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

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

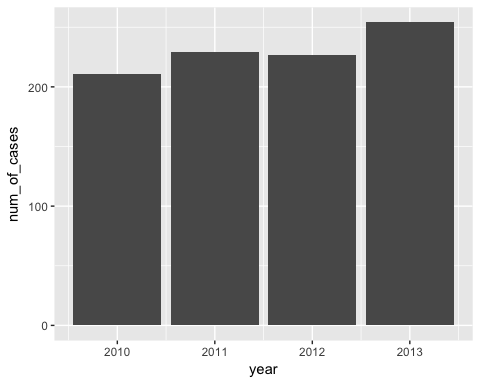
## Parsed with column specification:  
## cols(  
## X1 = col\_integer(),  
## Number = col\_integer(),  
## Name\_of\_Covered\_Entity = col\_character(),  
## State = col\_character(),  
## Business\_Associate\_Involved = col\_character(),  
## Individuals\_Affected = col\_integer(),  
## Date\_of\_Breach = 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)

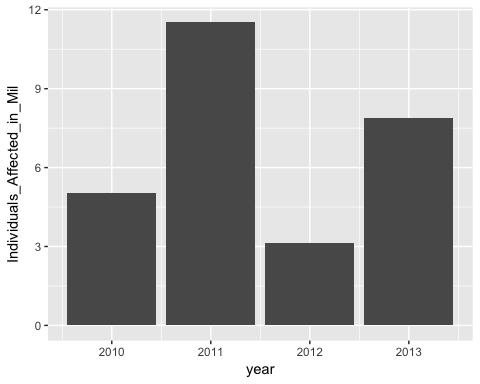
We may need to remove Date\_of\_Breach to tidy the dataset. Really only thing I've seen.

First we will start by exploring the dataset from a breadth first approach. These initial examinations will be where we glean our direction for the following exploration in this document.

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

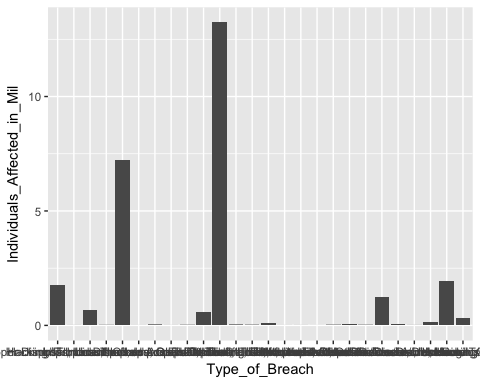


breaches %>%  
 filter(year > 2009 & year < 2014) %>%  
 group\_by(year) %>%  
 summarise(Individuals\_Affected\_in\_Mil=sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping=aes(x=year, y=Individuals\_Affected\_in\_Mil)) +  
 geom\_col()

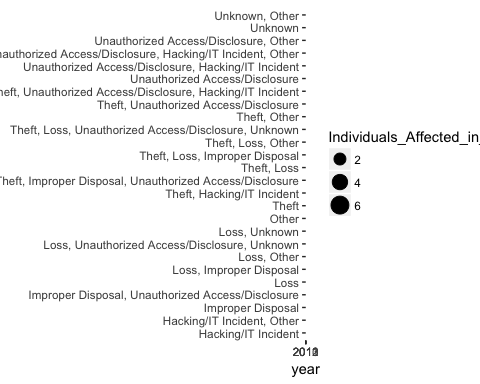


#geom\_smooth()

breaches %>%  
 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()

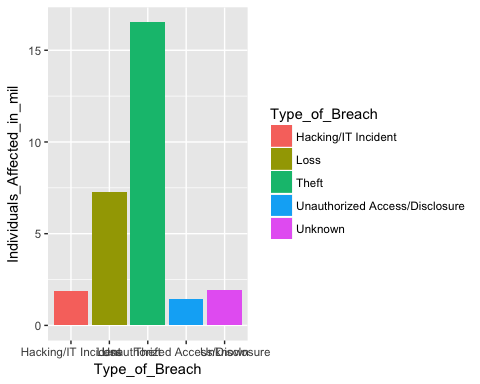


breaches %>%  
 filter(year > 2009 & year < 2014) %>%  
 group\_by(Type\_of\_Breach, year) %>%  
 summarise(Individuals\_Affected\_in\_mil=sum(Individuals\_Affected/1000000)) %>%  
 ggplot(mapping=aes(x=year, size=Individuals\_Affected\_in\_mil, y=Type\_of\_Breach)) +  
 geom\_point()



#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()

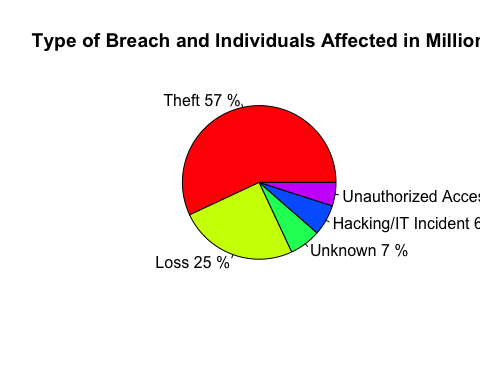
## Selecting by Individuals\_Affected\_in\_mil



#Top 5 types of breaches according to individuals affected by breach  
breach\_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)

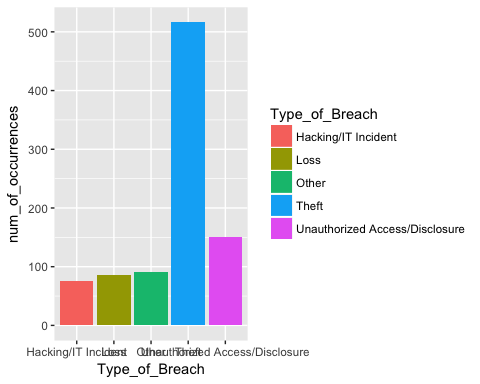
## Selecting by Individuals\_Affected\_in\_mil

slices <- breach\_for\_pie$Individuals\_Affected\_in\_mil  
lbls <- breach\_for\_pie$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 in Million")



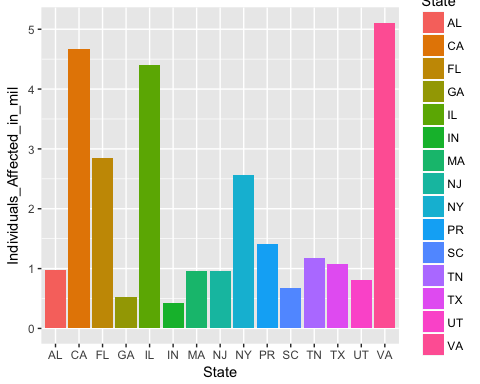
#Top 5 types of breaches according to number of occurrences  
breaches %>%  
 group\_by(Type\_of\_Breach) %>%  
 summarise(num\_of\_occurrences = n()) %>%  
 top\_n(5) %>%  
 ggplot(mapping=aes(x=Type\_of\_Breach, y=num\_of\_occurrences, fill=Type\_of\_Breach)) +  
 geom\_col()

## Selecting by num\_of\_occurrences



breaches %>%  
 group\_by(State) %>%  
 summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 top\_n(15) %>%  
 ggplot(mapping = aes(x=State, y=Individuals\_Affected\_in\_mil, fill=State)) +  
 geom\_col();

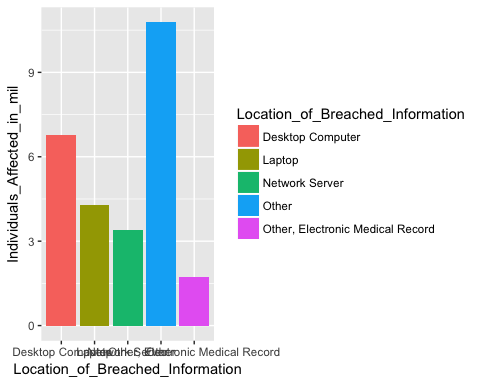
## Selecting by Individuals\_Affected\_in\_mil



#map(database = 'county')

breaches %>%  
 group\_by(Location\_of\_Breached\_Information) %>%  
 summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 top\_n(5) %>%  
 ggplot(mapping = aes(x=Location\_of\_Breached\_Information, y=Individuals\_Affected\_in\_mil, fill=Location\_of\_Breached\_Information)) +  
 geom\_col()

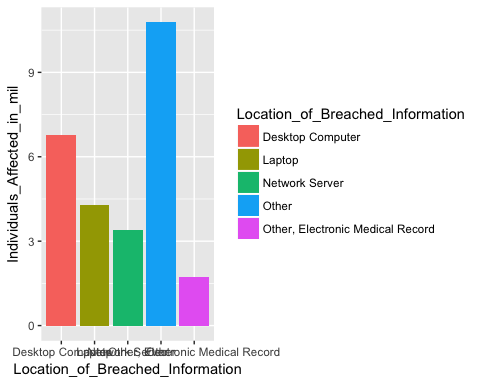
## Selecting by Individuals\_Affected\_in\_mil



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

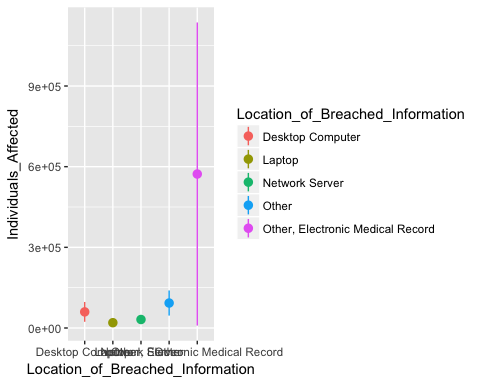
## Selecting by Individuals\_Affected\_in\_mil

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\_in\_mil, fill=Location\_of\_Breached\_Information)) +  
 geom\_col()



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()



As I begin to analyze this data further, it begins to become apparent that the credibility of this data and its source are not quite at the standard desired.

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

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

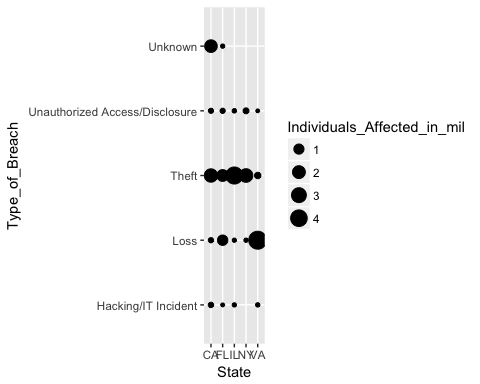
top\_States <- breaches %>%  
 group\_by(State) %>%  
 summarise(Individuals\_Affected\_in\_mil=sum(Individuals\_Affected/1000000)) %>%  
 arrange(desc(Individuals\_Affected\_in\_mil)) %>%  
 top\_n(5) %>%  
 select(State)

## Selecting by Individuals\_Affected\_in\_mil

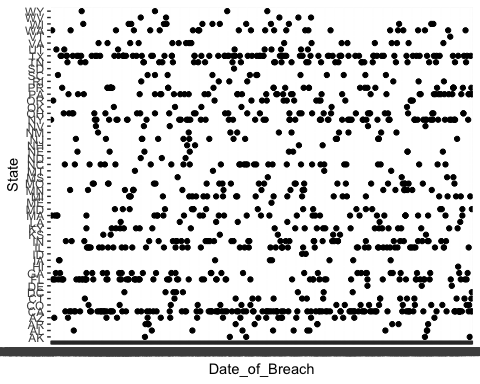
#group\_by(Type\_of\_Breach, State) %>%  
 #summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 #top\_n(1) %>%  
  
top\_States

## # 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()



breaches %>% ggplot(mapping=aes(x=Date\_of\_Breach, y=State)) + geom\_point()

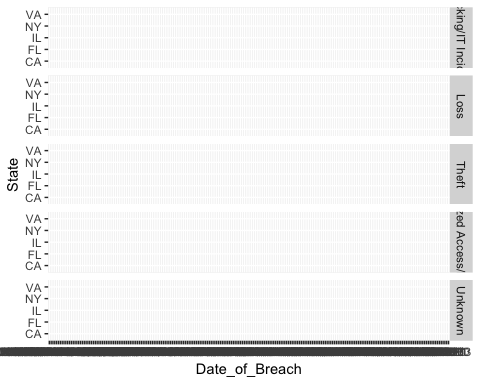


breaches\_by\_date <- both %>%  
 group\_by(Date\_of\_Breach)  
 #summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected))  
breaches\_by\_date

## Source: local data frame [243 x 14]  
## Groups: Date\_of\_Breach [218]  
##   
## X1 Number Name\_of\_Covered\_Entity State  
## <int> <int> <chr> <chr>  
## 1 5 4 L. Douglas Carlson, M.D. CA  
## 2 6 5 David I. Cohen, MD CA  
## 3 7 6 Michele Del Vicario, MD CA  
## 4 8 7 Joseph F. Lopez, MD CA  
## 5 9 8 Mark D. Lurie, MD CA  
## 6 10 9 City of Hope National Medical Center CA  
## 7 14 13 Kern Medical Center CA  
## 8 23 22 Kaiser Permanente Medical Care Program CA  
## 9 24 23 Blue Island Radiology Consultants IL  
## 10 29 28 Advocate Health Care IL  
## # ... with 233 more rows, and 10 more variables:  
## # Business\_Associate\_Involved <chr>, Individuals\_Affected <int>,  
## # Date\_of\_Breach <chr>, Type\_of\_Breach <chr>,  
## # Location\_of\_Breached\_Information <chr>, Date\_Posted\_or\_Updated <date>,  
## # Summary <chr>, breach\_start <date>, breach\_end <date>, year <int>

ggplot(data=breaches\_by\_date, mapping=aes(x=Date\_of\_Breach, y=State)) +  
 geom\_line() +  
 facet\_grid(Type\_of\_Breach ~ .)

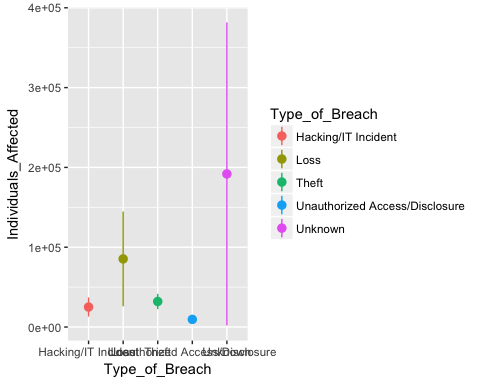
## geom\_path: Each group consists of only one observation. Do you need to  
## adjust the group aesthetic?



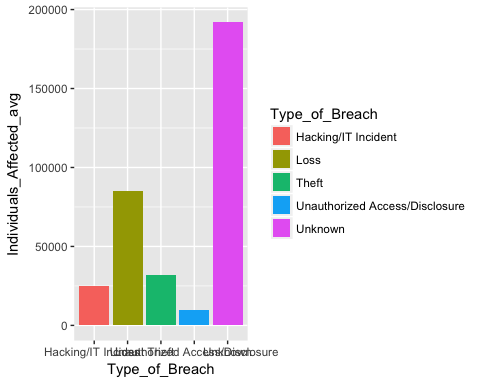
##maybe try grouping by date..

breaches %>% filter(Type\_of\_Breach %in% (top\_Breach\_Types$Type\_of\_Breach)) %>%  
 ggplot(mapping = aes(x=Type\_of\_Breach, y=Individuals\_Affected, color=Type\_of\_Breach)) + stat\_summary()

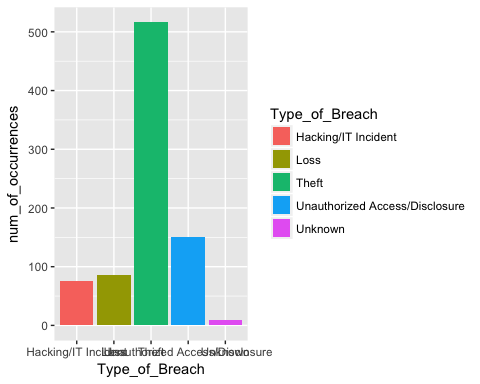
## No summary function supplied, defaulting to `mean\_se()



breaches %>% filter(Type\_of\_Breach %in% (top\_Breach\_Types$Type\_of\_Breach)) %>%  
 group\_by(Type\_of\_Breach) %>%  
 summarise(Individuals\_Affected\_avg = mean(Individuals\_Affected)) %>%  
 ggplot(mapping = aes(x=Type\_of\_Breach, y=Individuals\_Affected\_avg, fill=Type\_of\_Breach)) + geom\_col(size=5)

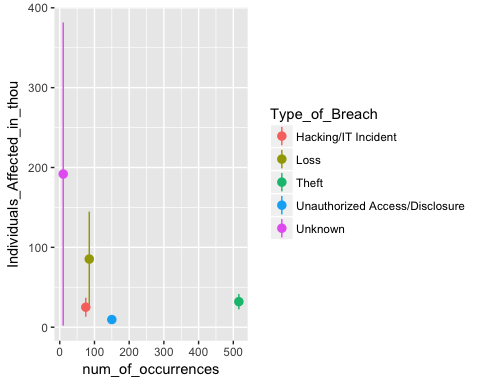


breaches %>% filter(Type\_of\_Breach %in% (top\_Breach\_Types$Type\_of\_Breach)) %>%  
 group\_by(Type\_of\_Breach) %>%  
 summarise(num\_of\_occurrences = n()) %>%  
 ggplot(mapping = aes(x=Type\_of\_Breach, y=num\_of\_occurrences, fill=Type\_of\_Breach)) + geom\_col(size = 5)



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()

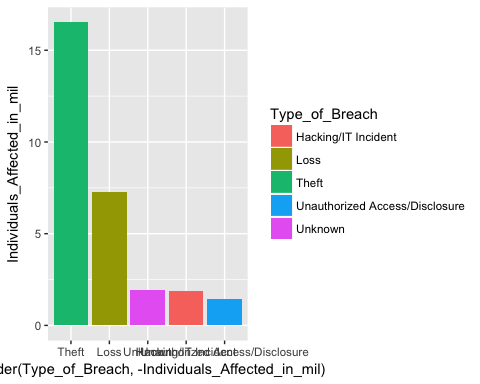
## No summary function supplied, defaulting to `mean\_se()



dataset <- 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

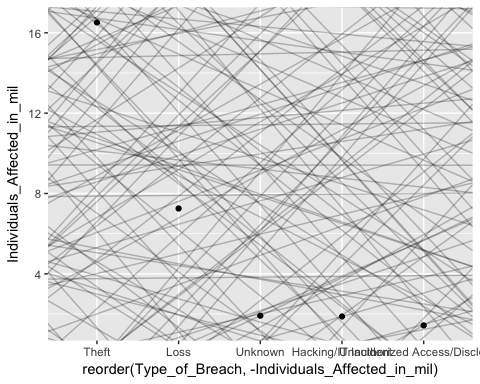
ggplot(data=dataset, mapping=aes(x=reorder(Type\_of\_Breach, -Individuals\_Affected\_in\_mil), y=Individuals\_Affected\_in\_mil, fill=Type\_of\_Breach)) +  
 geom\_col()



dataset

## # A tibble: 5 × 2  
## Type\_of\_Breach Individuals\_Affected\_in\_mil  
## <chr> <dbl>  
## 1 Theft 16.515554  
## 2 Loss 7.254286  
## 3 Unknown 1.918312  
## 4 Hacking/IT Incident 1.878870  
## 5 Unauthorized Access/Disclosure 1.434018

models <- tibble(  
 a0 = runif(250, -20, 40),  
 a1 = runif(250, -5, 5)  
 )  
  
p2 <- ggplot(dataset, aes(reorder(Type\_of\_Breach, -Individuals\_Affected\_in\_mil), Individuals\_Affected\_in\_mil)) + geom\_abline(aes(intercept = a0, slope=a1), data = models, alpha = 1/4) + geom\_point()  
 p2



dataset\_mod <- lm(reorder(Type\_of\_Breach, -Individuals\_Affected\_in\_mil) ~Individuals\_Affected\_in\_mil, data = dataset)

## Warning in model.response(mf, "numeric"): using type = "numeric" with a  
## factor response will be ignored

## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors

coef(dataset\_mod) %>% str()

## Named num [1:2] 4.238 -0.214  
## - attr(\*, "names")= chr [1:2] "(Intercept)" "Individuals\_Affected\_in\_mil"

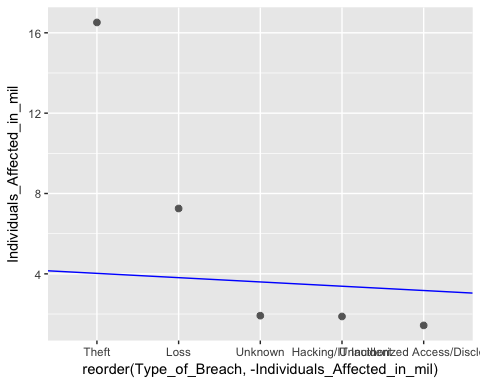
p3 <- summary(dataset\_mod)$r.squared \* 100

## Warning in Ops.factor(r, 2): '^' not meaningful for factors

p3

## [1] NA

p4 <- ggplot(dataset, aes(reorder(Type\_of\_Breach, -Individuals\_Affected\_in\_mil), Individuals\_Affected\_in\_mil)) +  
 geom\_point(size = 2, color = "grey40") +  
 geom\_abline(intercept = coef(dataset\_mod)[1], slope = coef(dataset\_mod)[2], color = "blue")  
 p4



#grid <- dataset %>%  
 #add\_predictions(dataset\_mod)  
 #print(grid)

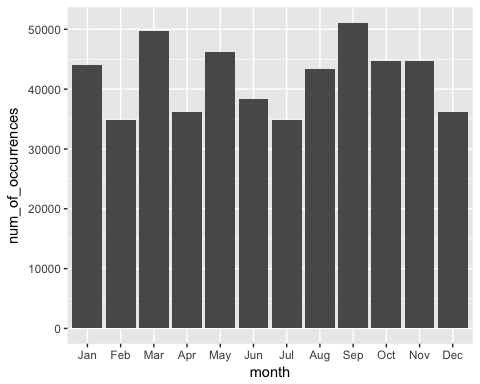
#Plot by month individuals affected  
#fit a model to this using MAD or something other than RMS  
require(lubridate)

## Loading required package: lubridate

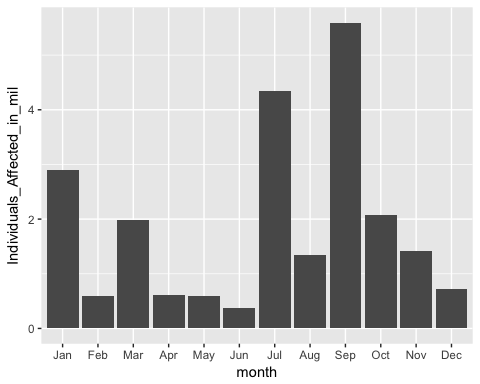
##   
## Attaching package: 'lubridate'

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

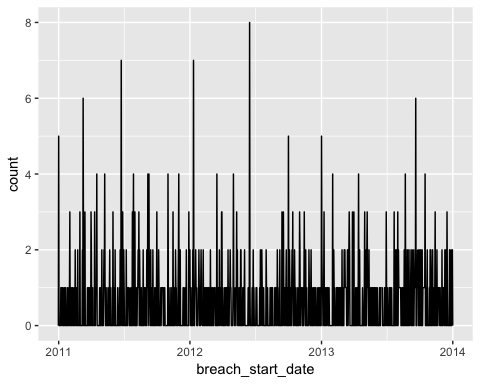
make\_date\_100 <- function(year, month, day) {  
 make\_date(year, month, day)  
}  
  
breaches\_date <- breaches %>% filter(!is.na(breach\_start) & year < 2014 & year > 2010) %>%  
 mutate(month = month(breach\_start, label=TRUE), num\_of\_occurrences = n()) %>%  
 ggplot(aes(x=month, y=num\_of\_occurrences)) + geom\_col(alpha = 1)  
breaches\_date



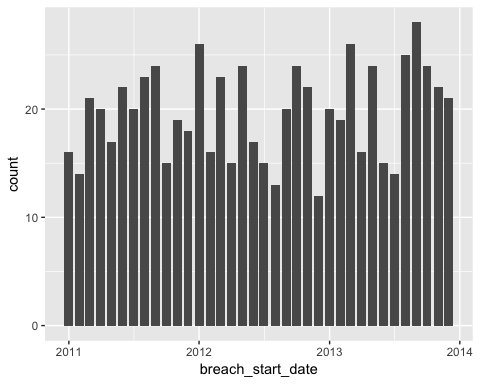
require(lubridate)  
  
breaches\_date\_affected <- breaches %>% filter(!is.na(breach\_start) & year < 2014 & year > 2010) %>%  
 mutate(month = month(breach\_start, label=TRUE)) %>%  
 group\_by(month) %>%  
 summarise(Individuals\_Affected\_in\_mil = sum(Individuals\_Affected/1000000)) %>%  
 ggplot(aes(x=month, y=Individuals\_Affected\_in\_mil)) + geom\_col()  
breaches\_date\_affected



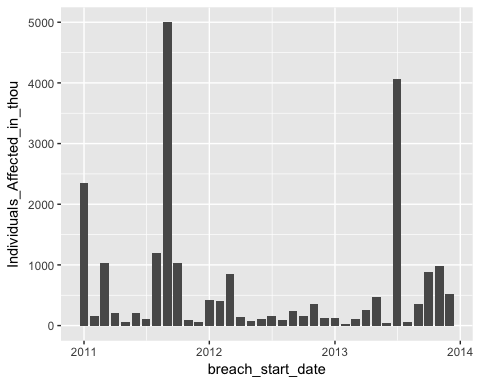
require(lubridate)  
  
make\_date\_100 <- function(year, month, day) {  
 make\_date(year, month, day)  
}  
  
breaches\_by\_date\_2 <- breaches %>% filter(!is.na(breach\_start) & year < 2014 & year > 2010) %>%  
 mutate(month = month(breach\_start, label = TRUE), day = day(breach\_start), breach\_start\_date = make\_date(year, month, day)) %>%  
 ggplot(aes(breach\_start\_date)) + geom\_freqpoly(binwidth = month(1) %>% as.numeric())  
breaches\_by\_date\_2



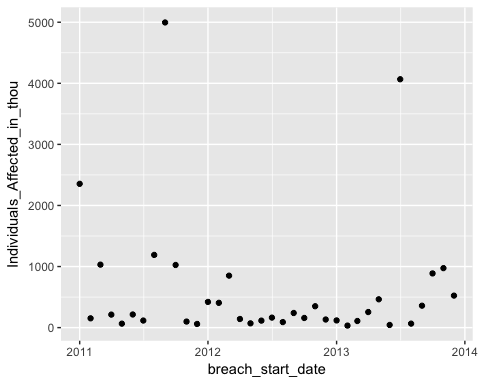
require(lubridate)  
  
make\_date\_100 <- function(year, month) {  
 make\_date(year, month, day)  
}  
  
breaches\_by\_date\_2 <- breaches %>% filter(!is.na(breach\_start) & year < 2014 & year > 2010) %>%  
 mutate(month = month(breach\_start, label = TRUE), day = day(breach\_start), breach\_start\_date = make\_date(year, month)) %>%  
 ggplot(aes(breach\_start\_date)) + geom\_bar()  
breaches\_by\_date\_2



require(lubridate)  
  
make\_date\_100 <- function(year, month, day) {  
 make\_date(year, month)  
}  
  
breaches\_by\_date\_3 <- breaches %>% filter(!is.na(breach\_start) & year < 2014 & year > 2010) %>%  
 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\_date\_3



require(lubridate)  
  
make\_date\_100 <- function(year, month, day) {  
 make\_date(year, month)  
}  
  
breaches\_by\_date4 <- breaches %>% filter(!is.na(breach\_start) & year < 2014 & year > 2010) %>%  
 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)  
 )  
  
p2 <- ggplot(breaches\_by\_date4, aes(breach\_start\_date, Individuals\_Affected\_in\_thou)) + geom\_abline(aes(intercept = a0, slope=a1), data = models, alpha = 1/4) + geom\_point()  
 p2



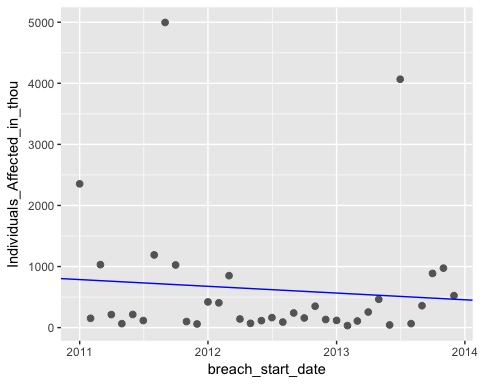
dataset\_mod <- lm(Individuals\_Affected\_in\_thou ~breach\_start\_date, data = breaches\_by\_date4)  
 coef(dataset\_mod) %>% str()

## Named num [1:2] 5326.513 -0.303  
## - attr(\*, "names")= chr [1:2] "(Intercept)" "breach\_start\_date"

p3 <- summary(dataset\_mod)$r.squared \* 100  
 p3

## [1] 0.8193493

p4 <- ggplot(breaches\_by\_date4, 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")  
 p4



#grid <- breaches\_by\_date4 %>%  
 #add\_predictions(dataset\_mod)  
 #print(grid)