Analysis On Cybersecurity Attacks

Huiting Wu

2024-11-12

```
# devtools::install_github("JLSteenwyk/ggpubfigs")
```

```
library(dplyr)
library(ggplot2)
library(here)
library(readr)
library(tidyverse)
library(gtsummary)
library(wesanderson)
library(ggpubfigs)
library(maps)
library(mapdata)
library(tidygeocoder)
library(mapproj)
library(viridis)
```

Part 1: Data Cleaning

```
cyber_original <- read_csv(here("cybersecurity_attacks.csv"))
```

```
# select the important variables only
cyber <- cyber_original |>
  select(Timestamp,
         `Attack Type`,
         `Severity Level`,
         `Action Taken`,
         `Geo-location Data`,
         `Device Information`) |>
  mutate_if(is.character, as.factor)

# Create a new variable for device type (Apple and Non-Apple)
cyber_device <- cyber |>
  mutate(Device = ifelse(
    grepl("Windows|Android|Linux", `Device Information`),
    "Non-Apple Device",
    ifelse(grepl("Mac|iPad|iPhone|iPod", `Device Information`),
           "Apple Device",
           "Others")))
```

Part 2: Exploratory Data Analysis

Missing Values: There is no missing values

```
knitr::kable(apply(cyber, 2, function(x) sum(is.na(x))))
```

	x
Timestamp	0
Attack Type	0
Severity Level	0
Action Taken	0
Geo-location Data	0
Device Information	0

```
# or

cyber |>
  summarise(across(everything(), ~ sum(is.na(.)))) |>
    pivot_longer(cols = everything(),
                 names_to = "Variable",
                 values_to = "Missing count")
```

```
## # A tibble: 6 x 2
##   Variable           `Missing count`
##   <chr>                        <int>
## 1 Timestamp                        0
## 2 Attack Type                      0
## 3 Severity Level                   0
## 4 Action Taken                     0
## 5 Geo-location Data                0
## 6 Device Information               0
```

```
#assigning names to columns
```

Summary Table

```
cyber_device |>
  select(`Attack Type`, `Severity Level`, `Action Taken`, Device) |>
  tbl_summary(by = Device)
```

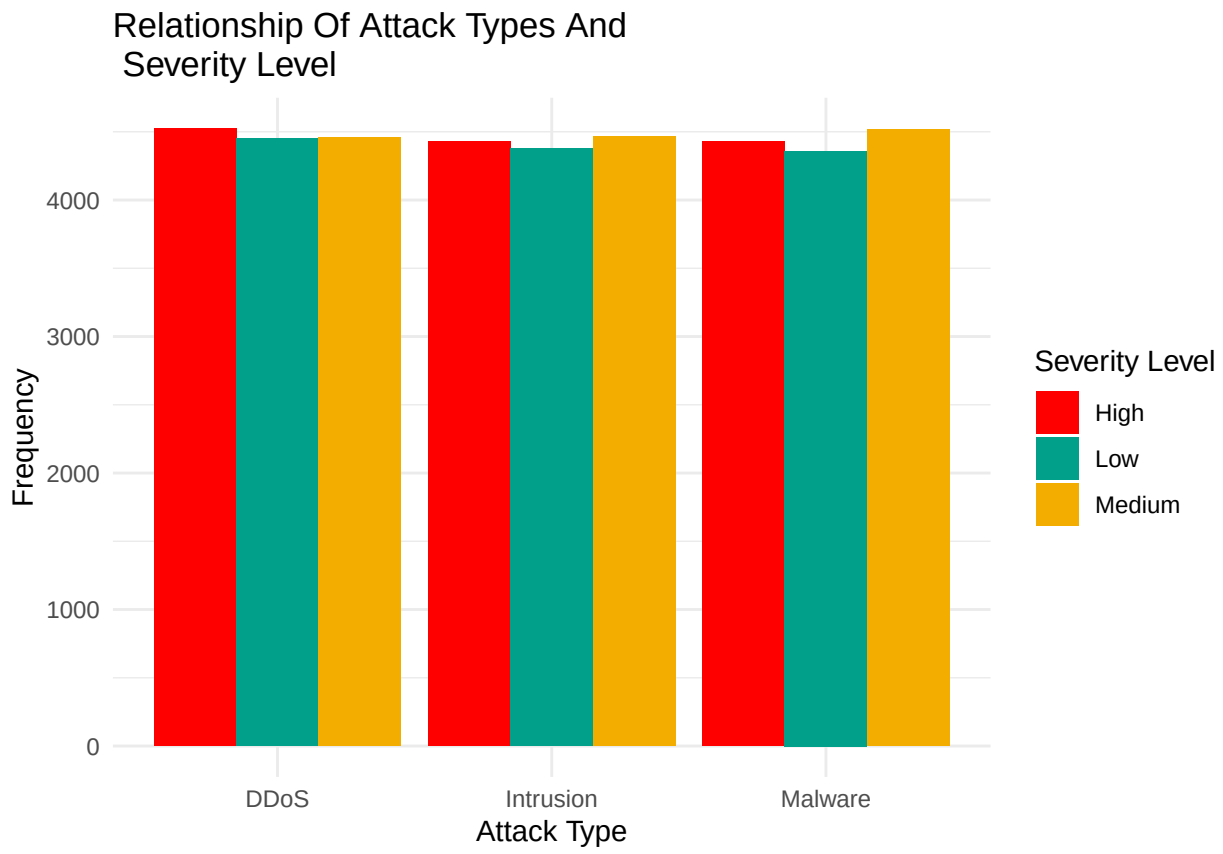
Plots

- **Relationship Of Attack Types And Severity Level**

Characteristic	**Apple Device** N = 11,587[1]	**Non-Apple Device** N = 28,413[1]
Attack Type		
DDoS	3,838 (33%)	9,590 (34%)
Intrusion	3,902 (34%)	9,363 (33%)
Malware	3,847 (33%)	9,460 (33%)
Severity Level		
High	3,894 (34%)	9,488 (33%)
Low	3,825 (33%)	9,358 (33%)
Medium	3,868 (33%)	9,567 (34%)
Action Taken		
Blocked	3,926 (34%)	9,603 (34%)
Ignored	3,832 (33%)	9,444 (33%)
Logged	3,829 (33%)	9,366 (33%)

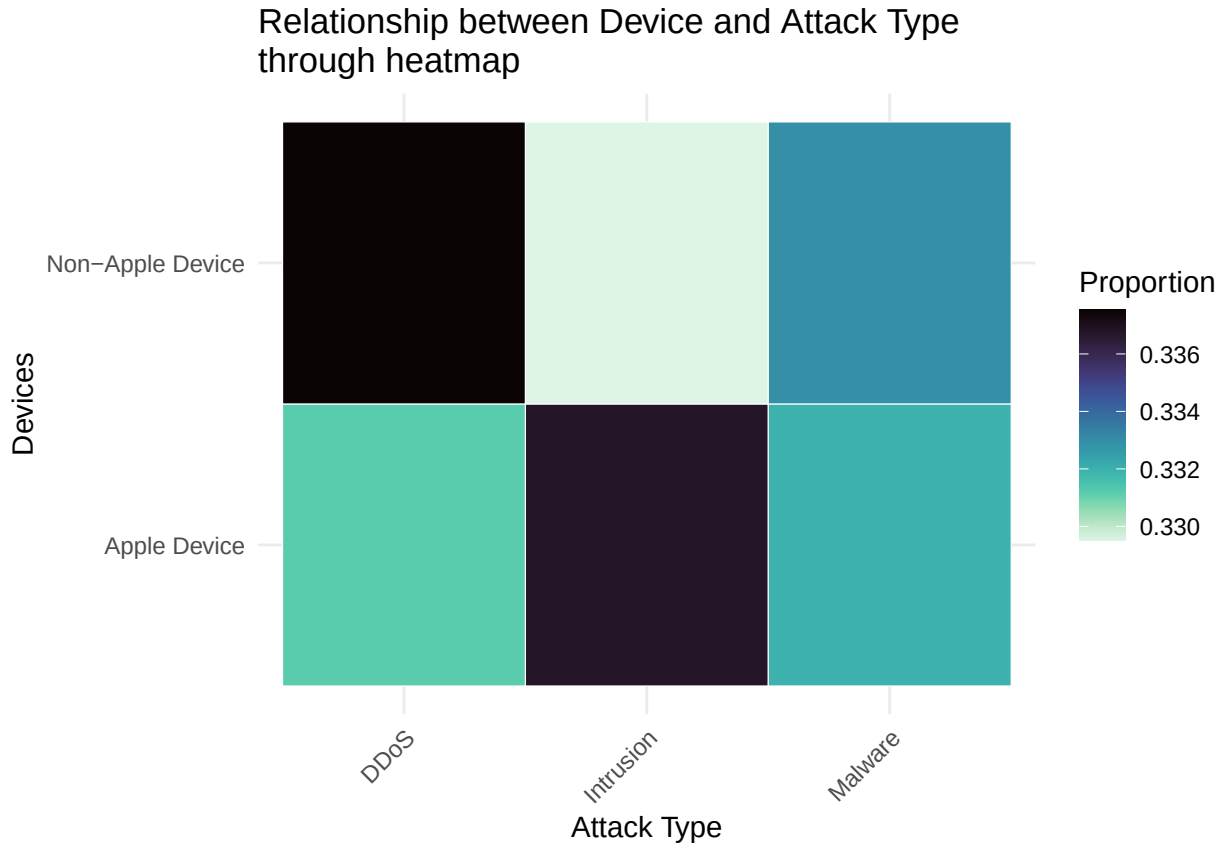
[1] n (%)

```
cyber |> ggplot(aes(x = `Attack Type`, fill = `Severity Level`)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = wes_palette("Darjeeling1")) +
  labs(title = "Relationship Of Attack Types And \n Severity Level",
       y = "Frequency") +
  theme_minimal()
```

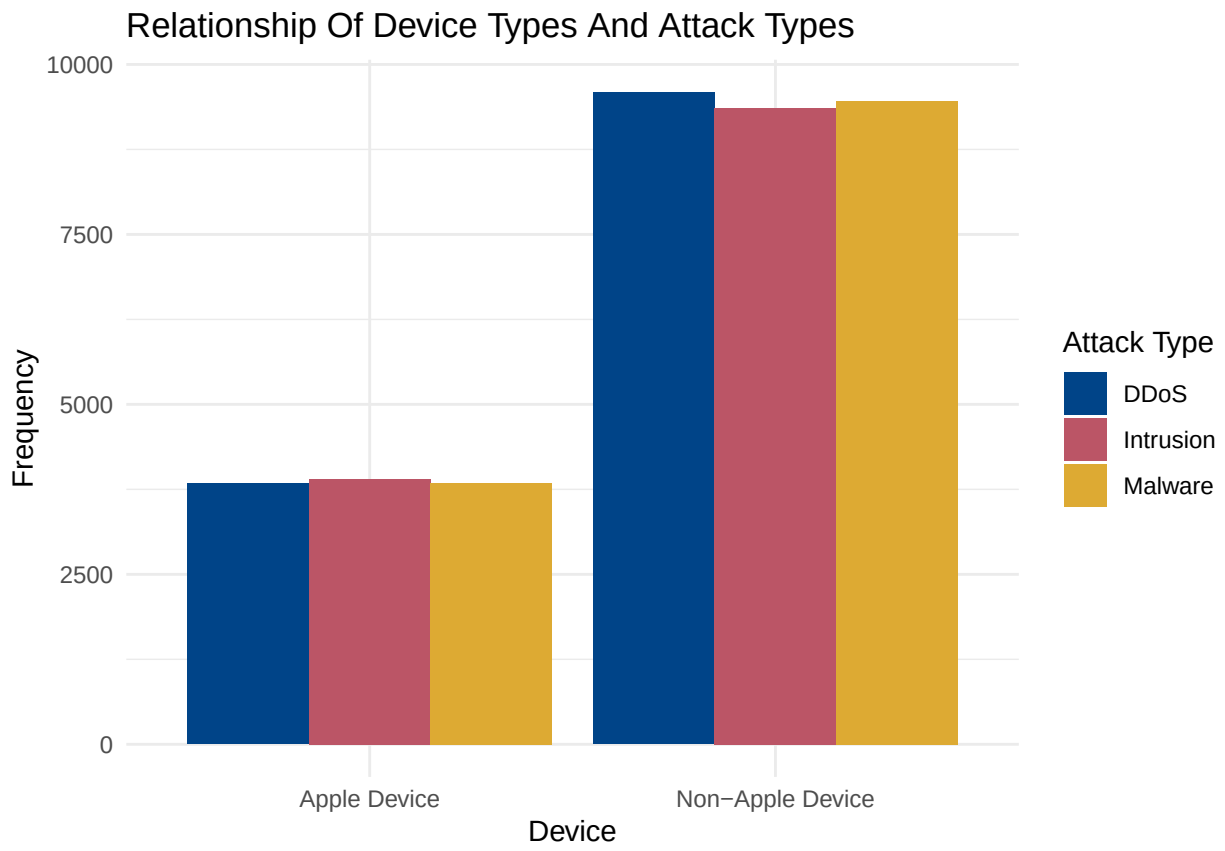


- **Relationship Of Device Types And Attack Types**

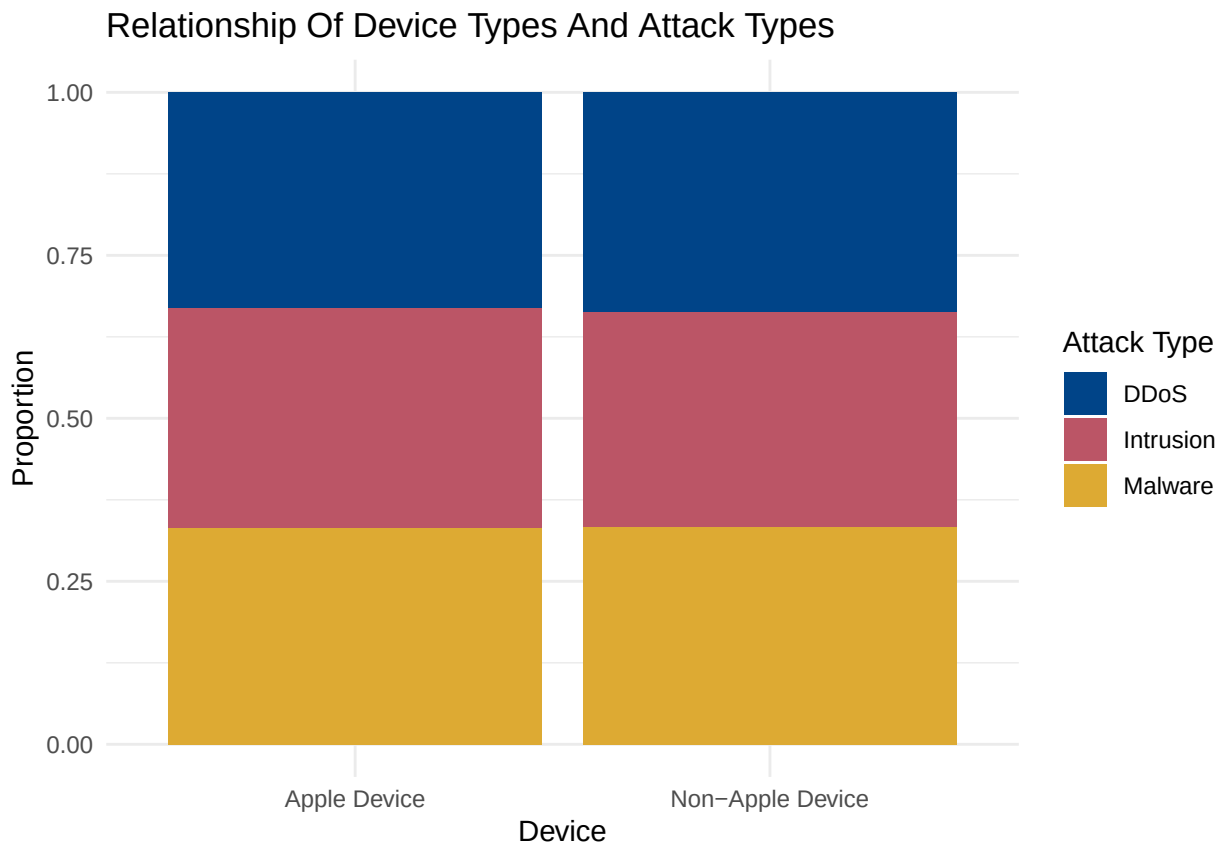
```
prop_cyber <- cyber_device |>
  group_by(Device, `Attack Type`) |>
  summarise(Count = n()) |>
  mutate(Proportion = Count / sum (Count)) |>
  ungroup ()

ggplot(prop_cyber, aes(x = `Attack Type`, y = Device , fill = Proportion)) +
  geom_tile(color = "white") +
  scale_fill_viridis(alpha = 1, begin = 1, end = 0, direction = 1, discrete = FALSE,
                     option = "G", aesthetics = "fill") +
  labs (
    title = "Relationship between Device and Attack Type \nthrough heatmap",
    x = "Attack Type",
    y="Devices",
    fill = "Proportion"
  )+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



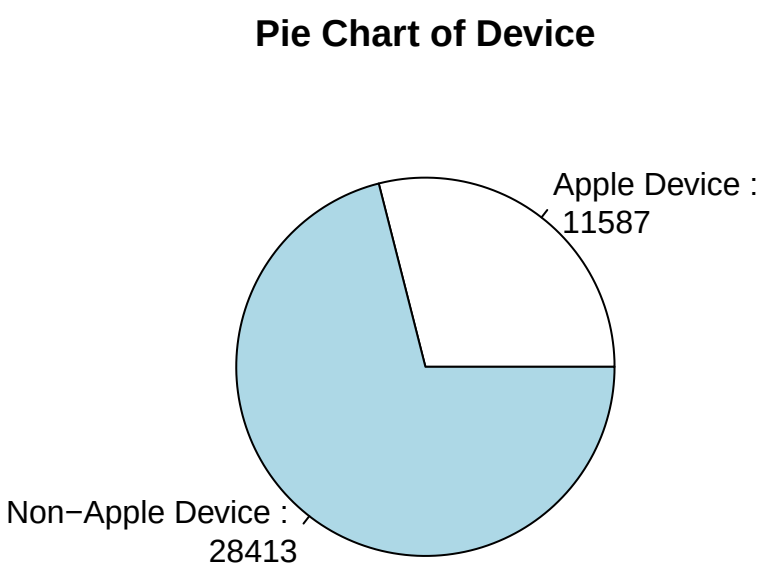

```
cyber_device |> ggplot(aes(x = Device, fill = `Attack Type`)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = friendly_pal("contrast_three")) +
  labs(title = "Relationship Of Device Types And Attack Types",
       y = "Frequency") +
  theme_minimal()
```




```
cyber_device |> ggplot(aes(x = Device, fill = `Attack Type`)) +
  geom_bar(position = "fill") +
  scale_fill_manual(values = friendly_pal("contrast_three")) +
  labs(title = "Relationship Of Device Types And Attack Types",
       y = "Proportion") +
  theme_minimal()
```

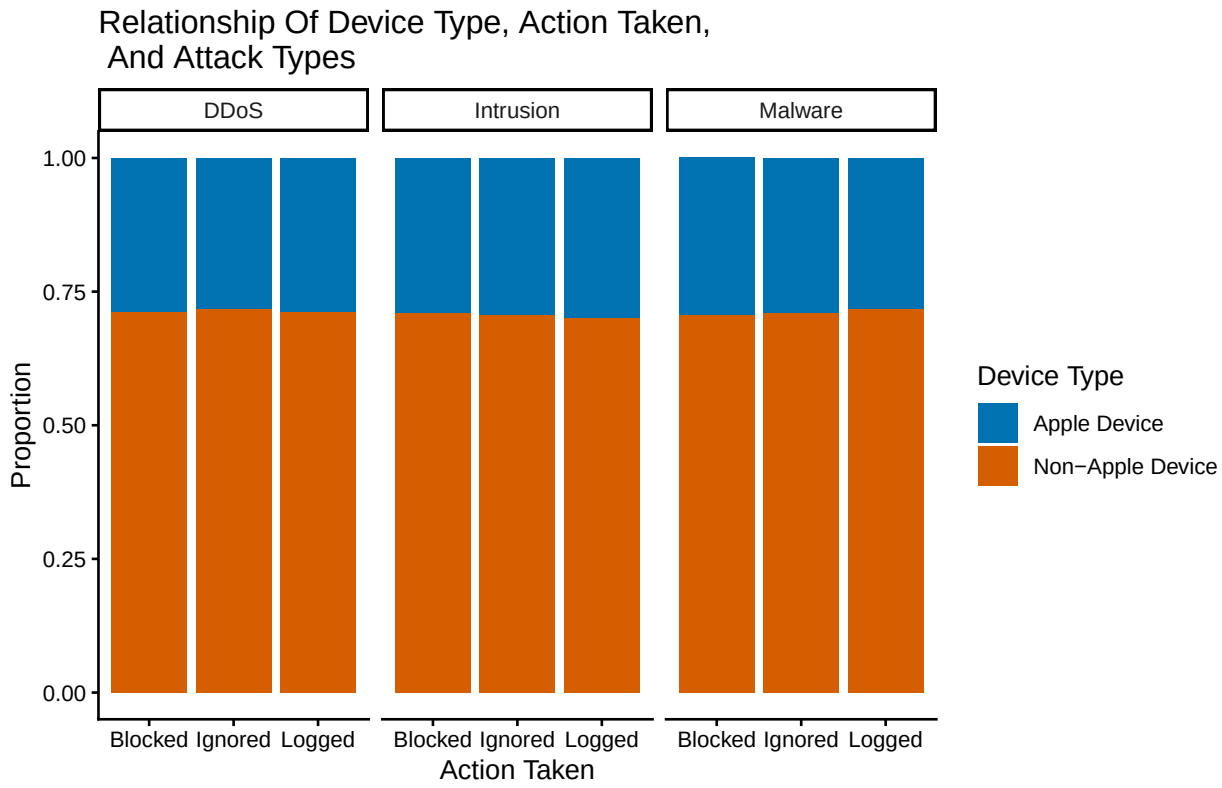



```r
# Unbalance Data set
pie(table(cyber_device$Device),
    main = "Pie Chart of Device",
    label = paste(names(table(cyber_device$Device)), ": \n", table(cyber_device$Device)))
```



- **Relationship Of Device Types, Action Taken, and Attack Type**

```
cyber_device |> ggplot(aes(fill = Device, x = `Action Taken`)) +
  geom_bar(position = "fill") +
  facet_wrap(~ `Attack Type`) +
  scale_fill_manual(values = friendly_pal("ito_seven")) +
  labs(title = "Relationship Of Device Type, Action Taken, \n And Attack Types",
       y = "Proportion",
       fill = "Device Type") +
  theme_classic()
```



- **Map Of States Of India With Number Of Attacks**

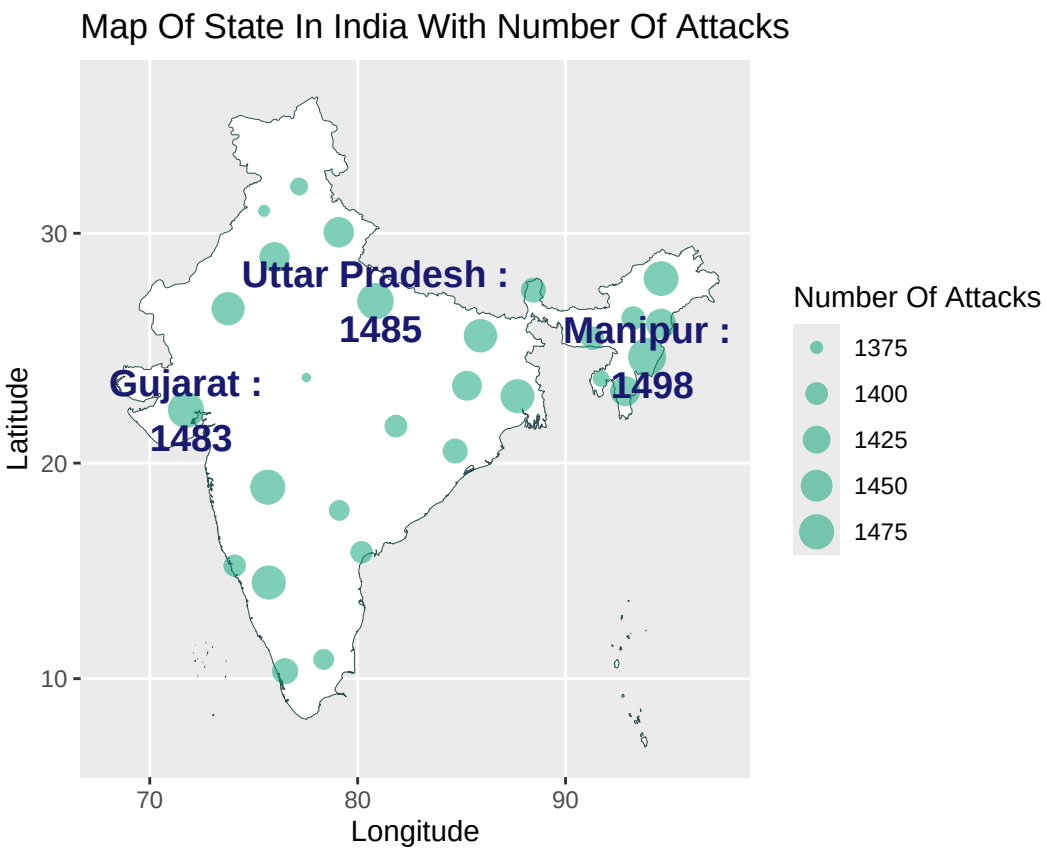
```
# use sub() to extract the state name from the geom location,
# count the total attacks for the state
locations <- cyber |>
  mutate(Location = sub(".*,", "", `Geo-location Data`)) |>
  group_by(Location) |>
  summarise(count = n()) |>
  arrange(desc(count))
```

```
# use tidygeocoder packages to create the longitudes and latitudes
# for the selected cities by OpenStreetMap
locations <- locations |>
  geocode(address = Location, method = "osm")
```

```
top_3_locations <- head(locations, 3)
```

```
# create an India map with maps and mapdata package
india_map <- map_data("worldHires", "India")
ggplot() +
  geom_polygon(data = india_map, aes(x = long, y = lat, group = group),
               fill = "white", color = "#2F4F4F", lwd = 0.1) +
  geom_point(data = locations,
             aes(x = long, y = lat, size = count),
             color = "#009E73", alpha = 0.5) +
  geom_text(data = top_3_locations,
```

```
            aes(x = long, y = lat,
                label = paste(Location, ": \n", count)),
            col = "#191970", size = 5, fontface = "bold") +
coord_map() +
labs(title = "Map Of State In India With Number Of Attacks",
     x = "Longitude",
     y = "Latitude",
     size = "Number Of Attacks")
```



Part 3: Data Analysis

Q1: Does the severity levels and attack types are associated?

Step 1 : Hypothesis

H_0: Attack type and severity level are independent.

H_A: Attack type and severity level are associated.

Step 2: Check conditions

1. Independence: Sample is generated randomly through AI.
2. Normality : All counts are greater than 5 so both conditions are satisfied.

```r
table (cyber_device$`Attack Type`, cyber_device$`Severity Level`)
```

```
##
##             High  Low Medium
##   DDoS      4523 4450   4455
##   Intrusion 4427 4374   4464
##   Malware   4432 4359   4516
```

```r
cyber_tab <- matrix(c(4523,4450,4455,
                      4427,4374,4464,
                      4432,4359,4516),
                    nrow = 3,
                    byrow = T)
colnames(cyber_tab) <- c("High", "Low", "Medium")
rownames (cyber_tab) <- c("DDoS", "Intrusion", "Malware")


test2 <- chisq.test(cyber_tab)
cat("Expected Counts : ", test2$expected)
```

```
## Expected Counts :  4492.337 4437.806 4451.857 4425.533 4371.812 4385.655 4510.13 4455.382 4469.489
```

Step 3: Test Statistic

The test statistic is 1.797.

Step 4: p-value

The p-value is 0.773.

Step 5: Decision

Fail to reject H_0.

Conclusion: We have enough evidence that there is no association between attack type and severity level.

Q2: Do Non-Apple devices have the more high severity level attacks than Apple devices?

Step 1:

Group 1: Proportion of Non-Apple Devices with cyber attacks high sever

Group 2: Proportion of Apple Devices

$H_0 : p_1 - p_2 = 0$

$H_A : p_1 - p_2 > 0$

```r
Apple <- cyber_device |> filter(Device == "Apple Device")

Non_Apple <- cyber_device |> filter(Device == "Non-Apple Device")

Apple_high <- Apple |> filter(`Severity Level` == "High")

Non_Apple_high <- Non_Apple |> filter(`Severity Level` == "High")
```

Step 2: check conditions

1. Independence: The data is a random sample generated by algorithm.
2. Large Sample Size:

```
x2 <- nrow(Apple_high)
x1 <- nrow(Non_Apple_high)

n2 <- nrow(Apple)
n1 <- nrow(Non_Apple)

p1 <- x1/n1
p2 <- x2/n2

n1*p1
```

```
## [1] 9488
```

```
n2*p2
```

```
## [1] 3894
```

```
n1*(1-p1)
```

```
## [1] 18925
```

```
n2*(1-p2)
```

```
## [1] 7693
```

All greater than 5.

```
pro_test <- prop.test(c(x1, x2), c(n1, n2), alternative = "greater", correct = FALSE)
pro_test1 <- prop.test(c(x1, x2), c(n1, n2), correct = FALSE)
pro_test$p.value
```

```
## [1] 0.659258
```

```
pro_test1
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 0.16846, df = 1, p-value = 0.6815
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.012334865  0.008065605
## sample estimates:
##    prop 1    prop 2
## 0.3339317 0.3360663
```

Step 3: Test Statistic

The test statistic is 0.168.

Step 4: P-value

The p-value is 0.659.

Step 5: Decision

p-value is greater than 0.05. We fail to reject the null hypothesis.

Conclusion: We have no enough evidence that not Apple devices have the more high severity level attacks than Apple devices.

Confidence Interval: The 95% confidence interval of $p_1 - p_2$ is -0.0123349, 0.0080656.

Q3: Does the action taken by Apple devices and non-Apple devices to the high severity level attacks associated?

Step 1:

H_0 : The device types and action taken to the high severity level attacks are independence.

H_A : The device types and action taken to the high severity level attacks are associated.

```
high_severity <- cyber_device |> filter(`Severity Level` == "High")

Device_vs_Action <- table(high_severity$Device, high_severity$`Action Taken`)

chi_test <- chisq.test(Device_vs_Action)
```

Step 2: check conditions

1. Independence: The data is a random sample generated by algorithm.
2. Expected Counts:

```
chi_test$expected
```

```
##
##                     Blocked  Ignored   Logged
##   Apple Device     1318.175 1297.806 1278.019
##   Non-Apple Device 3211.825 3162.194 3113.981
```

All expected values are greater than 5 and the data is a random sample. The conditions satisfied.

Step 3: Test Statistic

The test statistic is 2.884.

Step 4: p-value

The p-value is 0.236.

Step 5: Decision

p-value is greater than 0.05. We fail to reject the null hypothesis.

Conclusion: We have no enough evidence that the device types and action taken to the high severity level attacks are associated.

Part 4: Discuss the Results

References:

https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf

https://www.kaggle.com/datasets/teamincribo/cyber-security-attacks

https://incribo.com/

https://pmc.ncbi.nlm.nih.gov/articles/PMC9920136/#sec6-sensors-23-01231

package: wesanderson or ggbugfigs

These are a handful of color palettes that are color blind friendly.

https://github.com/karthik/wesanderson

https://github.com/JLSteenwyk/ggpubfigs

Method to due with unbalanced dataset:

https://github.com/dalpozz/unbalanced

https://rpubs.com/DeclanStockdale/799284

https://en.wikipedia.org/wiki/List_of_states_and_union_territories_of_India_by_population