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")+

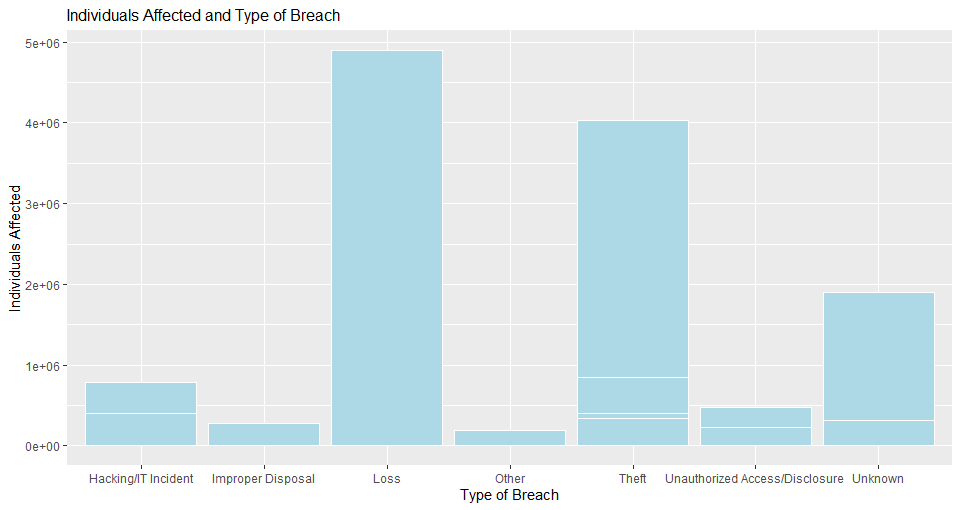
xlab("Years")

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") +

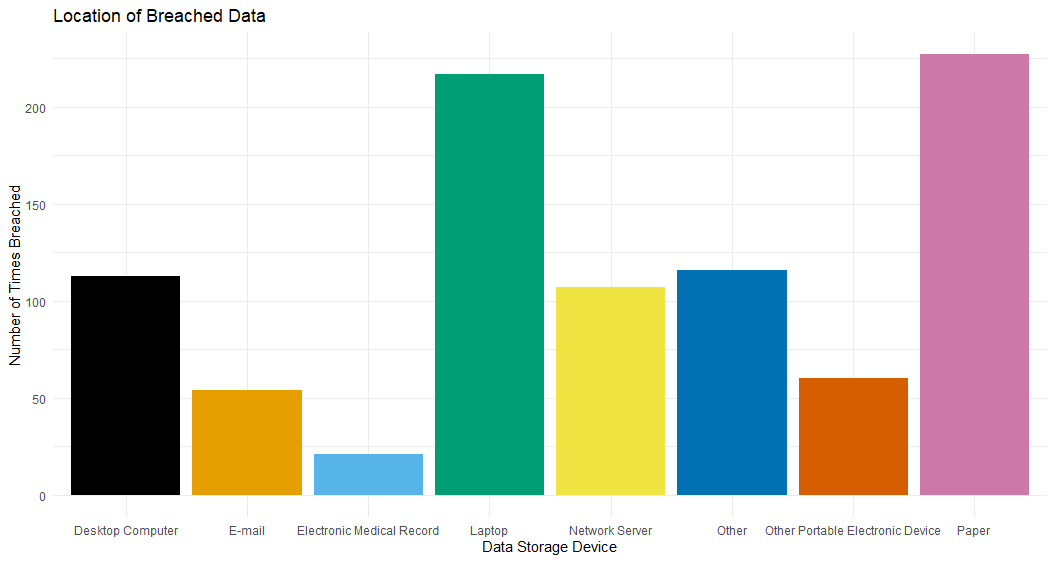
theme(plot.title = element\_text(size = 12))

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.

Question 2 (Zach Kovalenko): Are certain devices more susceptible to data breaches according to the data given?

Graph to answer the question:



Source Code:

#Read in the Data

df <- read.csv("CIS\_435\_project\_data-1.csv", header = T)

#Create a subset that will exclude all the rows that have multiple different Locations of Breached Information

#(If this is not done the Data is unreadable and it is not possible to determine which information storage spot is the worst)

df2 <- subset(df, Location\_of\_Breached\_Information == "Laptop"

| Location\_of\_Breached\_Information == "Desktop Computer"

| Location\_of\_Breached\_Information == "Other"

| Location\_of\_Breached\_Information == "Other Portable Electronic Device"

| Location\_of\_Breached\_Information == "E-mail"

| Location\_of\_Breached\_Information == "Paper"

| Location\_of\_Breached\_Information == "Network Server"

| Location\_of\_Breached\_Information == "Electronic Medical Record")

#color palette

cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

ggplot(df2, aes(Location\_of\_Breached\_Information)) +

geom\_bar(fill = cbbPalette) +

ggtitle("Location of Breached Data") +

xlab("Data Storage Device") +

ylab("Number of Times Breached") +

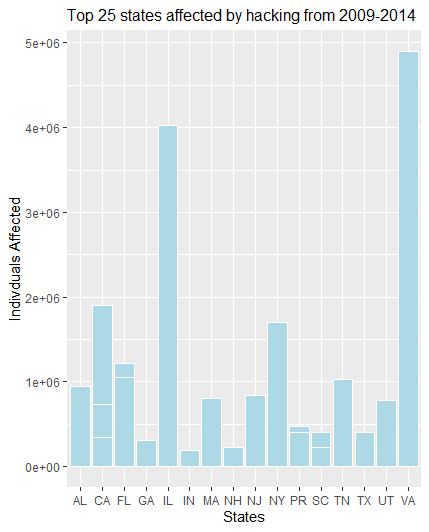
theme\_minimal()

Process Description:

After reading in the data a subset was created. This was done because in some of the breaches in the original data file there were multiple data storage places that were breached. For the purpose of figuring out which device was more susceptible to data breaches I took out all of the data that had multiple points of a breach. Also, the data was unreadable when multiple points of breach were added to the bar chart. Out of the different data storage types that were breached there were two that were breached the most. They were paper and laptop. Those two data storage devices were breached a lot more than any other data storage device.

Question 3(Nickolas Staub): Which states have the largest number of individuals affected by data breaches? What type of data breach was dominant?

The ways I tried and analyzed the dataset is to construct a visual bar graph and base the top 25 states affected by hacking from 2009-2015 and in each bar for each state each hacking attempted is grouped for each individual incident. I then clean the data further by taking care of all things related to subcategories for theft improper disposal and so and changed the categories to just those types. Finally, I used the summary on the data for the type of breach to give a list of categories and how many of each of those cases occur in the time period of 2009 to 2014. The things I will still have to do is to figure out what the trend is for the entire dataset and not just the top 25 states and to see if there are different patterns in the states with lower affected individuals.



Hacking/IT Incident

76

Improper Disposal

39

Loss

98

Other

91

Theft

571

Unauthorized Access/Disclosure

168

Unknown

12

Source Code:

top25 <-Projectdata[Projectdata$Individuals\_Affected >= 185000,]

ggplot(top25, aes(x=top25$State,y=top25$Individuals\_Affected)) + geom\_bar(stat = "identity",fill="lightblue",colour ="white",position = "dodge") + xlab("States") + ylab("Indivduals Affected") + ggtitle("Top 25 states affected by hacking from 2009-2014") + theme(plot.title = element\_text(size = 12))

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Loss, Unauthorized Access/Disclosure"] <- "Loss"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Loss, Improper Disposal"] <- "Loss

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach %in% c("Theft, Hacking/IT Incident","Theft, Improper Disposal, Unauthorized Access/Disclosure","Theft, Loss","Theft, Loss, Improper Disposal","Theft, Loss, Other","Theft, Loss, Unauthorized Access/Disclosure, Unknown","Theft, Other","Theft, Unauthorized Access/Disclosure","Theft, Unauthorized Access/Disclosure, Hacking/IT Incident"," Theft, Unauthorized Access/Disclosure, Other")] <- "Theft"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Theft, Unauthorized Access/Disclosure, Other"] <- "Theft"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach %in% c("Unauthorized Access/Disclosure, Hacking/IT Incident","Unauthorized Access/Disclosure, Hacking/IT Incident, Other","Unauthorized Access/Disclosure, Other")] <- "Unauthorized Access/Disclosure"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Unknown, Other"] <- "Unknown"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Hacking/IT Incident, Other"] <- "Hacking/IT Incident"

"Improper Disposal, Unauthorized Access/Disclosure"] <- "Improper Disposal"

summary(Projectdata$Type\_of\_Breach)

Due to technical difficulties with my computer I (John) could not answer question 4 due to my concerns within the email I sent to you (Dr.Q) so I did a regression model instead.

regressionModel <-ggplot(mod, mapping = aes( x=CIS\_435\_project\_data$Individuals\_Affected, y= CIS\_435\_project\_data$X1, colour = "#CC0000" )) + geom\_point() +

labs( title = "Regression Model" ,x = " Individuals Affected", y = "Total number of Breaches") + geom\_smooth()

regressionModel

