Names:

John Haag

Nick Staub

Sean Edwards

Zach Kovalenko

Kingsley Okereke

Question 1: Is there a relationship between number of breaches and breach month and year? if so is it statistically significant?

Graph to answer the question:

A screenshot of a cell phone

Description automatically generated

Source Code:

#count number of occurances for each month adn store it in its own tibble

count\_by\_date <- CIS\_435\_project\_data %>%

group\_by(month=floor\_date(Date\_of\_Breach, "month")) %>%

tally()

#first plot that shows breaches over time

ggplot(count\_by\_date, aes(x=count\_by\_date$month, y=count\_by\_date$n)) +

geom\_line() +

ggtitle("Breaches Over Time")+

ylab("Number of Breaches")+

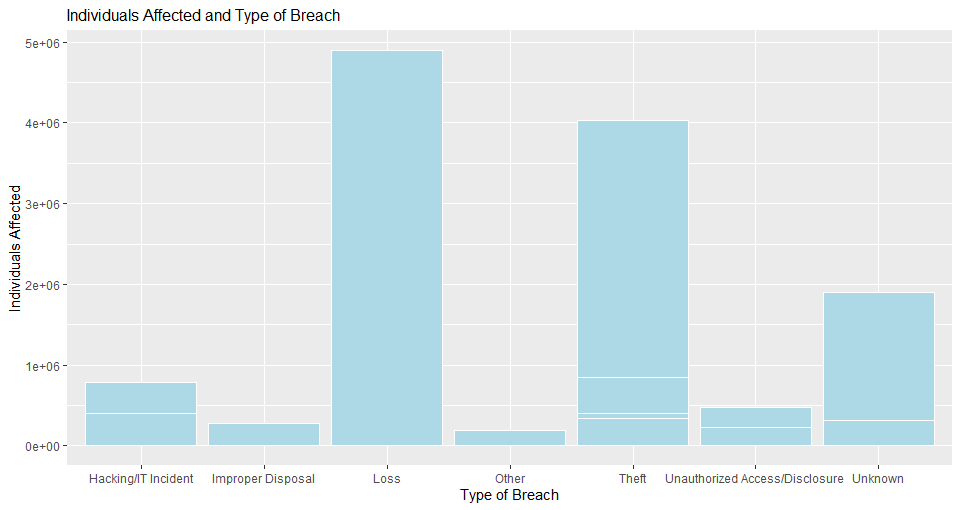
xlab("Years")

Process Description:

Before actually making the plot, the data had to be separated by months. A new tibble was created to do that. After this the breaches over time graph could be created. After observing the graph it does seem like some outliers could be removes. It is clear there were some breaches before 2007 but not very many as certain tech was probably not as advanced and as popular as in future years. It seems like removing these outliers and getting a better view of the main set of data would be the best solution.

Question 5: Is there a correlation between the type of breach and the number of individuals affected?

Graph to answer the question:



Source Code:

typeEffect <-CIS\_435\_project\_data[CIS\_435\_project\_data$Individuals\_Affected >= 185000,]

ggplot(typeEffect, aes(x=typeEffect$Type\_of\_Breach,y=typeEffect$Individuals\_Affected)) +

geom\_bar(stat = "identity",fill="lightblue",colour ="white",position = "dodge") +

xlab("Type of Breach") + ylab("Individuals Affected") +

ggtitle("Individuals Affected and Type of Breach") +

theme(plot.title = element\_text(size = 12))

Process Description:

The first step taken was to create a bar graph displaying the Types of Breach and the number of individuals affected. Every breach in the data could not be displayed without the graph becoming unreadable so a limit had to be set to maintain readability. Out of the known types of breaches the thefts and losses had a significantly larger impact on individuals than all of the other known types of breaches.