FinalProject

Jordan Donovan

April 4, 2017

# Final Project - Breaches

library(readr)  
library(tidyverse)

## Loading tidyverse: ggplot2  
## Loading tidyverse: tibble  
## Loading tidyverse: tidyr  
## Loading tidyverse: purrr  
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats  
## lag(): dplyr, stats

require(dplyr)  
require(lubridate)

## Loading required package: lubridate

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

breaches <- read\_csv("~/Downloads/breaches.csv")

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:  
## cols(  
## X1 = col\_integer(),  
## Number = col\_integer(),  
## Name\_of\_Covered\_Entity = col\_character(),  
## State = col\_character(),  
## Business\_Associate\_Involved = col\_character(),  
## Individuals\_Affected = col\_integer(),  
## Date\_of\_Breach = col\_character(),  
## Type\_of\_Breach = col\_character(),  
## Location\_of\_Breached\_Information = col\_character(),  
## Date\_Posted\_or\_Updated = col\_date(format = ""),  
## Summary = col\_character(),  
## breach\_start = col\_date(format = ""),  
## breach\_end = col\_date(format = ""),  
## year = col\_integer()  
## )

View(breaches)

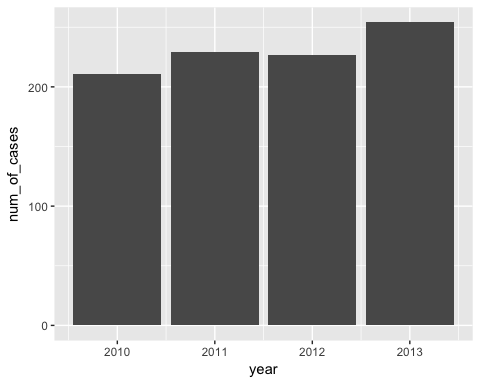
The dataset we will explore invludes data on cyber security breaches from September 2009 to June 2014. This data comes from the U.S. Department of Health and Human Services. A report is required to be submitted if more than five hundred individuals are affected by a breach; thus, all of these breaches involve five hundred or more individuals. We will cover what measures were taken to tidy the dataset before exploring the dataset as it pertains to several of the aspects included in the dataset.

Tidying this dataset was simple since, for the most part it had was already a tidy dataset. I removed Date\_of\_Breach and year from the dataset as there was already a breach\_start and breach\_end, so these columns were redundant to those. We remove these below.

breaches <- breaches %>% select(X1, Number, Name\_of\_Covered\_Entity, State, Business\_Associate\_Involved, Individuals\_Affected, Type\_of\_Breach, Location\_of\_Breached\_Information, Date\_Posted\_or\_Updated, Summary, breach\_start, breach\_end)

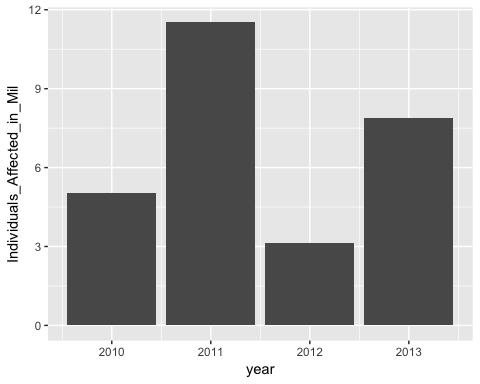
We will start by examining the data as it relates to date. We begin by observing the number of cases per year of the data. I filter out the years in which the data was collected over the entire year. Since 2009 and 2014 were the years where the data was only collected for a portion of the year these are our cut-offs. We can look at this data below.

#Date  
  
breaches %>% mutate(year=year(breach\_start)) %>%  
 filter(year > 2009 & year < 2014) %>%  
 group\_by(year) %>%  
 summarise(num\_of\_cases = n()) %>%  
 ggplot(mapping=aes(x=year, y=num\_of\_cases)) + geom\_col()



We can see from this data that the number of occurrences of cyber security breaches was steadily rising from 2010 to 2013 but not by a large margin. Also, does this mean that the number of individuals affected by cyber security breaches will match up to this? Will 2013 be the year with the most number of individuals affected by cyber security breaches since it was the year with the most number of breaches? This leads us to our next graph where we observe the individuals affected in millions over the four years.

breaches %>% mutate(year=year(breach\_start)) %>%  
 filter(year > 2009 & year < 2014) %>%  
 group\_by(year) %>%  
 summarise(Individuals\_Affected\_in\_Mil=sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping=aes(x=year, y=Individuals\_Affected\_in\_Mil)) +  
 geom\_col()



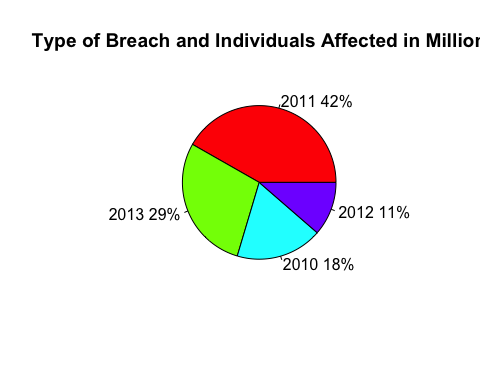
#geom\_smooth()

We can see from this graph that 2013 was not the year in which the most individuals were affected. 2011 seems to be the year with the most individuals affected by breaches. This year, as seen from the first graph, was the second highest in number of occurrences. This data would seem to suggest that the individuals affected per breach has some slight variation. Let's look at this in a pie chart with percentages to get a better idea of the spread of individuals affected over the 4 years.

#Top 5 years of breaches according to individuals affected by breach  
top\_years\_of\_breach\_by\_Individuals\_Affected\_for\_pie <- breaches %>%  
 mutate(year=year(breach\_start)) %>%  
 filter(year > 2009 & year < 2014) %>%  
 group\_by(year) %>%  
 summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 arrange(desc(Individuals\_Affected\_in\_mil)) %>%  
 top\_n(5)

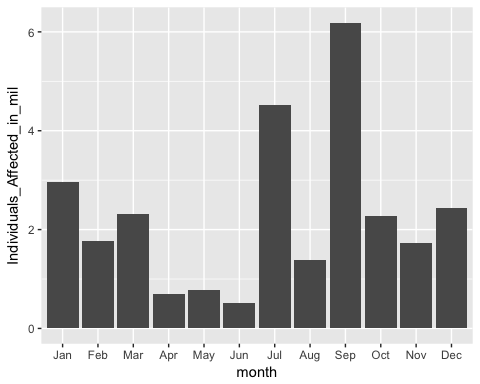
## Selecting by Individuals\_Affected\_in\_mil

slices <- top\_years\_of\_breach\_by\_Individuals\_Affected\_for\_pie$Individuals\_Affected\_in\_mil  
lbls <- top\_years\_of\_breach\_by\_Individuals\_Affected\_for\_pie$year  
pct <- round(slices/sum(slices) \* 100)  
lbls <- paste(lbls, pct)  
lbls <- paste(lbls, "%", sep="")  
 pie(slices, labels=lbls, col=rainbow(length(lbls)), main="Type of Breach and Individuals Affected in Million")



The above graph shows us that 2011 account for over 40% of the individuals affected by breaches over these four years. This is a large number of the overall individuals affected. After this, I wondered if the spread over a year was distributed evenly within each month. Below is a graph that groups the individuals affected by month, so that we can see if the distribution is even over a year's time.

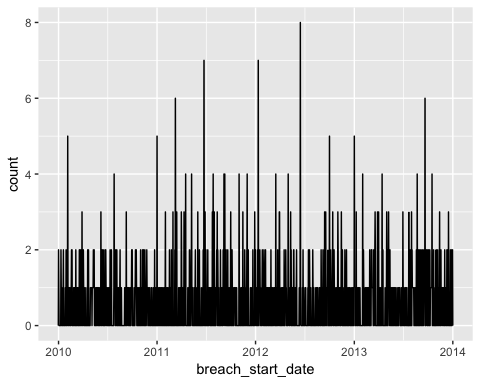
require(lubridate)  
  
breaches\_grouped\_by\_month\_Individuals\_Affected <- breaches %>% mutate(year=year(breach\_start)) %>%   
 filter(!is.na(breach\_start) & year < 2014 & year > 2009) %>%  
 mutate(month = month(breach\_start, label=TRUE)) %>%  
 group\_by(month) %>%  
 summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 ggplot(aes(x=month, y=Individuals\_Affected\_in\_mil)) + geom\_col()  
breaches\_grouped\_by\_month\_Individuals\_Affected



From the above graph, it looks as though April, May, and June are the months where the least amount of individuals were affected by breaches and that July and September are the months where the most individuals were affected. This is interesting. I wonder why that might be.

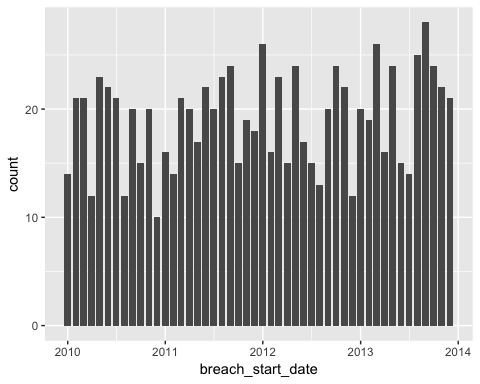
I wanted to see if there was anything that stuck out over the time period of these breaches that could suggest why the individuals affected was spread out the way it is over a year's time as well as over the time of this data's collection. We look at the number of occurrences per day over this time period below.

require(lubridate)  
  
make\_date\_100 <- function(year, month, day) {  
 make\_date(year, month, day)  
}  
  
breaches\_by\_day\_and\_num\_of\_occurrences <- breaches %>% mutate(year=year(breach\_start)) %>%  
 filter(!is.na(breach\_start) & year < 2014 & year > 2009) %>%  
 mutate(month = month(breach\_start, label = TRUE), day = day(breach\_start), breach\_start\_date = make\_date(year, month, day)) %>%  
 ggplot(aes(breach\_start\_date)) + geom\_freqpoly(binwidth = month(1) %>% as.numeric())  
breaches\_by\_day\_and\_num\_of\_occurrences



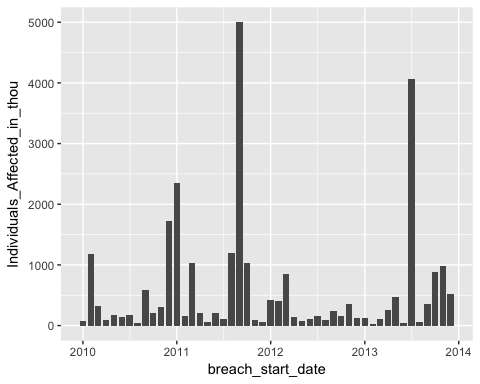
We can see here that, for the most part, there is either one or zero breaches on most days. We can also see that there are some days that have two, four, and even eight breaches on one day in 2012. This is odd. 2012 seems to have a high number of occurrences, but it also has the least number of individuals affected. This, again, seems to suggest that there is an inconsistency in the number of individuals affected by each breach. I wanted to see the above data grouped together by month to see if this suggested another direction or trend.

require(lubridate)  
  
make\_date\_100 <- function(year, month) {  
 make\_date(year, month, day)  
}  
  
breaches\_by\_month\_and\_num\_of\_occurrences <- breaches %>% mutate(year=year(breach\_start)) %>%  
 filter(!is.na(breach\_start) & year < 2014 & year > 2009) %>%  
 mutate(month = month(breach\_start, label = TRUE), day = day(breach\_start), breach\_start\_date = make\_date(year, month)) %>%  
 ggplot(aes(breach\_start\_date)) + geom\_bar()  
breaches\_by\_month\_and\_num\_of\_occurrences



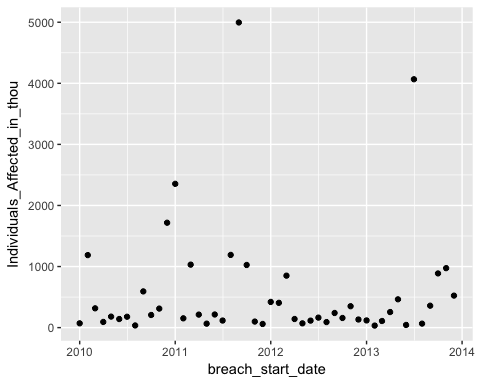
From this graph there appears to be several ups and downs to the number of occurrences, and it seems that the trend is slightly upward which is what we saw in the first graph observing number of occurrences of breaches per year. We have confirmed that metric with another graph. Let's look at the number of individuals affected broken down by month over this four year time span.

require(lubridate)  
  
make\_date\_100 <- function(year, month, day) {  
 make\_date(year, month)  
}  
  
breaches\_by\_month\_and\_individuals\_affected <- breaches %>% mutate(year=year(breach\_start)) %>%  
 filter(!is.na(breach\_start) & year < 2014 & year > 2009) %>%  
 mutate(month = month(breach\_start, label = TRUE), day = day(breach\_start), breach\_start\_date = make\_date(year, month)) %>%  
 group\_by(breach\_start\_date) %>%  
 summarise(Individuals\_Affected\_in\_thou = sum(Individuals\_Affected/1000)) %>%  
 ggplot(aes(x = breach\_start\_date, y = Individuals\_Affected\_in\_thou)) + geom\_col()  
breaches\_by\_month\_and\_individuals\_affected



Wow! We can see from this graph that there are a few very large spikes in the individuals affected that can probably be attributed to a single or possibly a couple of breaches within each of those months. We have already seen data suggesting that the number of individuals affected by each breach varies widely, so this assumption is not far-fetched. What does a model say about the trend of individuals affected over this four year span? Is it trending positively or negatively? We can see this from the following graph.

require(lubridate)  
  
make\_date\_100 <- function(year, month, day) {  
 make\_date(year, month)  
}  
  
breaches\_by\_date\_and\_individuals\_affected <- breaches %>% mutate(year=year(breach\_start)) %>%  
 filter(!is.na(breach\_start) & year < 2014 & year > 2009) %>%  
 mutate(month = month(breach\_start, label=TRUE), day=day(breach\_start), breach\_start\_date = make\_date(year, month)) %>%  
 group\_by(breach\_start\_date) %>%  
 summarise(Individuals\_Affected\_in\_thou = sum(Individuals\_Affected/1000))  
  
models <- tibble(  
 a0 = runif(250, -2000, 4000),  
 a1 = runif(250, -5000, 5000)  
 )  
  
plot1 <- ggplot(breaches\_by\_date\_and\_individuals\_affected, aes(breach\_start\_date, Individuals\_Affected\_in\_thou)) +  
 geom\_abline(aes(intercept = a0, slope=a1), data = models, alpha = 1/4) + geom\_point()  
plot1



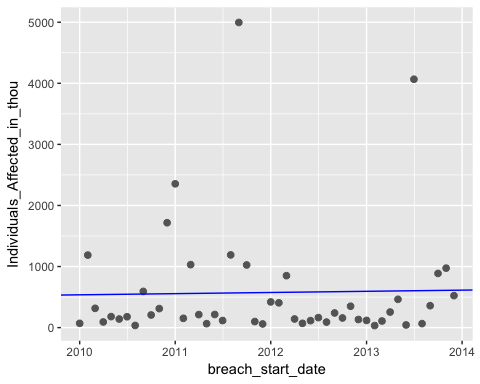
dataset\_mod <- lm(Individuals\_Affected\_in\_thou ~breach\_start\_date, data = breaches\_by\_date\_and\_individuals\_affected)  
 coef(dataset\_mod) %>% str()

## Named num [1:2] -230.8379 0.0526  
## - attr(\*, "names")= chr [1:2] "(Intercept)" "breach\_start\_date"

plot2 <- summary(dataset\_mod)$r.squared \* 100  
 plot2

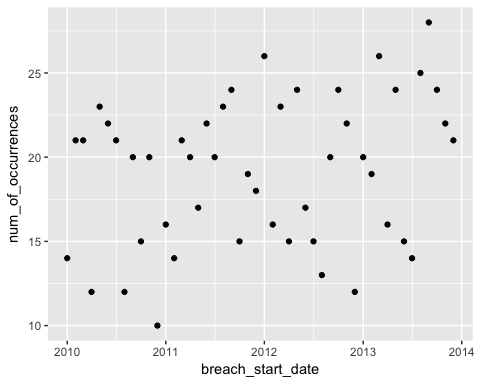
## [1] 0.05395494

plot3 <- ggplot(breaches\_by\_date\_and\_individuals\_affected, aes(breach\_start\_date, Individuals\_Affected\_in\_thou)) +  
 geom\_point(size = 2, color = "grey40") +  
 geom\_abline(intercept = coef(dataset\_mod)[1], slope = coef(dataset\_mod)[2], color = "blue")  
 plot3



It seems that the number of individuals affected by breaches over this four year span is trending ever-so-slightly positively. If this trend continues, then there will be more and more persons affected each year by cyber security breaches. What does this look like for the number of occurrences of cyber security breaches? Is it also trending upwards?

require(lubridate)  
  
make\_date\_100 <- function(year, month, day) {  
 make\_date(year, month)  
}  
  
breaches\_by\_date\_and\_num\_of\_occurrences <- breaches %>% mutate(year=year(breach\_start)) %>%  
 filter(!is.na(breach\_start) & year < 2014 & year > 2009) %>%  
 mutate(month = month(breach\_start, label=TRUE), day=day(breach\_start), breach\_start\_date = make\_date(year, month)) %>%  
 group\_by(breach\_start\_date) %>%  
 summarise(num\_of\_occurrences = n())  
  
models <- tibble(  
 a0 = runif(250, -2000, 4000),  
 a1 = runif(250, -5000, 5000)  
 )  
  
plot4 <- ggplot(breaches\_by\_date\_and\_num\_of\_occurrences, aes(breach\_start\_date, num\_of\_occurrences)) +  
 geom\_abline(aes(intercept = a0, slope=a1), data = models, alpha = 1/4) + geom\_point()  
plot4



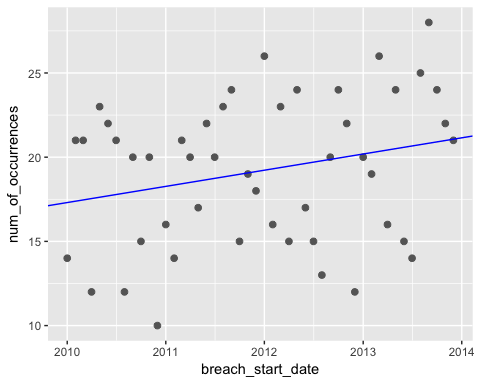
dataset\_mod <- lm(num\_of\_occurrences ~breach\_start\_date, data = breaches\_by\_date\_and\_num\_of\_occurrences)  
 coef(dataset\_mod) %>% str()

## Named num [1:2] -21.17696 0.00263  
## - attr(\*, "names")= chr [1:2] "(Intercept)" "breach\_start\_date"

plot5 <- summary(dataset\_mod)$r.squared \* 100  
 plot5

## [1] 6.559332

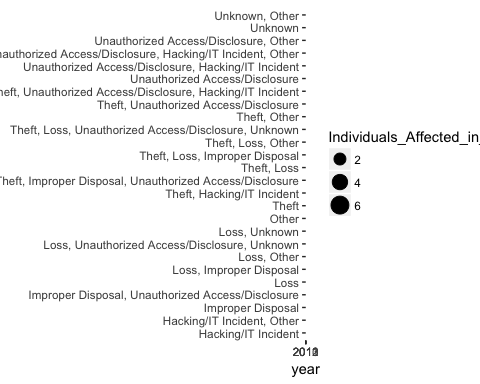
plot6 <- ggplot(breaches\_by\_date\_and\_num\_of\_occurrences, aes(breach\_start\_date, num\_of\_occurrences)) +  
 geom\_point(size = 2, color = "grey40") +  
 geom\_abline(intercept = coef(dataset\_mod)[1], slope = coef(dataset\_mod)[2], color = "blue")  
 plot6



Indeed. It would seem that the number of cyber security breaches is trending more positively than is the individuals affected. This would mean that the number of individuals affected per breach is trending negatively. Next I wanted to look at how the individuals affected fluctuated for types of breaches over the four years we are examining.

Type/Date

breaches %>%  
 mutate(year=year(breach\_start)) %>%  
 filter(year > 2009 & year < 2014) %>%  
 group\_by(Type\_of\_Breach, year) %>%  
 summarise(Individuals\_Affected\_in\_mil=sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping=aes(x=year, size=Individuals\_Affected\_in\_mil, y=Type\_of\_Breach)) +  
 geom\_point()

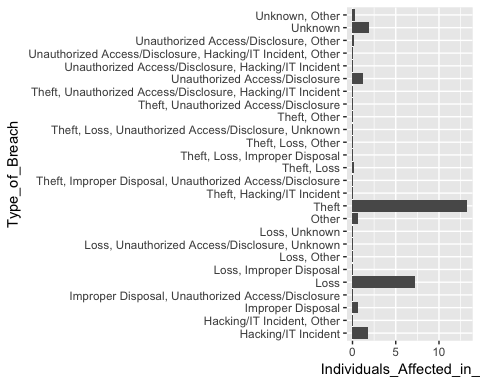


We can see that theft is the most consistent on the upside of individuals affected by million. We also see that in 2011 loss and unknown type of breaches were also on the upside of individuals affected by million. After looking at this metric, I wanted to research types of breaches and individuals affected as well as number of each type more in-depthly to see if there was some significant data there.

The first graph I generate was all of the different types of breaches and how many individuals were affected by each over the four year time span. We can see that there are some that potrude more than others in this respect.

Type:

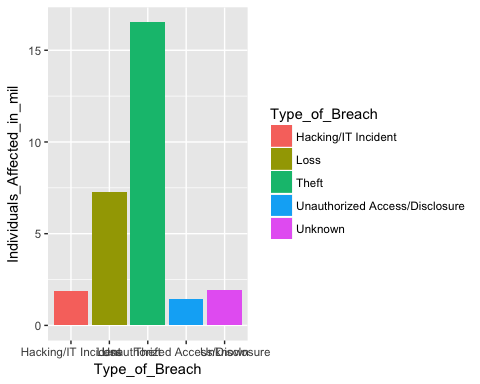
breaches %>%  
 mutate(year=year(breach\_start)) %>%  
 filter(year > 2009 & year < 2014) %>%  
 group\_by(Type\_of\_Breach) %>%  
 summarise(Individuals\_Affected\_in\_Mil=sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping=aes(x=Type\_of\_Breach, y=Individuals\_Affected\_in\_Mil)) +  
 geom\_col(position = position\_stack(reverse = TRUE)) +  
 coord\_flip()



I wanted to narrow down the top five of these types and graph them more closely. We can see this graph below.

#Top 5 types of breaches according to individuals affected by breach  
breaches %>%  
 group\_by(Type\_of\_Breach) %>%  
 summarise(Individuals\_Affected\_in\_mil=sum(Individuals\_Affected/1000000)) %>%  
 arrange(desc(Individuals\_Affected\_in\_mil)) %>%  
 top\_n(5) %>%  
 ggplot(mapping=aes(x=Type\_of\_Breach, y=Individuals\_Affected\_in\_mil, fill=Type\_of\_Breach)) +  
 geom\_col()

## Selecting by Individuals\_Affected\_in\_mil

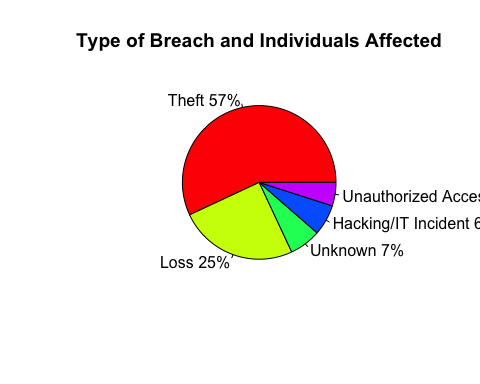


It seems theft is the type of breach that affected the most individuals, at least according to this dataset. Loss also seems to have affected a large number of individuals in comparison to the other types. I wanted to put this in a format with percentages of a whole. Below is a pie chart that shows this information.

#Top 5 types of breaches according to individuals affected by breach  
top\_types\_of\_breach\_by\_Individuals\_Affected\_for\_pie <- breaches %>%  
 group\_by(Type\_of\_Breach) %>%  
 summarise(Individuals\_Affected\_in\_mil=sum(Individuals\_Affected/1000000)) %>%  
 arrange(desc(Individuals\_Affected\_in\_mil)) %>%  
 top\_n(5)

## Selecting by Individuals\_Affected\_in\_mil

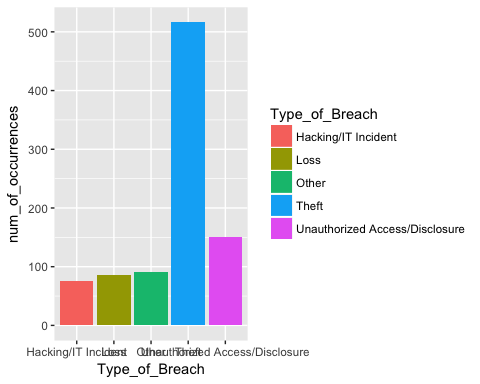
slices <- top\_types\_of\_breach\_by\_Individuals\_Affected\_for\_pie$Individuals\_Affected\_in\_mil  
lbls <- top\_types\_of\_breach\_by\_Individuals\_Affected\_for\_pie$Type\_of\_Breach  
pct <- round(slices/sum(slices) \* 100)  
lbls <- paste(lbls, pct)  
lbls <- paste(lbls, "%", sep="")  
 pie(slices, labels=lbls, col=rainbow(length(lbls)), main="Type of Breach and Individuals Affected")



We can see from this that of all the individuals affected by security breaches within this dataset, theft makes up a large majority with over fifty percent of all of the individuals affected being affected by theft. This pie chart also assumes that the same people were not affected by multiple breaches. Next I wanted to see similarly the top types of security breaches by number of occurrences to see if there was a difference.

#Top 5 types of breaches according to number of occurrences  
breaches %>%  
 group\_by(Type\_of\_Breach) %>%  
 summarise(num\_of\_occurrences = n()) %>%  
 top\_n(5) %>%  
 ggplot(mapping=aes(x=Type\_of\_Breach, y=num\_of\_occurrences, fill=Type\_of\_Breach)) +  
 geom\_col()

## Selecting by num\_of\_occurrences

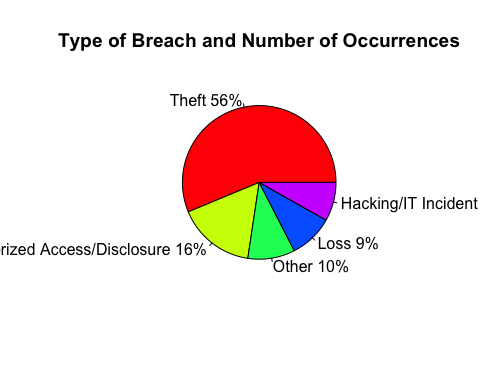


We can see here that theft also dominates the number of occurrences over the four years. The rest of the types of breaches do not even come close to Theft in this category. Let's see the percentages in a pie chart for this information.

#Top 5 types of breaches according to number of occurrences  
top\_types\_of\_breach\_by\_num\_of\_occurrences\_for\_pie <- breaches %>%  
 group\_by(Type\_of\_Breach) %>%  
 summarise(num\_of\_occurrences=n()) %>%  
 arrange(desc(num\_of\_occurrences)) %>%  
 top\_n(5)

## Selecting by num\_of\_occurrences

slices <- top\_types\_of\_breach\_by\_num\_of\_occurrences\_for\_pie$num\_of\_occurrences  
lbls <- top\_types\_of\_breach\_by\_num\_of\_occurrences\_for\_pie$Type\_of\_Breach  
pct <- round(slices/sum(slices) \* 100)  
lbls <- paste(lbls, pct)  
lbls <- paste(lbls, "%", sep="")  
 pie(slices, labels=lbls, col=rainbow(length(lbls)), main="Type of Breach and Number of Occurrences")



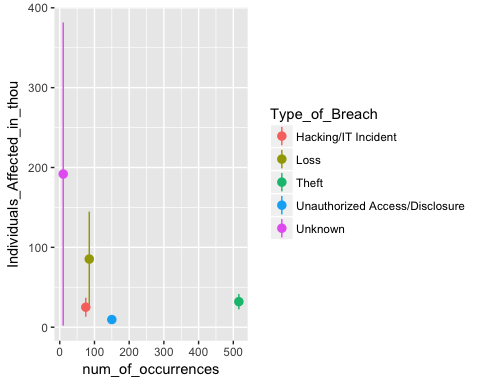
Wow! Theft makes up over half with the next closest type of breach being Unauthorized Access/Disclosure at a much lower sixteen percent. Theft is obviously a huge concern for cyber security. Can we view the number of occurrences and the individuals affected for each type of breach within the same graph to compare all of this information simultaneously? We see this below.

top\_Breach\_Types <- breaches %>% group\_by(Type\_of\_Breach) %>%  
 summarise(Individuals\_Affected\_in\_mil=sum(Individuals\_Affected/1000000)) %>%  
 arrange(desc(Individuals\_Affected\_in\_mil)) %>%  
 top\_n(5) %>%  
 select(Type\_of\_Breach)

## Selecting by Individuals\_Affected\_in\_mil

breaches %>% filter(Type\_of\_Breach %in% (top\_Breach\_Types$Type\_of\_Breach)) %>%  
 group\_by(Type\_of\_Breach) %>%  
 #summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 #mutate(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 mutate(num\_of\_occurrences = n()) %>%  
 mutate(Individuals\_Affected\_in\_thou = (Individuals\_Affected/1000)) %>%  
 ggplot(mapping=aes(x=num\_of\_occurrences, color=Type\_of\_Breach, y=Individuals\_Affected\_in\_thou)) +  
 stat\_summary()

## No summary function supplied, defaulting to `mean\_se()



Here, we can see the number of occurrences and the average number of individuals affected in thousand for each type of breach. We can also see the range of the individuals affected by different types of breaches. Unknown breach type has a wide range seemingly from one thousand individuals affected up to close to four-hundred thousand individuals affected. That is a large range of values. The other types do not vary nearly as much with loss being the other type that varies even significantly. We can also see theft as having many more occurrences but only affecting around twenty-five to thirty-five thousand persons in each occurrence. We can reason from this that the reason theft affected the most individuals overall is not because the occurrences individually affected a large number of people, but there were so many more occurrences compared to the other types of breaches that it affected many more individuals overall.

Next, I was wondering if there was a geographic area of the United States that stuck out within these security breaches. Firstly I looked at the top five states and the top five types of breaches according to individuals affected over the time span of the data. We can see this data below.

Type/State:

top\_Breach\_Types

## # A tibble: 5 × 1  
## Type\_of\_Breach  
## <chr>  
## 1 Theft  
## 2 Loss  
## 3 Unknown  
## 4 Hacking/IT Incident  
## 5 Unauthorized Access/Disclosure

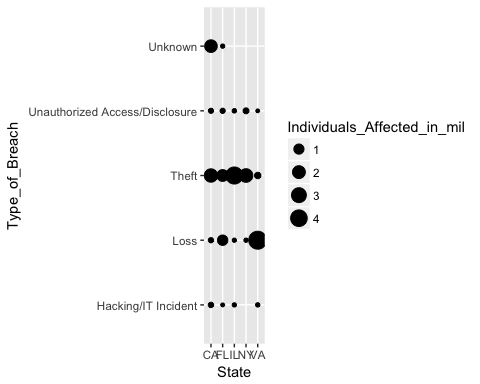
top\_States <- breaches %>%  
 group\_by(State) %>%  
 summarise(Individuals\_Affected\_in\_mil=sum(Individuals\_Affected/1000000)) %>%  
 arrange(desc(Individuals\_Affected\_in\_mil)) %>%  
 top\_n(5) %>%  
 select(State)

## Selecting by Individuals\_Affected\_in\_mil

#group\_by(Type\_of\_Breach, State) %>%  
 #summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 #top\_n(1) %>%  
  
top\_States

## # A tibble: 5 × 1  
## State  
## <chr>  
## 1 VA  
## 2 CA  
## 3 IL  
## 4 FL  
## 5 NY

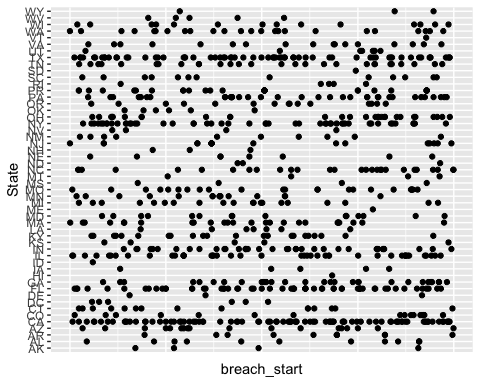
both <- breaches %>% filter(Type\_of\_Breach %in% (top\_Breach\_Types$Type\_of\_Breach) & State %in% (top\_States$State))  
  
both %>% group\_by(Type\_of\_Breach, State) %>%  
 summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping=aes(x=State, y=Type\_of\_Breach, size=Individuals\_Affected\_in\_mil)) +  
 geom\_point()



We can see from this that the top four states all had a large number of persons affected by theft breaches, where as the fifth state, Virginia, had a large number of individuals affected by loss breaches. We can also see that California, the first state (and also the state with the highest population overall), was affected largely by both theft and unknow breaches.

Next I wanted to look at all of the breaches across time within each state.

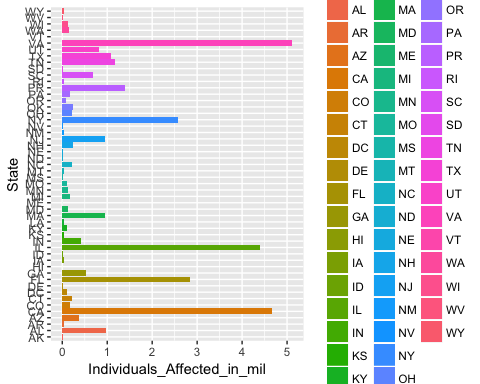
breaches %>% mutate(year=year(breach\_start)) %>% filter(year > 2009 & year< 2014) %>%  
 ggplot(mapping=aes(x=breach\_start, y=State)) + geom\_point() +  
 theme(axis.ticks.x=element\_blank(), axis.text.x=element\_blank())



We can see that there are some states with consisten breaches, while there are others with sporadic breaches. This is a very interesting graph to examine. Let's dive a little deeper into the breaches grouped by state. The graph below shows us the individuals affected in million for each state over the time span of this data.

State:

breaches %>%  
 group\_by(State) %>%  
 summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping = aes(x=State, y=Individuals\_Affected\_in\_mil, fill=State)) +  
 geom\_col(position = position\_stack(reverse = TRUE)) +  
 coord\_flip()



We can see from this that California had the most persons affected. We can also see how other states pale in comparison to this number. Vermont, for example, apperas to have no individuals affected at all.

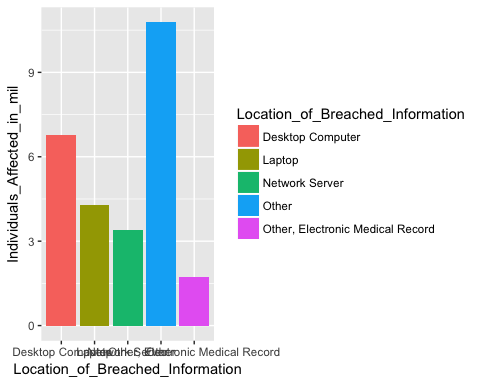
Next let's look briefly at the location from which this data was breached. We look at the top 5 locations below according to the number of individuals affected in million by data breached from these locations.

Location:

top\_Locations <- breaches %>%  
 group\_by(Location\_of\_Breached\_Information) %>%  
 summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 top\_n(5)

## Selecting by Individuals\_Affected\_in\_mil

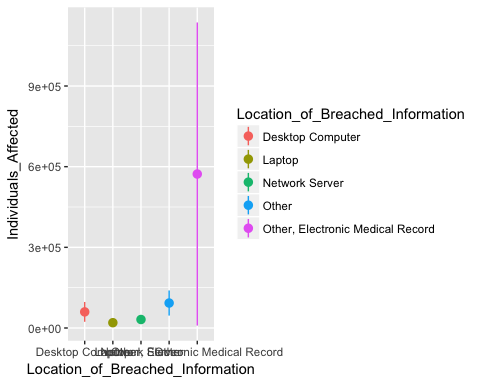
ggplot(top\_Locations, mapping = aes(x=Location\_of\_Breached\_Information, y=Individuals\_Affected\_in\_mil, fill=Location\_of\_Breached\_Information)) +  
 geom\_col()



It seems that locations not explicitly specified are the highest in the number of individuals affected. The second highest, and the highest of specified locations, is the number of individuals affected by data breached from desktop computers. We can see the other three in this graph as well as areas of interest. The graph below shows us the average number of individuals affected for each location as well as the range of individuals affected.

breaches %>% filter(Location\_of\_Breached\_Information %in% top\_Locations$Location\_of\_Breached\_Information) %>%  
 group\_by(Location\_of\_Breached\_Information) %>%  
 #summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping = aes(x=Location\_of\_Breached\_Information, y=Individuals\_Affected, color=Location\_of\_Breached\_Information)) +  
 stat\_summary()

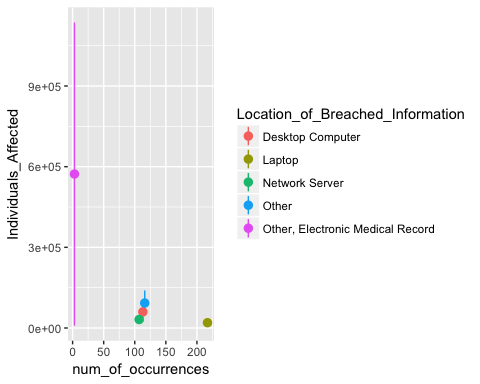
## No summary function supplied, defaulting to `mean\_se()



We can see that, for the most part there was not a wide range of the individuals affected by each breach for each location. Other, Electronic Medical Record seems to be the only one with fluctuation, but it does vary immensly. Let's look at this while also viewing the number of occurrences of breaches for each location.

breaches %>% filter(Location\_of\_Breached\_Information %in% top\_Locations$Location\_of\_Breached\_Information) %>%  
 group\_by(Location\_of\_Breached\_Information) %>%  
 mutate(num\_of\_occurrences = n()) %>%  
 mutate(Individuals\_Affected\_in\_thou = (Individuals\_Affected/1000)) %>%  
 #summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping = aes(x=num\_of\_occurrences, y=Individuals\_Affected, color=Location\_of\_Breached\_Information)) +  
 stat\_summary()

## No summary function supplied, defaulting to `mean\_se()



We can see from this that laptops had the most number of breaches but also had the fewest average individuals affected per case as well as very little variation in the individuals affected from case to case. This is the reason it was third in overall individuals affected. We can also see that Network Server, Desktop Computer, and Unknown locations have a similar number of cases with each successive one having slightly more cases and a slighly higher average of individuals affected per case. Network Server does not vary much in that respect, which is why it is fourth in overall individuals affected. We can also see that Other, Electronic Medical Record has very few cases, but it also has a very wide range of individuals affected per case as well as a very high average of individuals affected per case. This is the reason it makes the list of the top five locations of security breaches by overall individuals affected.

Why does this information matter? Personally, it matters for a couple of different reasons. It matters beacause I do not want my information to be vulnerable to being breach, but it also matters because I work for the U.S. Army Corps of Engineers which is a subset of the Department of Defense. With information such as the above, we can make efforts to protect our national security and the information that we do not want others to have. We have an entire team at ERDC that is dedicated to doing just that, and this information will be very helpful to them.

In conclusion, we have observed a dataset of cyber security breaches from the U.S. Department of Health and Human Services. The data was collected over a four year period from 2010-2013. We examined cyber security breaches and trends within from several different angles including from a perspective of type of breach, location of breached information, geographic location (U.S. states), as well as breaches over the time span of data collection. It is important information for both personal security as well as national security.