Protocol = Protocol,  
 `Source Port` = `Source Port`,  
 `Destination Port` = `Destination Port`  
 ) %>%  
 mutate(  
 Protocol = as.character(Protocol),  
 `Source Port` = as.numeric(`Source Port`),  
 `Destination Port` = as.numeric(`Destination Port`)  
 )  
)  
names(df\_aws\_cyberattacks)

## [1] "Protocol" "Source Port" "Destination Port"

df\_aws\_cyberdataset\_country <- bind\_rows(  
 aws\_attacks %>%  
 select(  
 Country = country  
 ) %>%  
 mutate(  
 Country = as.character(Country)  
 ),  
   
 cyber\_dataset %>%  
 select(  
 Country = `Geographical Location`  
 ) %>%  
 mutate(  
 Country = as.character(Country)  
 )  
)  
names(df\_aws\_cyberdataset\_country)

## [1] "Country"

df\_cyberdataset\_cyberattacks\_attack\_type <- bind\_rows(  
 cyber\_dataset %>%  
 select(  
 `Attack Type` = `Threat Category`  
 ) %>%  
 mutate(  
 `Attack Type` = as.character(`Attack Type`)  
 ),  
   
 cyber\_attacks %>%  
 select(  
 `Attack Type` = `Attack Type`  
 ) %>%  
 mutate(  
 `Attack Type` = as.character(`Attack Type`)  
 )  
)  
names(df\_cyberdataset\_cyberattacks\_attack\_type)

## [1] "Attack Type"

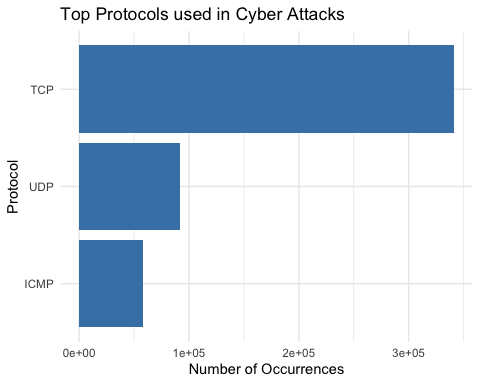
# Protocols and Ports used in Cyber Attacks

Understanding the protocols and ports involved in cyberattacks is crucial for identifying patterns in network exploitation.

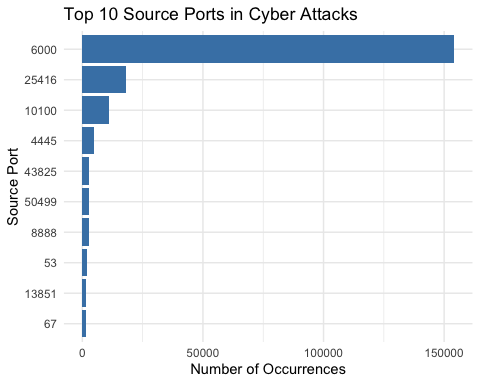
Top Protocols: The most frequently used protocols in cyberattacks were identified from a unified dataset then visualized using a bar chart. This analysis shows which specific network protocols are leveraged by attackers as opposed to other less-frequently used protocols. Some other popular protocols used in cyberattacks not shown on this analysis includes SMTP, DHCP, and ARP.

Source and Destination Ports: From the combined datasets, the top 10 source and destination ports that attackers commonly use for malicious traffic are shown. This is important because it shows that malicious traffic can target ports that are not commonly used. Protocols have default ports, for example in the format Protocol:Port; the following default configurations typically are HTTP:80, HTTPS:443, MySQL:3306, Telnet:23, DNS:53 just to name a few. What this data tells us is that not enough preventive measures were taken to mitigate this risk, by leaving unused ports open, attackers have a larger attack-surface. This can easily be mitigated by configuring proper firewall rules or “white-listing” specific ports which will deny ingress and egress traffic from all other non-whitelisted ports.

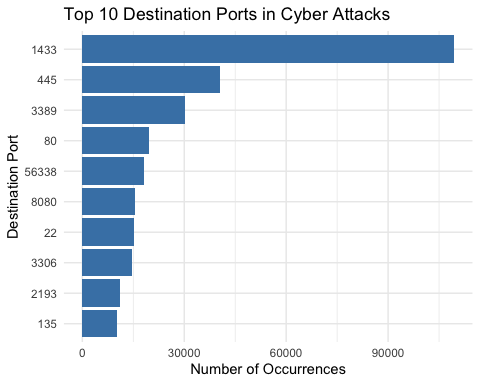
protocol\_analysis <- df\_aws\_cyberattacks %>%  
 count(`Protocol`, sort = TRUE)  
  
ggplot(protocol\_analysis, aes(x = reorder(`Protocol`, n), y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 coord\_flip() +  
 theme\_minimal() +  
 labs(  
 title = "Top Protocols used in Cyber Attacks",  
 x = "Protocol",  
 y = "Number of Occurrences"  
 )



source\_port\_analysis <- df\_aws\_cyberattacks %>%  
 filter(!is.na(`Source Port`)) %>%   
 count(`Source Port`, sort = TRUE) %>%  
 head(10)   
  
ggplot(source\_port\_analysis, aes(x = reorder(`Source Port`, n), y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 coord\_flip() +  
 theme\_minimal() +  
 labs(  
 title = "Top 10 Source Ports in Cyber Attacks",  
 x = "Source Port",  
 y = "Number of Occurrences"  
 )



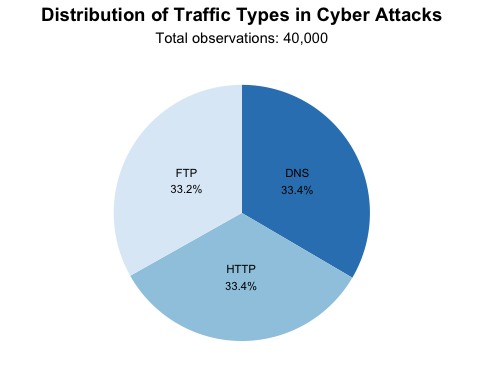
destination\_port\_analysis <- df\_aws\_cyberattacks %>%  
 filter(!is.na(`Destination Port`)) %>%  
 count(`Destination Port`, sort = TRUE) %>%  
 head(10)  
  
  
ggplot(destination\_port\_analysis, aes(x = reorder(`Destination Port`, n), y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 coord\_flip() +  
 theme\_minimal() +  
 labs(  
 title = "Top 10 Destination Ports in Cyber Attacks",  
 x = "Destination Port",  
 y = "Number of Occurrences"  
 )



# Traffic Type Analysis

A breakdown of traffic types in cyberattacks reveals the distribution of attack vectors. The polar bar chart displays the percentage of occurrences for each traffic type. This analysis helps identify dominant attack vectors and prioritize defensive measures aiding in the development of targeted mitigation strategies. With this knowledge, a cybersecurity professional is able to pay close attention to this traffic type and monitor it closely for any anomalies.

traffic\_type\_analysis <- cyber\_attacks %>%  
 count(`Traffic Type`, sort = TRUE) %>%  
 mutate(percentage = (n/sum(n))\*100)  
  
  
ggplot(traffic\_type\_analysis,   
 aes(x = "", y = n, fill = reorder(`Traffic Type`, n))) +  
 geom\_bar(stat = "identity", width = 1) +  
 coord\_polar("y", start = 0) +  
 theme\_minimal() +  
 scale\_fill\_brewer(palette = "Blues") +   
 geom\_text(aes(label = sprintf("%s\n%.1f%%", `Traffic Type`, percentage),  
 y = cumsum(n) - n/2),  
 size = 3) +  
 theme(  
 axis.text = element\_blank(),  
 axis.title = element\_blank(),  
 panel.grid = element\_blank(),  
 plot.title = element\_text(size = 14, face = "bold", hjust = 0.5),  
 plot.subtitle = element\_text(hjust = 0.5),  
 legend.position = "none"  
 ) +  
 labs(  
 title = "Distribution of Traffic Types in Cyber Attacks",  
 subtitle = paste("Total observations:", scales::comma(sum(traffic\_type\_analysis$n)))  
 )



# Analyzing Attack Types

From the combined datasets, we notice the most prevalent attack types including: DDoS, Malware, and Intrusion.

Top Attack Types: Although these 3 attack types are rather close to each other in terms of popularity, Distributed Denial of Service DDoS ranked at the top. We can assume this to be the case because DDoS is the only attack on here that does not rely on compromising or “breaking-into” the target asset. Rather, it leverages distributed computing resources, sometimes referred to as “zombie computers”, that floods the target with network requests ultimately exhausting the target’s resources causing it to stall, crash, or completely take it offline.

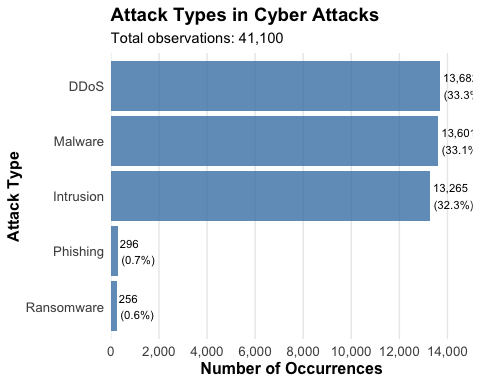
Malware is still a popular choice as there are many forms of malware such as spyware, RAT (Remote Access Trojan), Worms (a type of malware that spreads and infects several computers within a network) just to name a few.

Intrusion is also popular as it relies on exploiting a weakness in the target system allowing attackers to benefit from the compromise. This compromise typically leads to some sort of data ex-filtration such as stealing a website’s database containing usernames, emails, full\_name, credit card information and more.

Other attack types are shown but did not come close to the top 3. Phishing is a popular Social Engineering tactic that directly takes advantage of the decision making of individuals and/or groups. Ransomware is a very specific and popular type of malware where the victim will have their data encrypted rendering it obsolete and inaccessible. The victim will then have to pay a ransom to the attacker for the decryption key or decryption software.

Total Observations: Each attack type’s contribution to the overall dataset is annotated with its percentage and the total number of occurrences across all attack types.

attack\_type\_analysis <- df\_cyberdataset\_cyberattacks\_attack\_type %>%  
 count(`Attack Type`, sort = TRUE) %>%  
 mutate(percentage = (n/sum(n))\*100)  
  
ggplot(attack\_type\_analysis, aes(x = reorder(`Attack Type`, n), y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue", alpha = 0.8) +  
 coord\_flip() +  
 theme\_minimal() +  
 scale\_y\_continuous(  
 labels = scales::comma,  
 breaks = scales::pretty\_breaks(n = 10),   
 expand = expansion(mult = c(0, 0.1))   
 ) +  
 geom\_text(aes(label = sprintf("%s\n(%.1f%%)", scales::comma(n), percentage)),  
 hjust = -0.1,   
 size = 3) +   
 theme(  
 axis.text = element\_text(size = 10),  
 axis.title = element\_text(size = 12, face = "bold"),  
 plot.title = element\_text(size = 14, face = "bold"),  
 panel.grid.major.y = element\_blank(),   
 panel.grid.minor = element\_blank()   
 ) +  
 labs(  
 title = "Attack Types in Cyber Attacks",  
 subtitle = paste("Total observations:", scales::comma(sum(attack\_type\_analysis$n))),  
 x = "Attack Type",  
 y = "Number of Occurrences"  
 )



Analyzing Packet sizes by Attack Type

This analysis examines packet size characteristics for different attack types, providing insights into how attackers use network traffic.

Average Packet Length: The polar bar chart illustrates the average packet size for each attack type. DDoS attacks have the largest average packet size, reflecting their nature of flooding networks with large volumes of traffic.

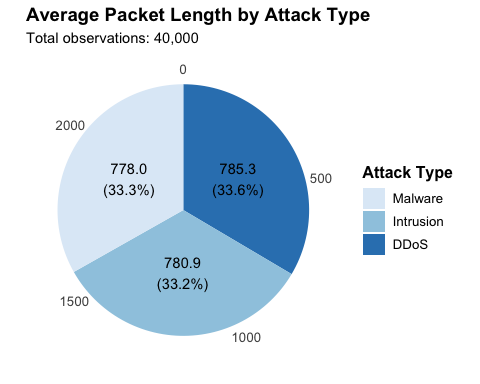
Statistical Summaries: For each attack type, metrics such as average, median, maximum, and minimum packet lengths were calculated, showing variations in packet sizes. These metrics help distinguish the traffic patterns of DDoS, Malware, and Intrusion attacks.

By understanding the variations of packet length, src+dst ports, protocol, and traffic type in cyberattacks; it may further equip cyber-defenders with theh knowledge necessary to prevent a system compromise.

packet\_analysis\_by\_attacktype <- cyber\_attacks %>%  
 group\_by(`Attack Type`) %>%  
 summarise(  
 n = n(),  
 avg\_packet\_length = mean(`Packet Length`, na.rm = TRUE),  
 med\_packet\_length = median(`Packet Length`, na.rm = TRUE),  
 max\_packet\_length = max(`Packet Length`, na.rm = TRUE),  
 min\_packet\_length = min(`Packet Length`, na.rm = TRUE),  
 .groups = 'drop'  
 ) %>%  
 arrange(desc(avg\_packet\_length)) %>%  
 mutate(percentage = (n/sum(n))\*100)  
  
packet\_analysis\_by\_attacktype

## # A tibble: 3 × 7  
## `Attack Type` n avg\_packet\_length med\_packet\_length max\_packet\_length  
## <chr> <int> <dbl> <dbl> <dbl>  
## 1 DDoS 13428 785. 786 1500  
## 2 Intrusion 13265 781. 785 1500  
## 3 Malware 13307 778. 777 1500  
## # ℹ 2 more variables: min\_packet\_length <dbl>, percentage <dbl>

ggplot(packet\_analysis\_by\_attacktype,   
 aes(x = "", y = avg\_packet\_length,   
 fill = reorder(`Attack Type`, avg\_packet\_length))) +  
 geom\_bar(stat = "identity", width = 1) +  
 coord\_polar("y", start = 0) +   
 theme\_minimal() +  
 scale\_fill\_brewer(palette = "Blues") +   
 geom\_text(aes(label = sprintf("%s\n(%.1f%%)",   
 scales::comma(avg\_packet\_length),   
 percentage)),  
 position = position\_stack(vjust = 0.5)) +  
 theme(  
 axis.text = element\_text(size = 10),  
 axis.title = element\_blank(),  
 plot.title = element\_text(size = 14, face = "bold"),  
 legend.title = element\_text(size = 12, face = "bold"),  
 legend.text = element\_text(size = 10),  
 panel.grid = element\_blank()  
 ) +  
 labs(  
 title = "Average Packet Length by Attack Type",  
 subtitle = paste("Total observations:",   
 scales::comma(sum(packet\_analysis\_by\_attacktype$n))),  
 fill = "Attack Type"   
 )



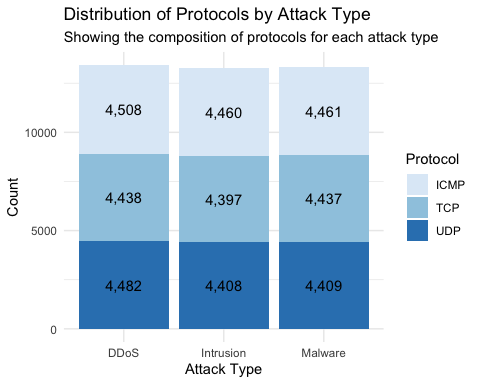
Attack Protocol Relationship

This analysis explores the relationship between attack types and the protocols used in executing cyberattacks.

Protocol Distribution by Attack Type: The stacked bar chart shows the number of protocols for each attack type. This helps identify the TCP, UDP, and ICMP as protocols of interest because of their prevalence with specific attacks such as DDoS, Malware, or Intrusion.

Insights: The protocol composition provides valuable insights for understanding the technical nature of different attack types which helps in tailoring defense mechanisms to safeguard these communication mediums.

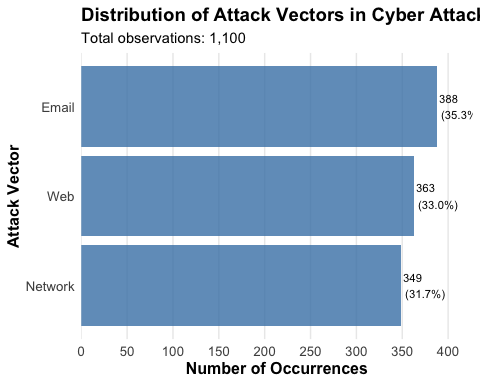
attack\_protocol\_relationship <- cyber\_attacks %>%  
 count(`Attack Type`, Protocol) %>%  
 spread(Protocol, n, fill = 0)  
  
ggplot(attack\_protocol\_relationship %>%   
 gather(Protocol, Count, -`Attack Type`)) +  
 geom\_bar(aes(x = `Attack Type`, y = Count, fill = Protocol),  
 stat = "identity", position = "stack") +  
 theme\_minimal() +  
 scale\_fill\_brewer(palette = "Blues") +  
 labs(  
 title = "Distribution of Protocols by Attack Type",  
 subtitle = "Showing the composition of protocols for each attack type",  
 x = "Attack Type",  
 y = "Count"  
 ) +  
 geom\_text(aes(x = `Attack Type`, y = Count,   
 label = scales::comma(Count), group = Protocol),  
 position = position\_stack(vjust = 0.5))



Top Attack Vectors

This analysis identifies the most common attack vectors in the Cybersecurity Dataset. A bar chart displays the frequency of different attack vectors (e.g., phishing, brute force, malware). The results highlight which attack vectors are most prevalent, providing insights into attacker behavior and common methods of exploitation. Together, these analyses provide actionable insights into the relationship between attack types, protocols, and vectors, helping to shape more effective cybersecurity strategies.

attack\_vector\_analysis <- cyber\_dataset %>%  
 count(`Attack Vector`, sort = TRUE) %>%  
 mutate(percentage = (n/sum(n))\*100)   
  
ggplot(attack\_vector\_analysis, aes(x = reorder(`Attack Vector`, n), y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue", alpha = 0.8) +  
 coord\_flip() +  
 theme\_minimal() +  
 scale\_y\_continuous(  
 labels = scales::comma,  
 breaks = scales::pretty\_breaks(n = 10),   
 expand = expansion(mult = c(0, 0.1))   
 ) +  
 geom\_text(aes(label = sprintf("%s\n(%.1f%%)", scales::comma(n), percentage)),  
 hjust = -0.1,   
 size = 3) +   
 theme(  
 axis.text = element\_text(size = 10),  
 axis.title = element\_text(size = 12, face = "bold"),  
 plot.title = element\_text(size = 14, face = "bold"),  
 panel.grid.major.y = element\_blank(),   
 panel.grid.minor = element\_blank()   
 ) +  
 labs(  
 title = "Distribution of Attack Vectors in Cyber Attacks",  
 subtitle = paste("Total observations:", scales::comma(sum(attack\_vector\_analysis$n))),  
 x = "Attack Vector",  
 y = "Number of Occurrences"  
 )

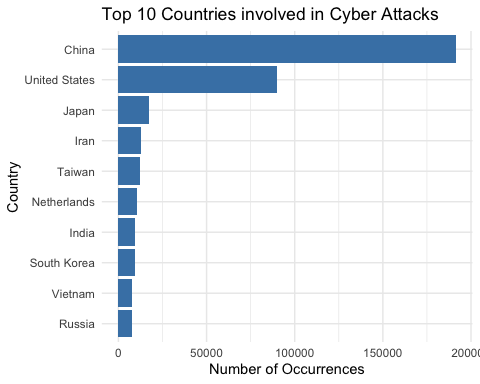


# Top Countries Involved in Cyber Attacks

This analysis identifies the top 10 countries contributing to offensive cyberattacks based on data from the AWS Honeypot and Cybersecurity Dataset:

Key Insights: The bar chart shows the frequency of attacks originating from each country while showcasing a trend on prominent geographical locations involved in cyberattacks. These findings emphasize the need for global collaboration, policy decisions, and defense strategies in order to mitigate cyber threats and strengthen international cybersecurity measures.

top\_countries <- df\_aws\_cyberdataset\_country %>%  
 count(`Country`, sort = TRUE) %>%  
 head(10)  
  
ggplot(top\_countries, aes(x = reorder(`Country`, n), y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 coord\_flip() +  
 theme\_minimal() +  
 labs(  
 title = "Top 10 Countries involved in Cyber Attacks",  
 x = "Country",  
 y = "Number of Occurrences"  
 )



# Text Mining and Sentiment Analysis in Hacker News

Sentiment and text mining were performed on the Hacker News Articles dataset to analyze the sentiment and key themes in article titles. The titles were tokenized into individual words and the analysis was conducted using Bing Lexicon which categorizes words into a positive or negative category.

Key Findings: A bar chart visualizes the top 20 words by sentiment and highlights the most frequently used positive and negative terms in article titles. Words associated with negative sentiment often reflect concerns or warnings, while positive terms suggest opportunities or solutions.

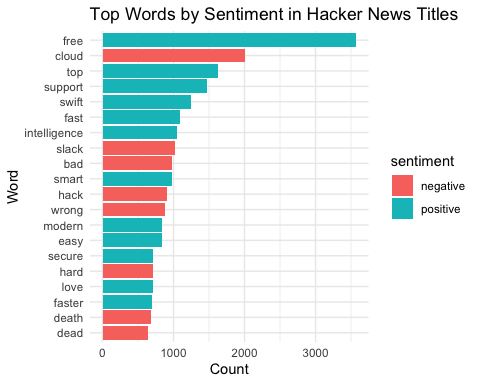
tidy\_titles <- hn\_titles %>%  
 unnest\_tokens(word, title) %>%  
 anti\_join(stop\_words)

## Joining with `by = join\_by(word)`

title\_sentiments <- tidy\_titles %>%  
 inner\_join(get\_sentiments("bing"), relationship = "many-to-many") %>%  
 count(word, sentiment, sort=T)

## Joining with `by = join\_by(word)`

ggplot(head(title\_sentiments, 20),   
 aes(x = reorder(word, n), y = n, fill = sentiment)) +  
 geom\_col() +  
 coord\_flip() +  
 theme\_minimal() +  
 labs(  
 title = "Top Words by Sentiment in Hacker News Titles",  
 x = "Word",  
 y = "Count"  
 )

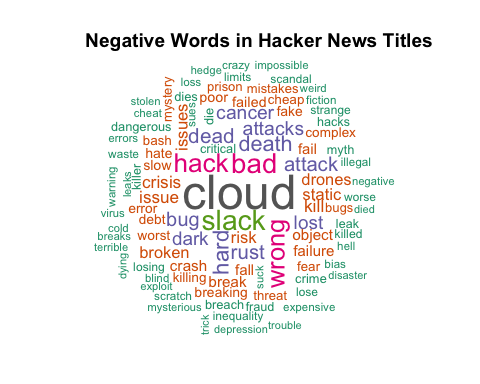


Word Cloud From Negative words found in Sentiment Analysis

A word cloud was generated to visualize the most frequent negative words extracted from sentiment analysis of Hacker News titles. The word cloud highlights terms commonly associated with negative sentiments. Prominent words such as “cloud,” “breach,” and “attack” indicate significant concerns about cloud-based assets and cyber threats.

Insights: The frequent mention of “cloud” suggests that cloud infrastructure is a primary target for cyberattacks as many organizations have moved to cloud based solutions. Other terms in the word cloud point to specific threats and vulnerabilities. This word cloud provides a visual summary of negative sentiment in Hacker News articles which serves as a guide on focus areas for research and defense.

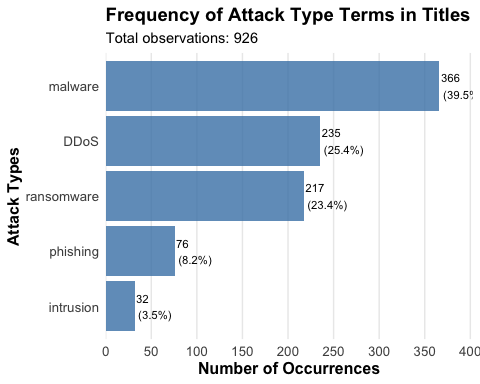
negative\_words <- title\_sentiments %>%  
 filter(sentiment == "negative") %>%  
 arrange(desc(n))  
  
wordcloud(words = negative\_words$word,   
 freq = negative\_words$n,   
 max.words = 100,   
 random.order = FALSE,   
 colors = brewer.pal(8, "Dark2"),   
 scale = c(3, 0.5))   
  
title(main = "Negative Words in Hacker News Titles")



Finding the Attack Type frequency in Hacker News article titles

This analysis showcases a bar chart related to the frequency of key attack types that were determined in the analysis above (e.g., DDoS, Malware, Phishing, Intrusion, Ransomware). The distribution of attack type mentions shows a trend into which cybersecurity issues are most frequent in news worthy articles. The visualization and data provide a clear view of how different attack types are represented on Hacker News.

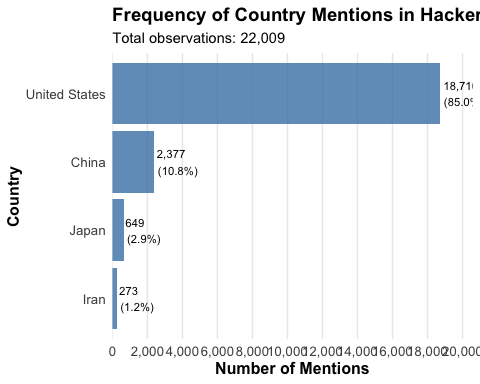
find\_security\_terms <- function(text, term) {  
 sum(grepl(term, text, ignore.case=T))  
}  
  
  
security\_terms <- data.frame(  
 term = c("DDoS", "malware", "phishing", "intrusion", "ransomware"),  
 count = c(  
 find\_security\_terms(hn\_titles$title, "DDoS"),  
 find\_security\_terms(hn\_titles$title, "malware"),  
 find\_security\_terms(hn\_titles$title, "phishing"),  
 find\_security\_terms(hn\_titles$title, "intrusion"),  
 find\_security\_terms(hn\_titles$title, "ransomware")  
 )  
)  
  
security\_terms <- security\_terms %>%  
 mutate(percentage = (count/sum(count))\*100)  
  
ggplot(security\_terms, aes(x = reorder(term, count), y = count)) +  
 geom\_bar(stat = "identity", fill = "steelblue", alpha = 0.8) +  
 coord\_flip() +  
 theme\_minimal() +  
 scale\_y\_continuous(  
 labels = scales::comma,  
 breaks = scales::pretty\_breaks(n = 10),  
 expand = expansion(mult = c(0, 0.1))  
 ) +  
 geom\_text(aes(label = sprintf("%s\n(%.1f%%)", scales::comma(count), percentage)),  
 hjust = -0.1,  
 size = 3) +  
 theme(  
 axis.text = element\_text(size = 10),  
 axis.title = element\_text(size = 12, face = "bold"),  
 plot.title = element\_text(size = 14, face = "bold"),  
 panel.grid.major.y = element\_blank(),  
 panel.grid.minor = element\_blank()  
 ) +  
 labs(  
 title = "Frequency of Attack Type Terms in Titles",  
 subtitle = paste("Total observations:", scales::comma(sum(security\_terms$count))),  
 x = "Attack Types",  
 y = "Number of Occurrences"  
 )



Finding The Frequency of Country Mentions in Hacker News Titles

This analysis examines the frequency of country mentions in Hacker News article titles. The bar chart shows the frequency of mentions for selected countries (e.g., China, United States, Japan, Iran) in the titles. Here we can see that the United States and China are mentioned most frequently. These findings highlight regions that attract the most attention in cybersecurity events and how frequently specific countries are mentioned in relation to cybersecurity topics

find\_country\_terms <- function(text, term) {  
 sum(grepl(term, text, ignore.case=T))  
}  
  
country\_terms <- data.frame(  
 term = c("China", "United States", "Japan", "Iran"),  
 count = c(  
 find\_country\_terms(hn\_titles$title, "China|Chinese"),  
 find\_country\_terms(hn\_titles$title, "United States|USA|U.S.|America|American"),  
 find\_country\_terms(hn\_titles$title, "Japan|Japanese"),  
 find\_country\_terms(hn\_titles$title, "Iran|Iranian")  
 )  
)  
  
country\_terms <- country\_terms %>%  
 mutate(percentage = (count/sum(count))\*100)  
  
ggplot(country\_terms, aes(x = reorder(term, count), y = count)) +  
 geom\_bar(stat = "identity", fill = "steelblue", alpha = 0.8) +  
 coord\_flip() +  
 theme\_minimal() +  
 scale\_y\_continuous(  
 labels = scales::comma,  
 breaks = scales::pretty\_breaks(n = 10),  
 expand = expansion(mult = c(0, 0.1))  
 ) +  
 geom\_text(aes(label = sprintf("%s\n(%.1f%%)", scales::comma(count), percentage)),  
 hjust = -0.1,  
 size = 3) +  
 theme(  
 axis.text = element\_text(size = 10),  
 axis.title = element\_text(size = 12, face = "bold"),  
 plot.title = element\_text(size = 14, face = "bold"),  
 panel.grid.major.y = element\_blank(),  
 panel.grid.minor = element\_blank()  
 ) +  
 labs(  
 title = "Frequency of Country Mentions in Hacker News Titles",  
 subtitle = paste("Total observations:", scales::comma(sum(country\_terms$count))),  
 x = "Country",  
 y = "Number of Mentions"  
 )



# Random Forests and Predicting

A Random Forest model was implemented to predict cyberattack types based on selected features from the cyber\_attacks dataset.

Feature Selection: Variables commonly associated with non-predictive or redundant information (e.g., IP addresses, timestamps, payload data) were excluded so that meaningful predictors like Packet Length, Traffic Type, and Severity Level could be used.

cyber\_data\_test <- cyber\_attacks %>%  
 select(-Timestamp,   
 -`Source IP Address`,   
 -`Destination IP Address`,   
 -`Payload Data`,   
 -`Malware Indicators`,   
 -`Alerts/Warnings`,   
 -`Attack Signature`,   
 -`Action Taken`,   
 -`IDS/IPS Alerts`,   
 -`Proxy Information`,  
 -`Firewall Logs`,   
 -`Geo-location Data`,   
 -`Log Source`,   
 -`Device Information`,   
 -`Network Segment`,   
 -`User Information`) %>%  
 rename(  
 Source\_Port = `Source Port`,   
 Destination\_Port = `Destination Port`,   
 Attack\_Type = `Attack Type`,  
 Packet\_Length = `Packet Length`,  
 Traffic\_Type = `Traffic Type`,  
 Anomaly\_Scores = `Anomaly Scores`,  
 Severity\_Level = `Severity Level`,   
 Packet\_Type = `Packet Type`  
 ) %>%  
 drop\_na()  
  
names(cyber\_data\_test)

## [1] "Source\_Port" "Destination\_Port" "Protocol" "Packet\_Length"   
## [5] "Packet\_Type" "Traffic\_Type" "Anomaly\_Scores" "Attack\_Type"   
## [9] "Severity\_Level"

Random Forest and Predictive Model

Model Implementation: The dataset was split into training (75%) and testing (25%) subsets. The Random Forest model was trained using 1,000 trees (ntree = 1000) and 4 features per split (mtry = 4).

Performance Insights: The Out-of-Bag (OOB) error rate was 67.25%, indicating that predicting attack types from the available data is challenging. This high error rate suggests that cyberattacks may be difficult to tell apart from normal traffic due to them containing similar features or well-masked malicious behavior.

The results highlight how complex it is to predict a cyberattack. This is evidence that there is a need for robust data, advanced modeling techniques, and well trained professionals with the knowledge necessary to implement defense mechanisims.

set.seed(123)  
sample = sample.split(seq(1:nrow(cyber\_data\_test)), SplitRatio = 0.75)  
cyber\_train <- subset(cyber\_data\_test, sample == TRUE)  
cyber\_test <- subset(cyber\_data\_test, sample == FALSE)  
rf\_model <- randomForest(  
 as.factor(Attack\_Type) ~ .,   
 data = cyber\_train,  
 mtry = 4,   
 ntree = 1000   
 )  
  
print(rf\_model)

##   
## Call:  
## randomForest(formula = as.factor(Attack\_Type) ~ ., data = cyber\_train, mtry = 4, ntree = 1000)   
## Type of random forest: classification  
## Number of trees: 1000  
## No. of variables tried at each split: 4  
##   
## OOB estimate of error rate: 67.25%  
## Confusion matrix:  
## DDoS Intrusion Malware class.error  
## DDoS 3388 3291 3359 0.6624826  
## Intrusion 3333 3179 3426 0.6801167  
## Malware 3431 3335 3258 0.6749800

# Conclusion

The rise in sophisticated cyber threats shows the need for more proactive security measures and data-driven decision-making in order to effectively predict, prevent, and respond to attacks.

Threat Categories and Defense Mechanisms: An analysis of the relationship between threat categories and suggested defense mechanisms provides valuable insights into how organizations can address specific threats. Threat categories like malware and DDoS attacks often have defined defense mechanisms that should be incorporated into any organization’s defensive approach.

Final Thoughts for Security Strategies: Organizations must adopt what the industry calls “Defense in Depth” which is a layered defense mechanism that focuses on integrating security via people, processes, and technology at various levels of the organization. By leveraging data analysis, we can refine and prioritize defenses which can significantly enhance the organization’s resilience against evolving cyberattacks.

threatcat\_defensemech\_relationship <- cyber\_dataset %>%   
 count(`Threat Category`, `Suggested Defense Mechanism`) %>%   
 spread(`Suggested Defense Mechanism`, n, fill = 0)  
  
ggplot(threatcat\_defensemech\_relationship %>%   
 gather(`Suggested Defense Mechanism`, Count, -`Threat Category`)) +  
 geom\_bar(aes(x = `Threat Category`, y = Count,   
 fill = `Suggested Defense Mechanism`),  
 stat = "identity",   
 position = "stack") +  
 theme\_minimal() +  
 scale\_fill\_brewer(palette = "Blues") +  
 labs(  
 title = "Distribution of Defense Mechanisms by Threat Category",  
 subtitle = "Showing suggested defenses for each threat category",  
 x = "Threat Category",  
 y = "Count"  
 ) +  
 geom\_text(aes(x = `Threat Category`, y = Count,   
 label = scales::comma(Count),   
 group = `Suggested Defense Mechanism`),  
 position = position\_stack(vjust = 0.5)) +  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 legend.position = "right"  
 )

