Final Project

Group 10: Huiting Wu, Sharmeen Kapporwala

2024-11-12

```
library(dplyr)
library(ggplot2)
library(here)
library(readr)
library(tidyverse)
library(gtsummary)
library(wesanderson)   #for color blind
library(ggpubfigs)    #for color blind
library(maps)
library(mapdata)
library(tidygeocoder)   #to create map
library(mapproj)
library(viridis)   #for color blind
```

```
cyber_original <- read_csv("cybersecurity_attacks.csv")   #retrieving Dataset
```

Part 1: Data Cleaning

extracting 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) #changing categorical variables to factor
```

Renaming the columns for better understanding

```
cyber <- cyber |>
  rename(Attack_Type = `Attack Type`,
         Severity_level =  `Severity Level`,
         Action_taken = `Action Taken`,
         Geo_location = `Geo-location Data`,
         Device_Information = `Device Information`
         )

#colnames(cyber)
```

Create a new variable for device type (Apple and Non-Apple)

```
cyber_device <- cyber |>
  mutate(Device = ifelse(
    #extracting words from string
    grepl("Windows|Android|Linux", Device_Information),
    "Non-Apple Device",
    ifelse(grepl("Mac|iPad|iPhone|iPod", Device_Information),
           "Apple Device",
           "Others"))) # assign them into Apple or non-Apple device
```

Part 2: Exploratory Data Analysis

Missing Values: There is no missing values in the selected columns

```
cyber |>
  summarise(across(everything(), ~ sum(is.na(.)))) |>
  #adding all counts of a single column
    pivot_longer(cols = everything(),
                 names_to = "Variable",
                 values_to = "Missing count")   #assigning names to columns
```

```
## # 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                     0
## 6 Device_Information               0
```

Summary Table

```
cyber_device |>
  select(Attack_Type, Action_taken, Severity_level, Device) |>
  tbl_summary(by = Device)   #to get proportion grouped by device type
```

Plots

package: wesanderson or ggbugfigs

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

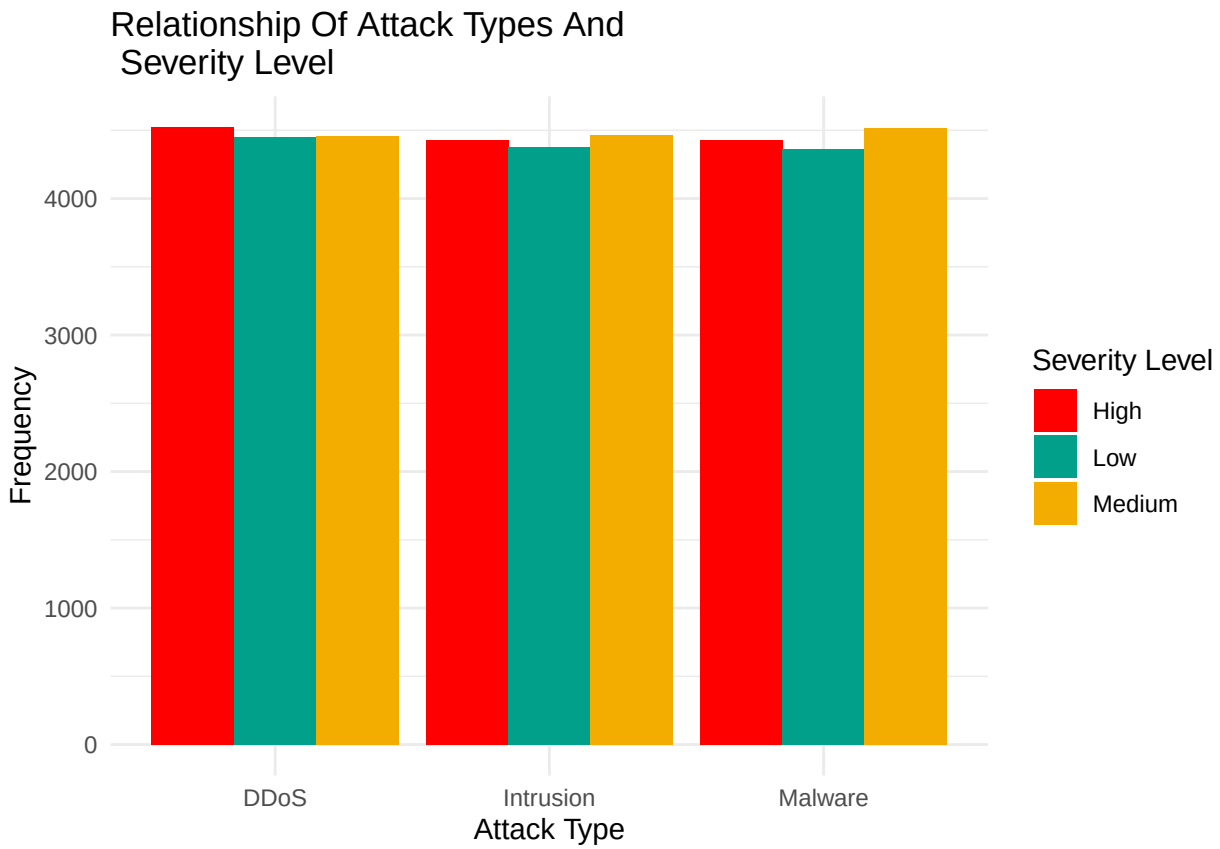
- **Relationship Of Attack Types And Severity Level**

```
#bar plot to show relationship

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",
       x = "Attack Type",
       fill = "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%)
Action_taken		
Blocked	3,926 (34%)	9,603 (34%)
Ignored	3,832 (33%)	9,444 (33%)
Logged	3,829 (33%)	9,366 (33%)
Severity_level		
High	3,894 (34%)	9,488 (33%)
Low	3,825 (33%)	9,358 (33%)
Medium	3,868 (33%)	9,567 (34%)

[1] n (%)

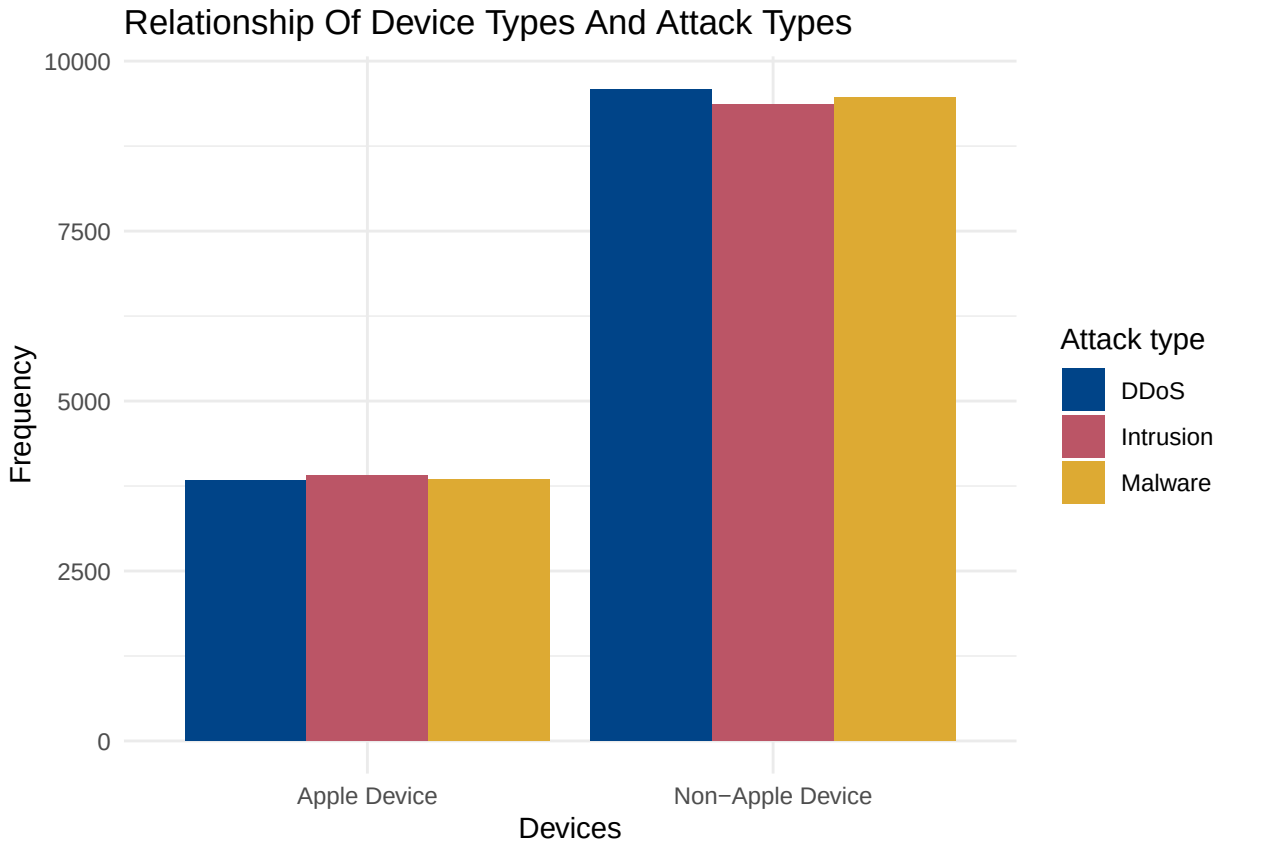
```
      y = "Frequency") +
  theme_minimal()
```



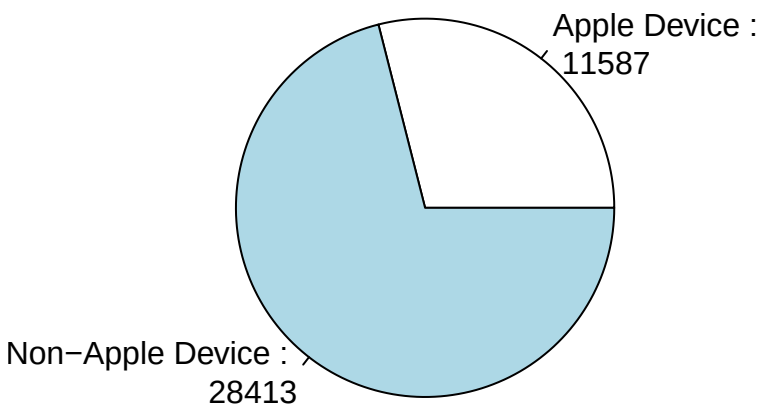
- **Relationship Of Device Types And Attack Types**

```
#bar plot to show relationship w.r.t frequency
```

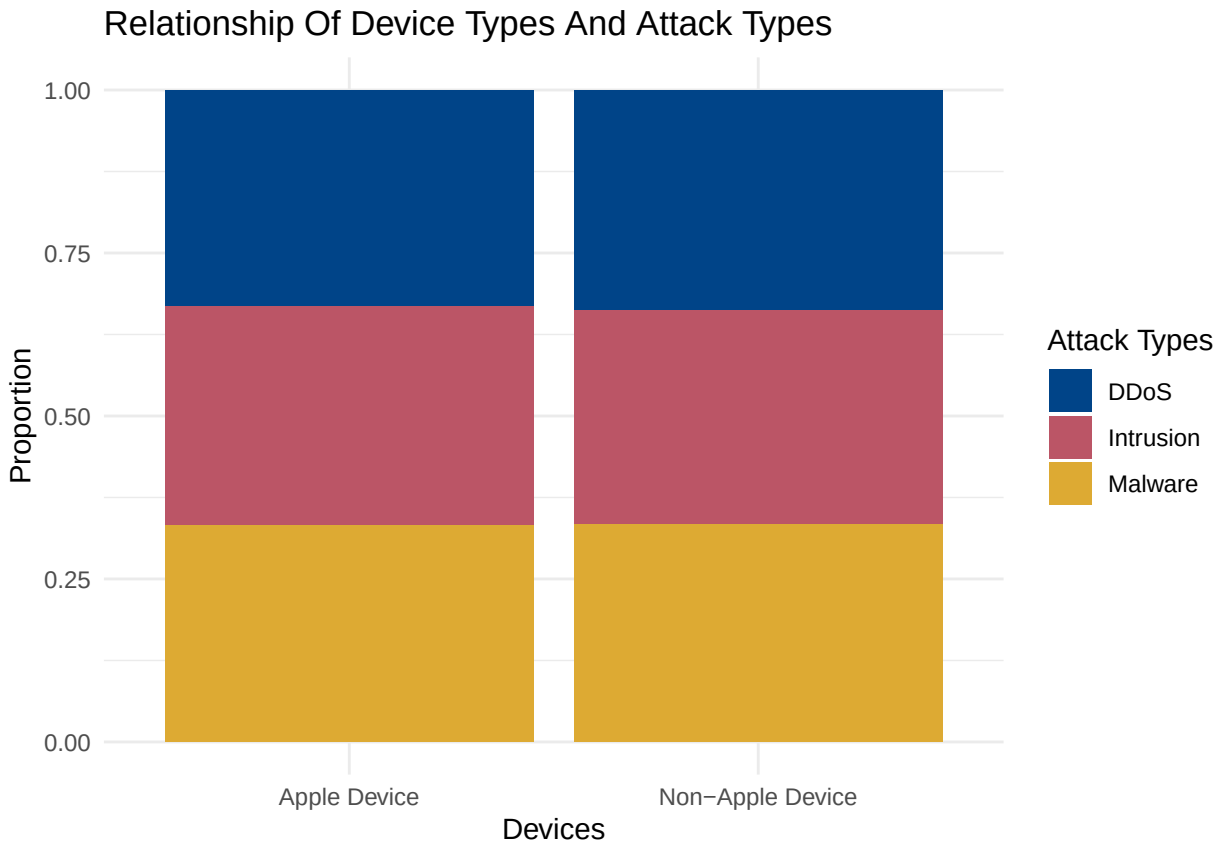
```r
cyber_device |> ggplot(aes(x = Device, fill = Attack_Type)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = friendly_pal("contrast_three")) +
  labs(title = "Relationship Of Device Types And Attack Types",
       fill = "Attack type",
       x = "Devices",
       y = "Frequency")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



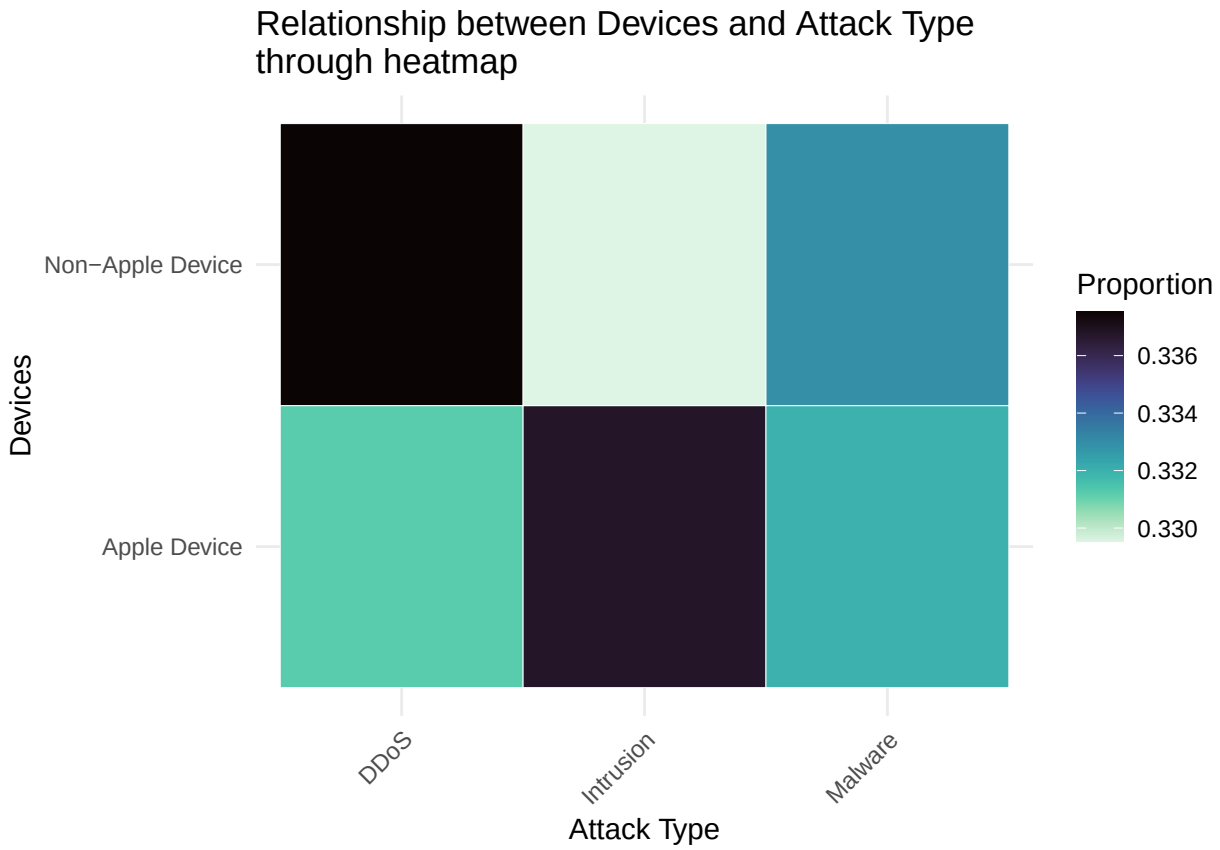

```r
# Unbalance Data set

pie(table(cyber_device$Device),
    main = "Pie Chart of Device Types",
    label = paste(names(table(cyber_device$Device)), ": \n",
                  table(cyber_device$Device)))
```




```
#bar plot to show relationship w.r.t proportion


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",
       fill = "Attack Types",
       x = "Devices",
       y = "Proportion") +
  theme_minimal()
```

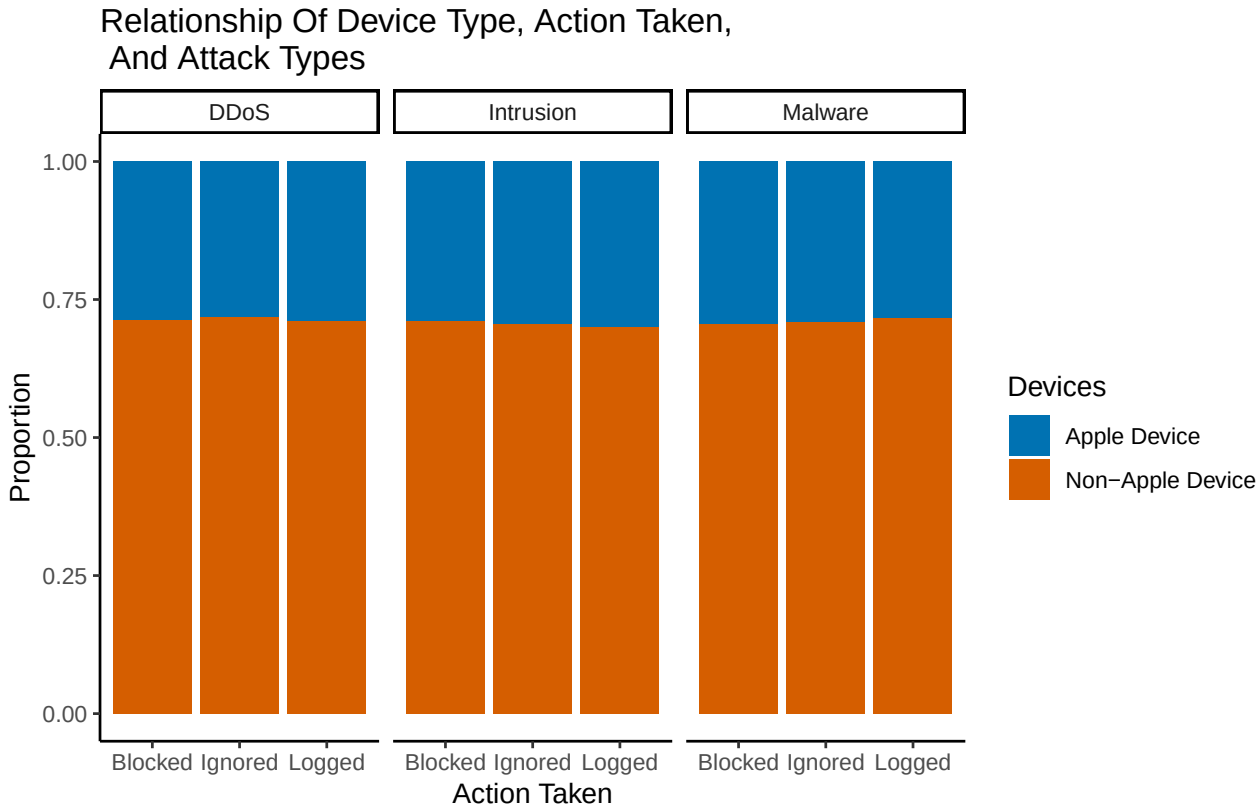



```
#showing relationship through heatmap to get better
# visualization of proportions of attacks

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") +
          #used viridis for color blind friendly
  labs (
    title = "Relationship between Devices and Attack Type \nthrough heatmap",
    x = "Attack Type",
    y="Devices",
    fill = "Proportion"
  )+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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

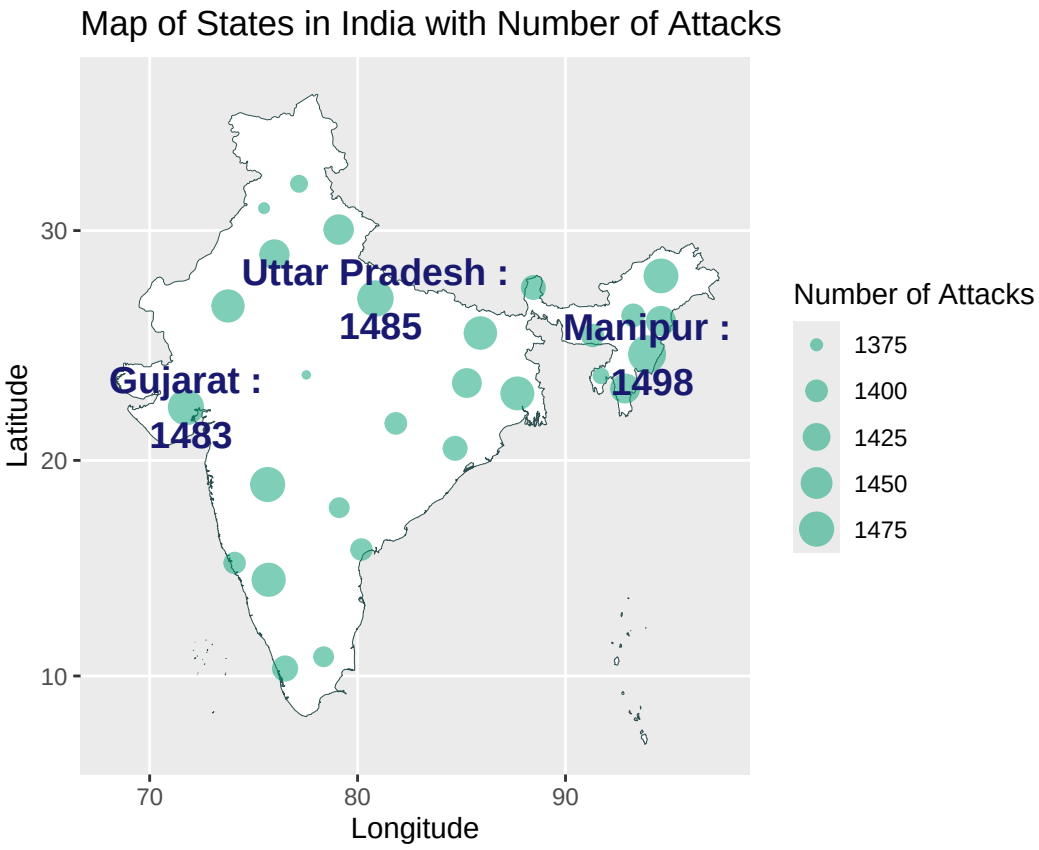
```
#barplot w.r.t proportion

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",
       x = "Action Taken",
       fill = "Devices") +
  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)) |>
  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) #select the top 3 states

# 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) +
  # create a shape of India
  geom_point(data = locations,
             aes(x = long, y = lat, size = count),
             color = "#009E73", alpha = 0.5) +
  # add the attack numbers to the plot by location
  geom_text(data = top_3_locations,
```

```
          aes(x = long, y = lat,
              label = paste(Location, ": \n", count)),
          col = "#191970", size = 5, fontface = "bold") +
# label the state names and the number of attacks
coord_map() +
labs(title = "Map of States 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?

Performing chi- squared test for independence of attack type and severity level

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 synthetic data
2. Expected counts

```r
#table (cyber_device$Attack_Type, cyber_device$Severity_level)

cat ("Observed Counts : \n")
```

```
## Observed Counts :
```

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

cyber_tab
```

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

```r
cat ("\n\n")
```

```r
chi1_test<- chisq.test(cyber_tab)
cat("Expected Counts : ", chi1_test$expected)
```

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

All counts are greater than 5

Step 3 : *Test Statistics*

```r
chi1_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  cyber_tab
## X-squared = 1.7971, df = 4, p-value = 0.773
```

```r
cat("The test statistic is " , round (chi1_test$statistic,3))
```

```
## The test statistic is  1.797
```

Step 4 : $P - value$

```r
cat ("The p-value is ", round(chi1_test$p.value,3))
```

```
## The p-value is  0.773
```

Step 5 : *Decision*

Decision : fail to reject H_0

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

Q2: Does Non-Apple devices have the greater proportion of high severity level attacks than Apple devices?

Performing Two Proportion Hypothesis Test for devices with high severity level attacks

Step 1 : *Hypothesis*

Group 1: Proportion of Non-Apple Devices with attacks of high severity level

Group 2: Proportion of Apple Devices with attacks of high severity level

$H_0 : p_1 - p_2 = 0$

$H_A : p_1 - p_2 > 0$

```
#apple devices
Apple <- cyber_device |> filter(Device == "Apple Device")

#non-apple devices
Non_Apple <- cyber_device |> filter(Device == "Non-Apple Device")

#apple devices with high severity level attacks
Apple_high <- Apple |> filter(Severity_level == "High")

#non apple devices with high severity level attacks
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.

both conditions are satisfied

```
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
```

```
##
##  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 = 0.6593
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.01069494  1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.3339317 0.3360663
```

Step 3 : *Test Statistics*

```
cat("The test statistic is " , round (pro_test$statistic,3))
```

```
## The test statistic is  0.168
```

Step 4 : $P - value$

```
cat ("The p-value is ", round(pro_test$p.value,3))
```

```
## The p-value is  0.659
```

Step 5 : *Decision*

Decision: Fail to reject the H_0.

Conclusion: We have no enough evidence that Non Apple devices have the more high severity level attacks than Apple devices.

Confidence Interval :

```
cat("The 95% confidence interval of p1 - p2 is",pro_test1$conf.int)
```

```
## The 95% confidence interval of p1 - p2 is -0.01233487 0.008065605
```

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

Step 1 : *Hypothesis*

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*

Decision : Fail to reject the H_0

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