# Trabajo final Data Science - Grupo 1

## Brechas de ciberseguridad - Estados Unidos

-Ernesto Guarda
-Christian Vadillo
-Cristhian Medina
-Jorge Cabrera
-Royer Rojas

## 1 Objetivo

El objetivo de esta práctica es realizar un estudio de las brechas de ciberseguridad que se han dado en Estados Unidos entre los años 2000 y 2014.

Luego de presentar la estructura y los trabajos realizados en el dataset se procederá a responder las preguntas que se han planteado con el fin de obtener información para orientarnos a tomar decisiones.

### 1.1 Librerías utilizadas

Para poder elaborar este script hemos utilizado las siguientes librerias de R:

1. readr
2. dplyr
3. ggplot2
4. tidyverse
5. ggthemes
6. lubridate
7. lattice
8. survival
9. Formula
10. Hmisc
11. rmarkdown
12. knirt

### 1.2 Dataset

El dataset utilizado se llama “Cyber Security Breaches” y puede se encontrado dando click aquí

### 1.3 Preparación del dataset

Antes de comenzar a trabajar con el dataset ajustaremos los tipos de variables para poder obtener resultados correctos

```r
#La variable "State" la convertiremos a tipo factor
cyberb$State <- as.factor(cyberb$State)
#La variable "Type_of_Breach" la convertiremos a tipo factor
cyberb$Type_of_Breach <- as.factor(cyberb$Type_of_Breach)
#La variable "Location_of_Breached_Information" la convertiremos a tipo factor
cyberb$Location_of_Breached_Information <- as.factor(cyberb$Location_of_Breached_Information)
#La variable "Date_Posted_or_Updated" la convertiremos a tipo fecha
cyberb$Date_Posted_or_Updated <- as.Date(cyberb$Date_Posted_or_Updated,format="%d/%m/%Y")
#La variable "breach_start" la convertiremos a tipo fecha
cyberb$breach_start <- as.Date(cyberb$breach_start,format="%d/%m/%Y")
#La variable "breach_end" la convertiremos a tipo fecha
cyberb$breach_end <- as.Date(cyberb$breach_end,format="%d/%m/%Y")
```

### 1.4 Descripción de las variables del dataset

```
## cyberb
##
##  10  Variables      1055  Observations
## ---------------------------------------------------------------------------
## Name_of_Covered_Entity
##        n  missing distinct
##     1055        0      963
##
## lowest : 101 FAMILY MEDICAL GROUP
## highest: Yale University
## ---------------------------------------------------------------------------
## State
##        n  missing distinct
##     1055        0       52
##
## lowest : AK AL AR AZ CA, highest: VT WA WI WV WY
## ---------------------------------------------------------------------------
## Business_Associate_Involved
##        n  missing distinct
##      271      784      214
##
## lowest : Accretive Health                                      Accuprint
## highest: Xand Corporation                                      Xforia Web Services
## ---------------------------------------------------------------------------
## Individuals_Affected
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1055        0      809        1    30262    55209      550      629
##      .25      .50      .75      .90      .95
##     1000     2300     6941    20446    55062
##
## lowest :     500     501     502     504     505
## highest: 1220000 1700000 1900000 4029530 4900000
## ---------------------------------------------------------------------------
## Type_of_Breach
##        n  missing distinct
##     1055        0       28
##
## lowest : Hacking/IT Incident                                   Hacking/IT Incident, Other
```

```
## highest: Unauthorized Access/Disclosure, Hacking/IT Incident        Unauthorized Access/Disclosure, H
## ---------------------------------------------------------------------------
## Location_of_Breached_Information
##        n  missing distinct
##     1055        0       41
##
## lowest : Desktop Computer                                                    Desktop Computer
## highest: Other Portable Electronic Device, Other                             Other Portable El
## ---------------------------------------------------------------------------
## Date_Posted_or_Updated
##          n    missing   distinct       Info       Mean        Gmd        .05
##       1055          0         43      0.719 2014-02-23      47.55 2014-01-23
##        .10        .25        .50        .75        .90        .95
## 2014-01-23 2014-01-23 2014-01-23 2014-03-24 2014-06-03 2014-06-19
##
## lowest : 2014-01-23 2014-01-24 2014-01-31 2014-02-11 2014-02-12
## highest: 2014-06-19 2014-06-20 2014-06-24 2014-06-27 2014-06-30
## ---------------------------------------------------------------------------
## Summary
##        n  missing distinct
##      142      913      141
##
## lowest :
##
## OCR opened an investigation of the covered entity (CE), Paul G. Klein DPM, after it reported an encry
##
##
##
##
##
## The covered entity (CE), Medco Health Solutions, mailed letters with incorrect addresses after a prog
##
##
##
##
## highest: Two unencrypted desktop computers containing the electronic protected health information (el
## ---------------------------------------------------------------------------
## breach_start
##          n    missing   distinct       Info       Mean        Gmd        .05
##       1055          0        732          1 2011-12-09      612.9 2009-10-31
##        .10        .25        .50        .75        .90        .95
## 2010-02-17 2010-11-08 2012-01-11 2013-03-07 2013-10-17 2014-01-09
##
## lowest : 1997-01-01 2002-05-06 2003-03-29 2004-04-21 2004-05-01
## highest: 2014-04-19 2014-05-13 2014-05-27 2014-05-30 2014-06-02
## ---------------------------------------------------------------------------
## breach_end
##          n    missing   distinct       Info       Mean        Gmd        .05
##        145        910        121          1 2012-10-28      279.6 2011-11-17
##        .10        .25        .50        .75        .90        .95
## 2011-12-19 2012-04-22 2012-10-29 2013-05-29 2013-08-15 2013-10-03
##
## lowest : 2007-06-14 2011-02-28 2011-08-05 2011-08-18 2011-09-20
## highest: 2013-10-15 2013-10-31 2013-11-06 2013-11-08 2013-11-30
```

```
## ------------------------------------------------------------------------
```

### 1.5 Resumen del dataset

```
##  Name_of_Covered_Entity     State     Business_Associate_Involved
##  Length:1055            CA     :113   Length:1055
##  Class :character       TX     : 83   Class :character
##  Mode  :character       FL     : 66   Mode  :character
##                         NY     : 58
##                         IL     : 49
##                         IN     : 40
##                         (Other):646
##  Individuals_Affected                         Type_of_Breach
##  Min.   :    500      Theft                          :516
##  1st Qu.:   1000      Unauthorized Access/Disclosure:150
##  Median :   2300      Other                          : 91
##  Mean   :  30262      Loss                           : 85
##  3rd Qu.:   6941      Hacking/IT Incident            : 75
##  Max.   :4900000      Improper Disposal              : 38
##                       (Other)                        :100
##                  Location_of_Breached_Information Date_Posted_or_Updated
##  Paper                              :227          Min.   :2014-01-23
##  Laptop                             :217          1st Qu.:2014-01-23
##  Other                              :116          Median :2014-01-23
##  Desktop Computer                   :113          Mean   :2014-02-23
##  Network Server                     :107          3rd Qu.:2014-03-24
##  Other Portable Electronic Device: 60             Max.   :2014-06-30
##  (Other)                            :215
##    Summary           breach_start          breach_end
##  Length:1055        Min.   :1997-01-01   Min.   :2007-06-14
##  Class :character   1st Qu.:2010-11-08   1st Qu.:2012-04-22
##  Mode  :character   Median :2012-01-11   Median :2012-10-29
##                     Mean   :2011-12-09   Mean   :2012-10-28
##                     3rd Qu.:2013-03-07   3rd Qu.:2013-05-29
##                     Max.   :2014-06-02   Max.   :2013-11-30
##                                          NA's   :910
```

## 2 Preguntas

### 2.1 ¿Cuáles son los tipos de brechas que afectarona más personas?

```
a1 <- cyberb %>%
    group_by(Type_of_Breach) %>%
    summarise(count = n())
a1$ID <- paste("B",c(1:length(a1$Type_of_Breach)), sep = "")
```

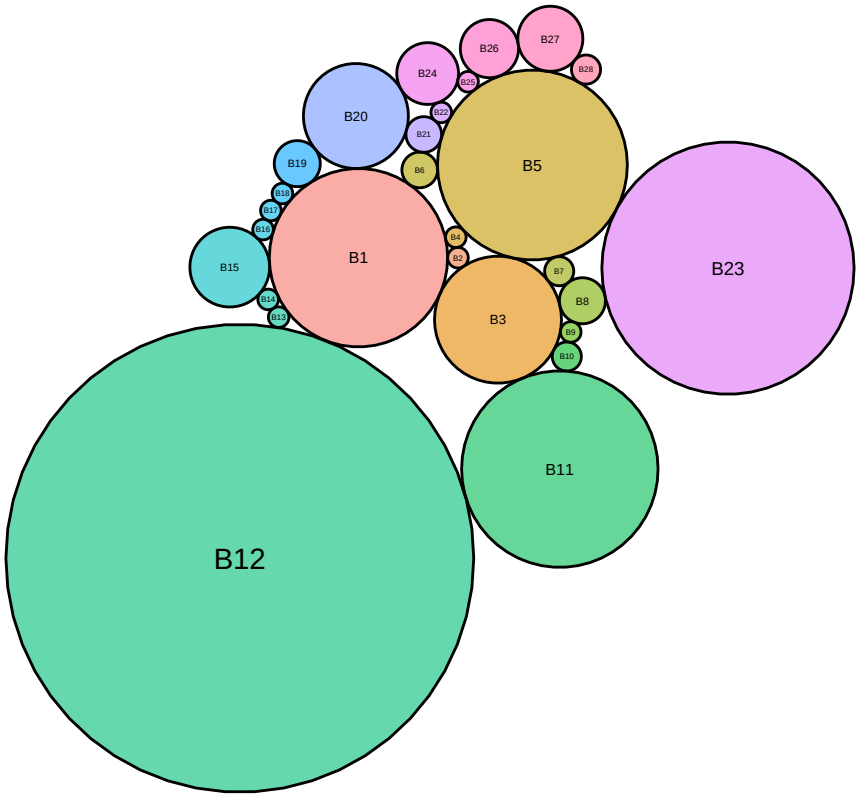


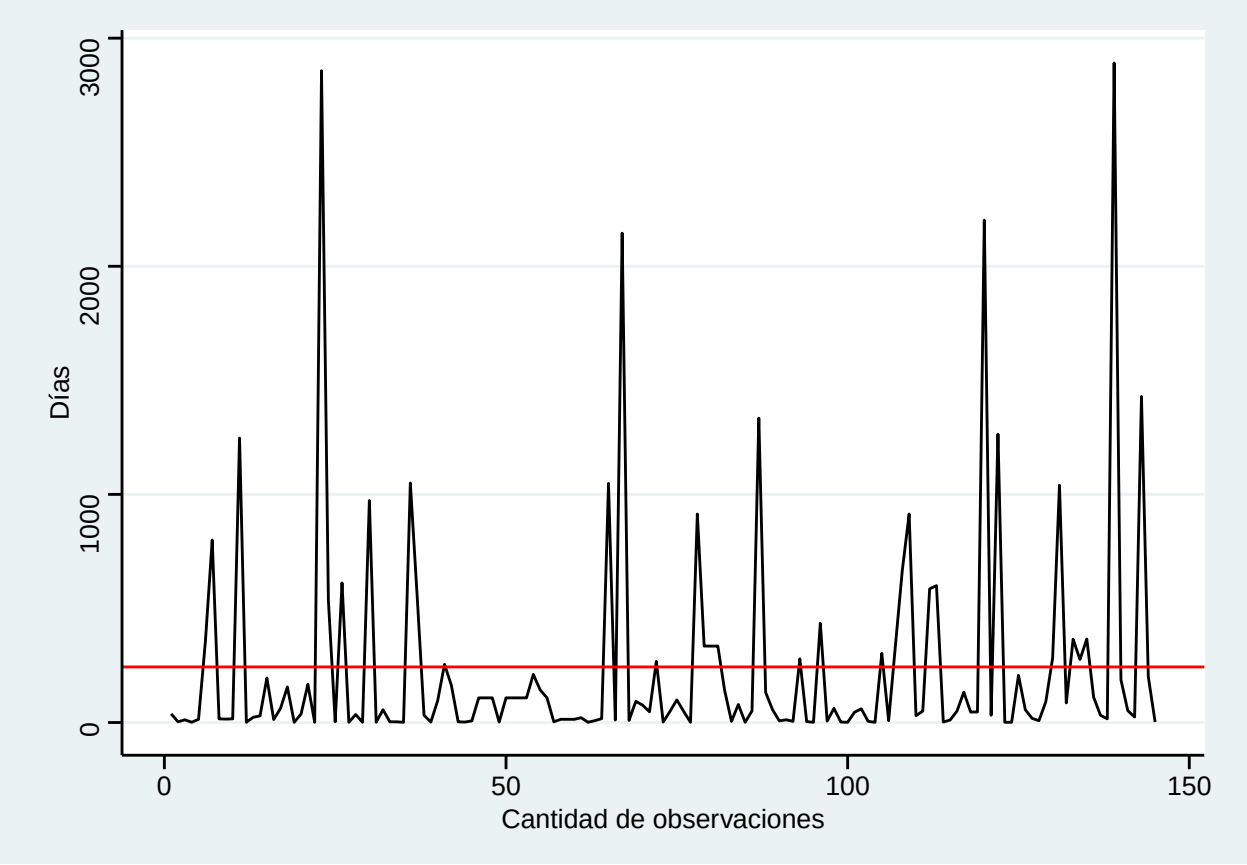

Table 1: Tipos de brechas

| Id | Tipo.de.brecha | Cantidad |
|---|---|---|
| B1 | Hacking/IT Incident | 75 |
| B2 | Hacking/IT Incident, Other | 1 |
| B3 | Improper Disposal | 38 |
| B4 | Improper Disposal, Unauthorized Access/Disclosure | 1 |
| B5 | Loss | 85 |
| B6 | Loss, Improper Disposal | 3 |
| B7 | Loss, Other | 2 |
| B8 | Loss, Unauthorized Access/Disclosure | 5 |
| B9 | Loss, Unauthorized Access/Disclosure, Unknown | 1 |
| B10 | Loss, Unknown | 2 |
| B11 | Other | 91 |
| B12 | Theft | 516 |
| B13 | Theft, Hacking/IT Incident | 1 |
| B14 | Theft, Improper Disposal, Unauthorized Access/Disclosure | 1 |
| B15 | Theft, Loss | 15 |
| B16 | Theft, Loss, Improper Disposal | 1 |
| B17 | Theft, Loss, Other | 1 |
| B18 | Theft, Loss, Unauthorized Access/Disclosure, Unknown | 1 |
| B19 | Theft, Other | 5 |
| B20 | Theft, Unauthorized Access/Disclosure | 26 |
| B21 | Theft, Unauthorized Access/Disclosure, Hacking/IT Incident | 3 |
| B22 | Theft, Unauthorized Access/Disclosure, Other | 1 |

| Id | Tipo.de.brecha | Cantidad |
|---|---|---|
| B23 | Unauthorized Access/Disclosure | 150 |
| B24 | Unauthorized Access/Disclosure, Hacking/IT Incident | 9 |
| B25 | Unauthorized Access/Disclosure, Hacking/IT Incident, Other | 1 |
| B26 | Unauthorized Access/Disclosure, Other | 8 |
| B27 | Unknown | 10 |
| B28 | Unknown, Other | 2 |

### 2.2 ¿Cual es el tiempo promedio para superar la brecha?

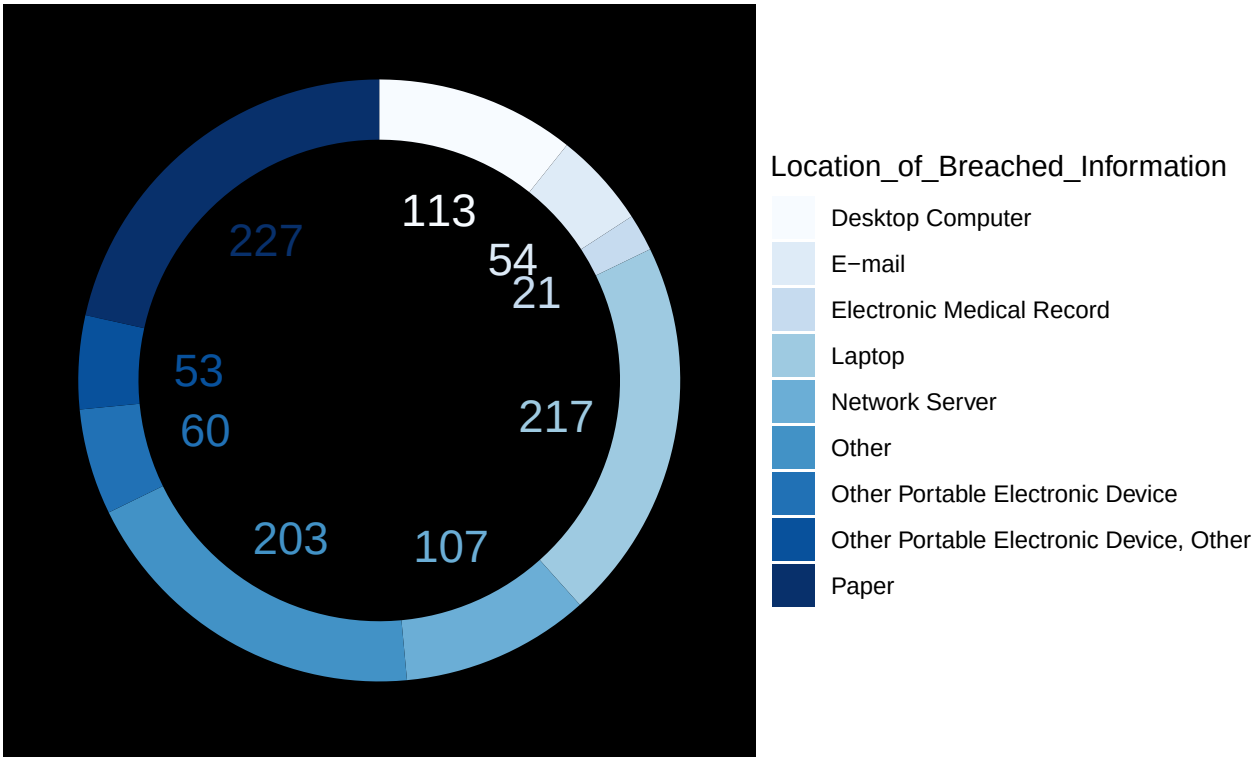
```
#Se crea un nuevo dataframe en el cual almacenamos los días que duran las brechas
a2 <- data.frame(day=na.omit(cyberb$breach_end - cyberb$breach_start))
#luego calculamos el promedio
media <- as.numeric(mean(a2$day))
```




### 2.3 ¿Qué tipo de almacenamiento de la información tuvo mas vulnerabilidades?

```
#Creamos un nuevo dataframe en el cual esten agrupados los datos de la columna
#"Location_of_Breached_Information" y sume la cantidad de repeticiones que tiene
#dicho valor
a3 <- cyberb %>%
```

```
    group_by(Location_of_Breached_Information) %>%
    summarise(count = n())
#Todo aquellos valores que sean menores a 20 seran ingresados en el grupo
#"Other"  para disminuir la cantidad de observaciones de la columna de a3
#"Location_of_Breached_Information"
a3[a3$Location_of_Breached_Information=="Other", "count"] <-
  a3[a3$Location_of_Breached_Information=="Other", "count"]+sum(a3[a3$count<20,
                                                                  "count"])
#Asignamos a a3 el dataframe cuyas observaciones sean mayores o iguales a 20
#en la columna "count"
a3 <- a3[a3$count>=20, ]
```

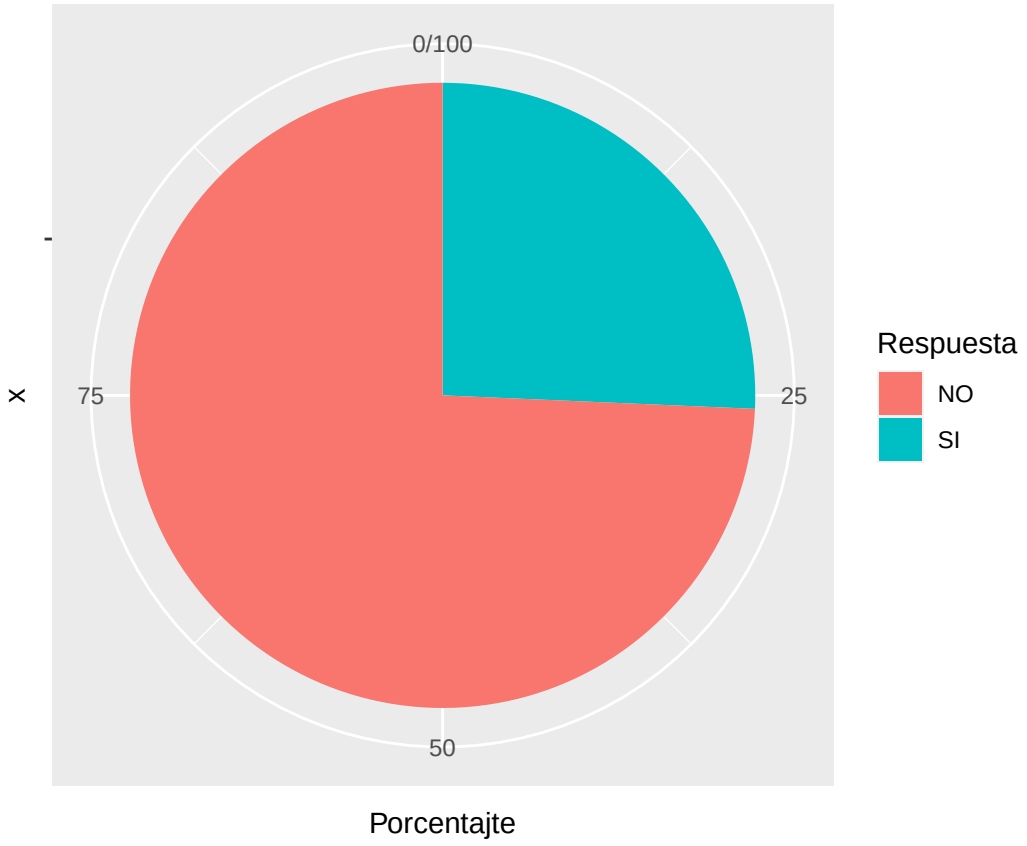



### 2.4 ¿Cuáles son los Estados más atacados?

### 2.5 ¿Cuántas empresas afectaron a terceros tras un ciberataque?

```
si <- (sum(!is.na(cyberb$Business_Associate_Involved))*100)/length(cyberb$Business_Associate_Involved)
no <- (sum(is.na(cyberb$Business_Associate_Involved))*100)/length(cyberb$Business_Associate_Involved)

a5 <- data.frame(
  Respuesta=c("SI","NO"),
  Porcentajte=c(si,no)
)
```




### 2.6 Las 10 empresas que tuvieron la mayor cantidad de afectados

```
a6 <- (data.frame (ID = paste("E",c(1:length(cyberb$Name_of_Covered_Entity)), sep = ""),
                   Entidad = cyberb$Name_of_Covered_Entity,
                   Individuos_afectados = trunc((cyberb$Individuals_Affected)/1000))) %>%
  arrange(desc(Individuos_afectados)) %>%
  slice(1:10)
```

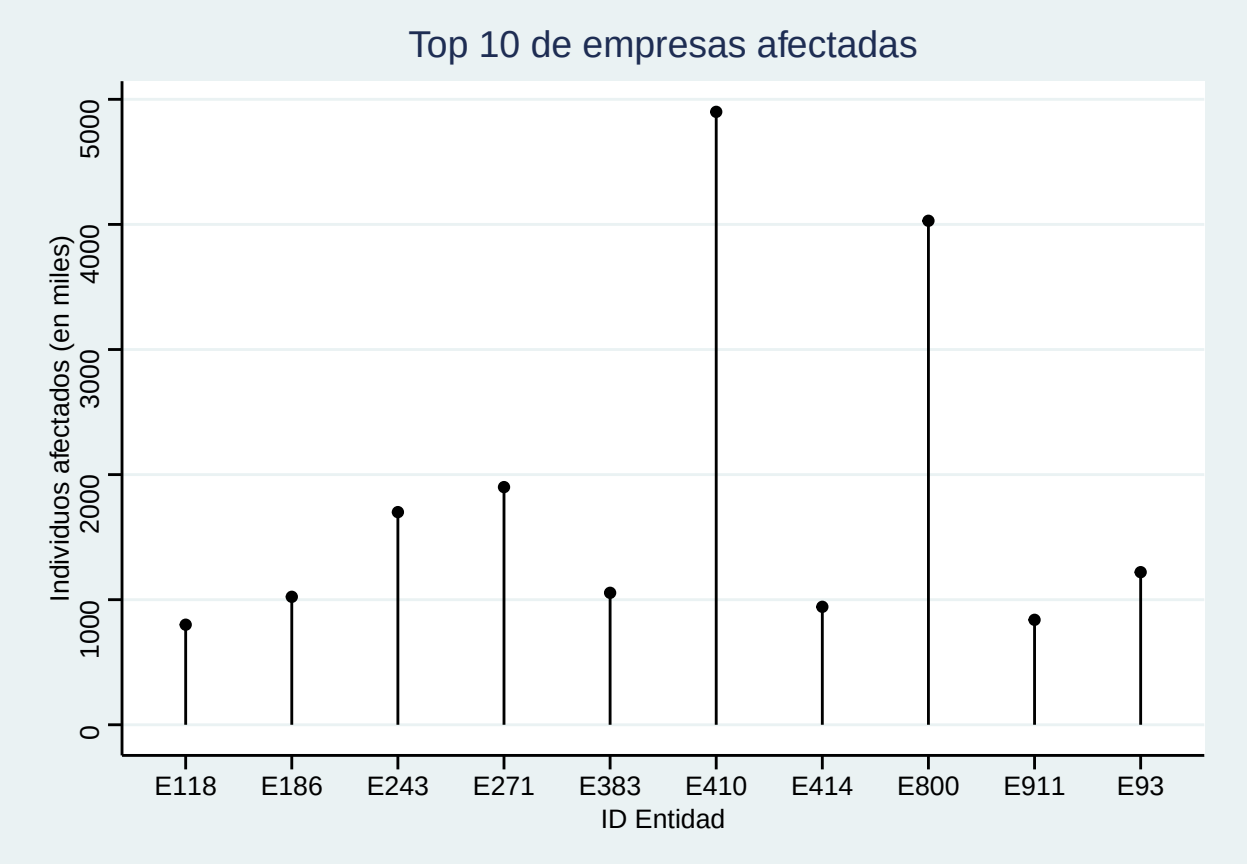



Table 2: Top 5 empresas afectadas

| ID | Entidad | Individuos_afectados |
|---|---|---|
| E410 | TRICARE Management Activity (TMA) | 4900 |
| E800 | Advocate Health and Hospitals Corporation, d/b/a Advocate Medical Group | 4029 |
| E271 | Health Net, Inc. | 1900 |
| E243 | New York City Health & Hospitals Corporation's North Bronx Healthcare Network | 1700 |
| E93 | AvMed, Inc. | 1220 |
| E383 | The Nemours Foundation | 1055 |
| E186 | BlueCross BlueShield of Tennessee, Inc. | 1023 |
| E414 | Sutter Medical Foundation | 943 |
| E911 | Horizon Healthcare Services, Inc., doing business as Horizon Blue Cross Blue Shield of New Jersey, and its affiliates | 839 |
| E118 | South Shore Hospital | 800 |