1. Análisis Exploratorio:

1.1. Bibliotecas:

Para el análisis exploratorio decidimos usar las bibliotecas de tidyverse, skimr y nortest, dado que nos sirven para el análisis de datos, obtención de resúmenes concisos de los datos y para realizar pruebas de normalidad.

1.2. Función de análisis exploratorio

La función `analisis_exploratorio` se define para analizar una columna del conjunto de datos, la cuál realiza las siguientes operaciones para conocer las cualidades de los distintos datos de esta.

- `Tipo y subtipo de dato:` Esta parte del código determina si cada columna es cualitativa o cuantitativa, y dentro de estas categorías, nos ayuda a determinar si el subtipo del atributo es discreta, continua, nominal u otra.
- `Niveles y frecuencias:` Esta parte del código es especial para datos cualitativos. Se obtienen los niveles y frecuencias de cada categoría. Por ejemplo, "Muertes por bombardeo nuclear : 9".
- `Porcentaje de valores perdidos:` Para cada columna de atributos del dataset, calcula el porcentaje de valores NA o vacíos.
- `Valores permitidos:` Esta parte define el rango de datos que son permitidos por el atributo, es decir, si son enteros, flotantes, cadenas, etc.
- `Estadísticas para la descripción:` Esta parte del código, es exclusiva para datos del tipo cuantitativos. Calcula el mínimo,

máximo, media y desviación estándar del conjunto de datos de la columna.

- `Evaluación de la distribución:` Esta parte realiza una prueba de Anderson-Darling para evaluar si la distribución de los datos es normal o no. Se puede usar otros métodos tanto visuales como pruebas para determinar si es otro tipo de distribución.

- `Valores atípicos:` Calcula la cantidad de valores atípicos usando el método del rango intercuartílico (IQR). Así pudiendo analizar esos casos y/o tener en cuenta el dato para la imputación de valores faltantes.

Se aplica la función `analisis_exploratorio` a cada columna del conjunto de datos original usando lapply. Por otro lado, construye una cadena de texto con todos los resultados del análisis y guarda esta cadena en un archivo de texto llamado `analisis_exploratorio.txt`.

1.3. Código Análisis Exploratorio:

```
library(tidyverse)
library(skimr)
library(nortest)

# Cargar el archivo CSV
data <- read.csv("path/globalterrorismdb_0718dist.csv",
   sep=",", header = T, stringsAsFactors = F ,encoding = "
   ISO-8859-1")

# Seleccionar las columnas de interes (69 a 102)
#selected_data <- data[, 69:102]

# Funcion para analizar cada columna
analisis_exploratorio <- function(column) {
  result <- list()
```

```r

  # Tipo de dato
  get_type <- function() {
    if (is.factor(column)) {
      return("Cualitativo")
    } else if (is.numeric(column)) {
      return("Cuantitativo")
    } else {
      return("Cualitativo")
    }
  }

  result$Type <- get_type()

  # Subtipo del datp
  get_subtype <- function() {
    if (is.numeric(column)) {
      if (all(!is.na(column) & column %% 1 == 0))  {
        return("Discreto")
      } else {
        return("Continuo")
      }
    } else if (is.character(column)) {
      return("Nominal")
    } else {
      return("Otro")
    }
  }

  result$SubType <- get_subtype()

  # Sacamos los niveles y frecuencia
  if (any(result$Type %in% c("Cualitativo", "factor", "
     character"))) {
    result$Levels <- levels(factor(column))
    result$Frequencies <- table(column)
  }

```

```r
  # Numero de porcentaje de valores perdidos
  result$MissingPercentage <- sum(is.na(column) | column == "") / length(column) * 100

  # Sacamos los valores permitidos:
  get_allowed_values <- function() {
    if (is.numeric(column)) {
      if (all(!is.na(column) & column %% 1 == 0))  {
        return("Enteros")
      } else {
        return("Flotantes")
      }
    } else if (is.character(column)) {
      return("Cadenas")
    } else {
      return("Otro")
    }
  }
  result$AllowedValues = get_allowed_values()

  # Sacamos el min, max, media y desviacion estandar
  if (any(result$Type %in% c("Cuantitativo","integer", "numeric"))) {
    result$Min <- min(column, na.rm = TRUE)
    result$Max <- max(column, na.rm = TRUE)
    result$Mean <- mean(column, na.rm = TRUE)
    result$SD <- sd(column, na.rm = TRUE)

    # Evaluacion de la distribucion
    if (length(column[!is.na(column)]) > 2) {
      ad_test <- ad.test(column[!is.na(column)])
      result$Distribution <- ifelse(ad_test$p.value > 0.05, "Normal", "Non-Normal")
    } else {
      result$Distribution <- NA
    }

    # valores atipicos
```

```
    Q1 <- quantile(column, 0.25, na.rm = TRUE)
    Q3 <- quantile(column, 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1
    result$Outliers <- sum(column < (Q1 - 1.5 * IQR) |
       column > (Q3 + 1.5 * IQR), na.rm = TRUE)
  }

  return(result)
}

# Aplicar la funcion a cada columna y almacenar los
   resultados
analysis_results <- lapply(data, analisis_exploratorio)

# Imprimir los resultados
#print(analysis_results)

# Guardar datos
output_text <- ""
for (col in names(analysis_results)) {
  output_text <- paste(output_text, "\n######\nColumna:",
     col, "\n", sep = "")
  output_text <- paste(output_text, paste(names(analysis_
     results[[col]]), analysis_results[[col]], sep = ": ",
      collapse = "\n"), sep = "\n")
}

# Guardar los resultados en un archivo de texto
write(output_text, file = "./analisis_exploratorio.txt")
```

1.4. Tabla con resultados del análisis exploratorio:

A continuación se muestra una tabla que representa un condensado del el análisis exploratorio realizado con el código anteriormente enunciado.

Columna	Type	SubType	LevelFrecuency	MissingPercentage	AllowedValues	Min	Max	Mean	SD	Distribution	Outliers
eventid	Cuantitativo	Discreto	No Aplica	0	Enteros	197000000001	201712310032	200270523949.246	1325957057.16345	Geometrica	FALSE
iyear	Cuantitativo	Discreto	No Aplica	0	Enteros	1970	2017	2002.63899697839	13.2594304662506	Geometrica	FALSE
imonth	Cuantitativo	Discreto	No Aplica	0	Enteros	0	12	6.46727686016368	3.38830339448391	Geometrica	FALSE
iday	Cuantitativo	Discreto	No Aplica	0	Enteros	0	31	15.5056441981166	8.81404475236334	Geometrica	FALSE
approxdate	Cualitativo	Nominal	Muchas relaciones	94.9149930376298	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
extended	Cuantitativo	Discreto	No Aplica	0	Enteros	0	1	0.0453462196806666	0.208062919098971	Bernoulli	TRUE
resolution	Cualitativo	Nominal	Muchas relaciones	98.7781453126462	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
country	Cuantitativo	Discreto	No Aplica	0	Enteros	4	1004	131.968501466776	112.414535335087	Geometrica	TRUE
country_txt	Cualitativo	Nominal	Muchas relaciones	0	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
region	Cuantitativo	Discreto	No Aplica	0	Enteros	1	12	7.16093807618429	2.93340791490636	Geometrica	FALSE
region_txt	Cualitativo	Nominal	Muchas relaciones	0	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
provstate	Cualitativo	Nominal	Muchas relaciones	0.23171208260178	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
city	Cualitativo	Nominal	Muchas relaciones	0.238867087527726	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
latitude	Cuantitativo	Continuo	No Aplica	2.50755403404682	Flotantes	-53.154613	74.633553	23.4983429592853	18.5692424210256	Normal	TRUE
longitude	Cuantitativo	Continuo	No Aplica	2.50810441904112	Flotantes	-86185896	179.366667	-458.69565302484	204778.988611396	Normal	TRUE
specificity	Cuantitativo	Continuo	No Aplica	0.00330230996582109	Flotantes	1	5	1.45145168836173	0.995429521505464	Normal	TRUE
vicinity	Cuantitativo	Discreto	No Aplica	0	Enteros	-9	1	0.068297273943123 2	0.284552858506402	Bernoulli	TRUE
location	Cualitativo	Nominal	Muchas relaciones	69.4563847411264	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
summary	Cualitativo	Nominal	Muchas relaciones	36.3964092882972	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
crit1	Cuantitativo	Discreto	No Aplica	0	Enteros	0	1	0.988529976718715	0.106482506792187	Bernoulli	TRUE
crit2	Cuantitativo	Discreto	No Aplica	0	Enteros	0	1	0.993092668321491	0.0828230535667552	Bernoulli	TRUE
crit3	Cuantitativo	Discreto	No Aplica	0	Enteros	0	1	0.875668029786836	0.329960801646127	Bernoulli	TRUE
doubterr	Cuantitativo	Continuo	No Aplica	0.00055038499430351 5	Flotantes	-9	1	-0.523171335791733	2.45581906434694	Normal	TRUE
alternative	Cuantitativo	Continuo	No Aplica	84.0327809302607	Flotantes	1	5	1.29292337389266	0.703728612195706	Normal	TRUE
alternative_txt	Cualitativo	Nominal	Muchas relaciones	84.0327809302607	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
multiple	Cuantitativo	Continuo	No Aplica	0.00055038499430351 5	Flotantes	0	1	0.13777313005669	0.344662659001423	Normal	TRUE
success	Cuantitativo	Discreto	No Aplica	0	Enteros	0	1	0.889598273992658	0.313390691399009	Bernoulli	TRUE
suicide	Cuantitativo	Discreto	No Aplica	0	Enteros	0	1	0.0365070366721522	0.187548571150398	Bernoulli	TRUE
attacktype1	Cuantitativo	Discreto	No Aplica	0	Enteros	1	9	3.24754665888789	1.91577151399266	Geometrica	TRUE
attacktype1_txt	Cualitativo	Nominal	Muchas relaciones	0	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
attacktype2	Cuantitativo	Continuo	No Aplica	96.5248691459676	Flotantes	1	9	3.71951219512195	2.27202269734836	Normal	FALSE
attacktype2_txt	Cualitativo	Nominal	Muchas relaciones	96.5248691459676	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
attacktype3	Cuantitativo	Continuo	No Aplica	99.7644352224381	Flotantes	1	8	5.24532710280374	2.24664238375437	Normal	FALSE
attacktype3_txt	Cualitativo	Nominal	Muchas relaciones	99.7644352224381	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
targtype1	Cuantitativo	Discreto	No Aplica	0	Enteros	1	22	8.43971908349891	6.65383774489877	Non-Normal	FALSE
targtype1_txt	Cualitativo	Nominal	Muchas relaciones	0	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
targsubtype1	Cuantitativo	Continuo	No Aplica	5.70914354591036	Flotantes	1	113	46.971474100795	30.9533569707158	Asimetrica	FALSE
targsubtype1_txt	Cualitativo	Nominal	Muchas relaciones	5.70914354591036	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
corp1	Cualitativo	Nominal	Muchas relaciones	23.4166799676374	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
target1	Cualitativo	Nominal	Muchas relaciones	0.348944086388429	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
natlty1	Cuantitativo	Continuo	No Aplica	0.85805020611918	Flotantes	4	1004	127.686441054338	89.2991199397951	Non-Normal	TRUE
natlty1_txt	Cualitativo	Nominal	Muchas relaciones	0.85805020611918	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
targtype2	Cuantitativo	Continuo	No Aplica	93.8665096234816	Flotantes	1	22	10.2472182340273	5.70907599826356	Non-Normal	FALSE
targtype2_txt	Cualitativo	Nominal	Muchas relaciones	93.8665096234816	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
targsubtype2	Cuantitativo	Continuo	No Aplica	94.1191363358669	Flotantes	1	113	55.3116518483856	25.6403103477152	Normal	FALSE
targsubtype2_txt	Cualitativo	Nominal	Muchas relaciones	94.1191363358669	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
corp2	Cualitativo	Nominal	Muchas relaciones	94.4317550126313	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
target2	Cualitativo	Nominal	Muchas relaciones	93.9347573627753	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
natlty2	Cuantitativo	Continuo	No Aplica	94.0404312816815	Flotantes	4	1004	131.179442186923	125.951484772792	Asimetrica	TRUE
natlty2_txt	Cualitativo	Nominal	Muchas relaciones	94.0404312816815	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
targtype3	Cuantitativo	Continuo	No Aplica	99.3527472466991	Flotantes	1	22	10.0212585034014	5.72344700017021	Normal	FALSE
targtype3_txt	Cualitativo	Nominal	Muchas relaciones	99.3527472466991	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
targsubtype3	Cuantitativo	Continuo	No Aplica	99.396227661249	Flotantes	1	113	55.5487693710119	26.2889551209084	Normal	FALSE
targsubtype3_txt	Cualitativo	Nominal	Muchas relaciones	99.396227661249	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
corp3	Cualitativo	Nominal	Muchas relaciones	99.4353049958446	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
target3	Cualitativo	Nominal	Muchas relaciones	99.3532976316934	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
natlty3	Cuantitativo	Continuo	No Aplica	99.3687084115339	Flotantes	4	1004	144.564952048823	163.299294573692	Non-Normal	TRUE
natlty3_txt	Cualitativo	Nominal	Muchas relaciones	99.3687084115339	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
gname	Cualitativo	Nominal	Muchas relaciones	0	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
gsubname	Cualitativo	Nominal	Muchas relaciones	96.7582323835523	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
gname2	Cualitativo	Nominal	Muchas relaciones	98.892075006467	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
gsubname2	Cualitativo	Nominal	Muchas relaciones	99.9119384009114	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
gname3	Cualitativo	Nominal	Muchas relaciones	99.8216752618457	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
gsubname3	Cualitativo	Nominal	Muchas relaciones	99.9889923001139	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
motive	Cualitativo	Nominal	Muchas relaciones	72.17198430302	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
guncertain1	Cuantitativo	Continuo	No Aplica	0.209146297835336	Flotantes	0	1	0.081440177374787	0.273510671600518	Normal	TRUE
guncertain2	Cuantitativo	Continuo	No Aplica	98.9239973361366	Flotantes	0	1	0.265473145780051	0.441697802219161	Normal	FALSE
guncertain3	Cuantitativo	Continuo	No Aplica	99.8238768018229	Flotantes	0	1	0.19375	0.395854300165159	Normal	TRUE
individual	Cuantitativo	Discreto	No Aplica	0	Enteros	0	1	0.00295006356946684	0.0542344621370945	Sesgada der	TRUE
nperps	Cuantitativo	Continuo	No Aplica	39.1406288698945	Flotantes	-99	25000	-65.3611543192013	216.5366334108	Sesgada izq	TRUE
nperpcap	Cuantitativo	Continuo	No Aplica	38.245702869157	Flotantes	-99	406	-1.51772695673874	12.8303464785711	Sesgada izq	TRUE
claimed	Cuantitativo	Continuo	No Aplica	36.3914558233484	Flotantes	-9	1	0.0496664388125049	1.09319524256963	Geometrica	TRUE
claimmode	Cuantitativo	Continuo	No Aplica	89.497003153706	Flotantes	1	10	7.02284756065608	2.4768505658773	Non-Normal	TRUE
claimmode_txt	Cualitativo	Nominal	Muchas relaciones	89.497003153706	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
claim2	Cuantitativo	Continuo	No Aplica	98.9597723607664	Flotantes	-9	1	0.247619047619048	0.974017751403902	Non-Normal	TRUE
claimmode2	Cuantitativo	Continuo	No Aplica	99.660962843509	Flotantes	1	10	7.17694805194805	2.78372525228763	Sesgada izq	FALSE
claimmode2_txt	Cualitativo	Nominal	Muchas relaciones	99.660962843509	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
claim3	Cuantitativo	Continuo	No Aplica	99.8249775718115	Flotantes	0	1	0.411949685534591	0.492961792323454	Asimetrica	FALSE
claimmode3	Cuantitativo	Continuo	No Aplica	99.9267987957576	Flotantes	1	10	6.72932330827068	2.90800311457482	Sesgada izq	FALSE
claimmode3_txt	Cualitativo	Nominal	Muchas relaciones	99.9267987957576	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
compclaim	Cuantitativo	Continuo	No Aplica	97.3366870125653	Flotantes	-9	1	-6.29634221946683	4.23461994131615	Non-Normal	FALSE
weaptype1	Cuantitativo	Discreto	No Aplica	0	Enteros	1	13	6.44732540412018	2.17343478048134	Non-Normal	TRUE
weaptype1_txt	Cualitativo	Nominal	Muchas relaciones	0	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
weapsubtype1	Cuantitativo	Continuo	No Aplica	11.4303955616954	Flotantes	1	31	11.1171616238822	6.49561155554381	Asimetrica	FALSE

weapsubtype1_txt	Cualitativo	Nominal	Muchas relaciones	11.4303955616954	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
weaptype2	Cuantitativo	Continuo	No Aplica	92.7750961797777	Flotantes	1	13	6.81252380589624	2.27708144006315	Asimetrica	TRUE
weaptype2_txt	Cualitativo	Nominal	Muchas relaciones	92.7750961797777	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
weapsubtype2	Cuantitativo	Continuo	No Aplica	93.6474563957488	Flotantes	1	31	10.7540287645122	7.5945741948772	Non-Normal	FALSE
weapsubtype2_txt	Cualitativo	Nominal	Muchas relaciones	93.6474563957488	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
weaptype3	Cuantitativo	Continuo	No Aplica	98.9746327556125	Flotantes	2	13	6.91143317230274	2.17795646114256	Non-Normal	FALSE
weaptype3_txt	Cualitativo	Nominal	Muchas relaciones	98.9746327556125	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
weapsubtype3	Cuantitativo	Continuo	No Aplica	99.0681982046441	Flotantes	1	28	11.6432368576491	8.4931663206445	Non-Normal	FALSE
weapsubtype3_txt	Cualitativo	Nominal	Muchas relaciones	99.0681982046441	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
weaptype4	Cuantitativo	Continuo	No Aplica	99.9598218954158	Flotantes	5	12	6.24657534246575	1.50721249255601	Non-Normal	TRUE
weaptype4_txt	Cualitativo	Nominal	Muchas relaciones	99.9598218954158	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
weapsubtype4	Cuantitativo	Continuo	No Aplica	99.9614730503988	Flotantes	2	28	10.8428571428571	8.19267207792412	Non-Normal	FALSE
weapsubtype4_txt	Cualitativo	Nominal	Muchas relaciones	99.9614730503988	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
weapdetail	Cualitativo	Nominal	Muchas relaciones	37.2445525645189	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
nkill	Cuantitativo	Continuo	No Aplica	5.67612044625215	Flotantes	0	1570	2.40327229866144	11.5457405603185	Geometrica	TRUE
nkillus	Cuantitativo	Continuo	No Aplica	35.4701113428843	Flotantes	0	1360	0.045980638833212 5	5.68185442204366	Geometrica	TRUE
nkillter	Cuantitativo	Continuo	No Aplica	36.8526784485748	Flotantes	0	500	0.508057838634046	4.19993703761121	Asimetrica	TRUE
nwound	Cuantitativo	Continuo	No Aplica	8.97732964208464	Flotantes	0	8191	3.16766840004837	35.9493918057585	Sesgada der	TRUE
nwoundus	Cuantitativo	Continuo	No Aplica	35.6110099014261	Flotantes	0	751	0.0389438323261161	3.05736149781253	Sesgada Izq	TRUE
nwoundte	Cuantitativo	Continuo	No Aplica	38.055269661128	Flotantes	0	200	0.107163165938089	1.48888121157181	Sesgada izq	TRUE
property	Cuantitativo	Discreto	No Aplica	0	Enteros	-9	1	-0.544556417213841	3.12288900003788	Sesgada izq	TRUE
propextent	Cuantitativo	Continuo	No Aplica	64.7395853399453	Flotantes	1	4	3.29540310622024	0.486911871066809	Geometrica	TRUE
propextent_txt	Cualitativo	Nominal	Muchas relaciones	64.7395853399453	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
propvalue	Cuantitativo	Continuo	No Aplica	78.5410394571002	Flotantes	-99	2700000000	208811.86872733	15524630.3114258	Non-Normal	TRUE
propcomment	Cualitativo	Nominal	Muchas relaciones	68.1002361151626	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
ishostkid	Cuantitativo	Continuo	No Aplica	0.0979685289860257	Flotantes	-9	1	0.0590536215036939	0.46124428237116	Sesgada izq	TRUE
nhostkid	Cuantitativo	Continuo	No Aplica	92.5301748573127	Flotantes	-99	17000	4.5332301797819	202.316385835137	Asimetrica	TRUE
nhostkidus	Cuantitativo	Continuo	No Aplica	92.5604460319994	Flotantes	-99	86	-0.353998668343567	6.83564459438277	Asimetrica	TRUE
nhours	Cuantitativo	Continuo	No Aplica	97.7637857681448	Flotantes	-99	999	-46.7939330543933	82.8004052186323	Sesgada izq	TRUE
ndays	Cuantitativo	Continuo	No Aplica	95.5286723062782	Flotantes	-99	2454	-32.5163712456918	121.209205116511	Sesgada izq	TRUE
divert	Cualitativo	Nominal	Muchas relaciones	99.8216752618457	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
kidhijcountry	Cualitativo	Nominal	Muchas relaciones	98.1809775938269	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
ransom	Cuantitativ	Continu	No Aplic	57.410658755799	Flotante	-		-0.14581098719323	1.2078607372127	Asimetrica	TRUE
ransomamt	Cuantitativo	Continuo	No Aplica	99.2569802576903	Flotantes	-99	1000000000	3172529.88717778	30211571.2702668	Asimetrica	TRUE
ransomamtus	Cuantitativo	Continuo	No Aplica	99.6901332482071	Flotantes	-99	132000000	578486.530461812	7077923.89058757	Asimetrica	TRUE
ransompaid	Cuantitativo	Continuo	No Aplica	99.5740020144091	Flotantes	-99	275000000	717943.701485788	10143919.926645	Asimetrica	TRUE
ransompaidus	Cuantitativo	Continuo	No Aplica	99.6961874831445	Flotantes	-99	48000	240.378623188406	2940.96729333219	Asimetrica	TRUE
ransomnote	Cualitativo	Nominal	Muchas relaciones	99.717102112928	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
hostkidoutcome	Cuantitativo	Continuo	No Aplica	93.9507185276101	Flotantes	1	7	4.6292421071786	2.03535988520627	Sesgada der	FALSE
hostkidoutcome_txt	Cualitativo	Nominal	Muchas relaciones	93.9507185276101	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
nreleased	Cuantitativo	Continuo	No Aplica	94.2759960592434	Flotantes	-99	2769	-29.0182692307692	65.7201187312035	Asimetrica	TRUE
addnotes	Cualitativo	Nominal	Muchas relaciones	84.4301588961479	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
scite1	Cualitativo	Nominal	Muchas relaciones	36.430533157944	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
scite2	Cualitativo	Nominal	Muchas relaciones	57.6572312332476	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
scite3	Cualitativo	Nominal	Muchas relaciones	76.0494465878882	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
dbsource	Cualitativo	Nominal	Muchas relaciones	0	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica
INT_LOG	Cuantitativo	Discreto	No Aplica	0	Enteros	-9	1	-4.54373083972239	4.54354684888787	Geometrica	FALSE
INT_IDEO	Cuantitativo	Discreto	No Aplica	0	Enteros	-9	1	-4.46439834664348	4.63715195724299	Geometrica	FALSE
INT_MISC	Cuantitativo	Discreto	No Aplica	0	Enteros	-9	1	0.0900099619683969	0.568457289734643	Geometrica	TRUE
INT_ANY	Cuantitativo	Discreto	No Aplica	0	Enteros	-9	1	-3.94595219355939	4.69132464123207	Geometrica	FALSE
related	Cualitativo	Nominal	Muchas relaciones	86.2194605126286	Cadenas	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica	No Aplica

2. Preprocesamiento de Datos:

1. eventid
    - Significado: Identificador del evento
    - Atributo seleccionado: Si, para mantener el identificador
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: NA
    - Normalización: NA
2. iyear
    - Significado: Año del evento
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA
3. imonth
    - Significado: Mes del evento
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA

4. iday

    - Significado: Día del evento
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA

5. approxdate

    - Significado: Aproximación de la fecha cuando esta no es clara
    - Atributo seleccionado: no
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA

    extended

    - Significado: Duración mayor o menor a 24 horas
    - Atributo seleccionado: Si, para calcular la duración del evento
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA

resolution

- Significado: Fecha de la resolución del conflicto
- Atributo seleccionado: no
- Imputación de valores perdidos: NA
- Eliminación de valores atípicos: NA
- Discretización de atributos numéricos: NA
- Normalización: NA

country

- Significado:Número identificador del País
- Atributo seleccionado: Si, dato geográfico
- Imputación de valores perdidos: criterios IQR
- Eliminación de valores atípicos: NA
- Discretización de atributos numéricos: NA
- Normalización: NA

countrytxt

- Significado:País
- Atributo seleccionado: no
- Imputación de valores perdidos: NA
- Eliminación de valores atípicos: NA
- Discretización de atributos numéricos: NA
- Normalización: NA

regiontxt

- Significado:Región del mundo
- Atributo seleccionado: Si, dato geografico
- Imputación de valores perdidos: NA
- Eliminación de valores atípicos: NA
- Discretización de atributos numéricos: NA
- Normalización: NA

provstate

- Significado:No
- Atributo seleccionado: no
- Imputación de valores perdidos: Relleno de dato cualitativo
- Eliminación de valores atípicos: NA
- Discretización de atributos numéricos: NA
- Normalización: NA

city

- Significado:Ciudad
- Atributo seleccionado: no
- Imputación de valores perdidos: Relleno de dato cualitativo
- Eliminación de valores atípicos: NA
- Discretización de atributos numéricos: NA
- Normalización: NA

latitude

- Significado:latitud geografica

- Atributo seleccionado: si, dato geografico
- Imputación de valores perdidos: media
- Eliminación de valores atípicos: criterios IQR
- Discretización de atributos numéricos: NA
- Normalización: NA

longitude

- Significado:longitud geografica
- Atributo seleccionado: si, dato geografico
- Imputación de valores perdidos: media
- Eliminación de valores atípicos: criterios IQR
- Discretización de atributos numéricos: NA
- Normalización: NA

specificity

- Significado: Que tan específica es la locación
- Atributo seleccionado: Si, especificidad geografica
- Imputación de valores perdidos: moda
- Eliminación de valores atípicos: criterios IQR
- Discretización de atributos numéricos: NA
- Normalización: NA

6. vicinity
    - Significado: Dentro de la ciudad o afuera de a ciudad.
    - Atributo seleccionado: Si, especificidad geografica

- Imputación de valores perdidos: NA
- Eliminación de valores atípicos: NA
- Discretización de atributos numéricos: NA
- Normalización: NA

7. location
    - Significado: Texto sobre la locación del evento
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA
8. summary
    - Significado: Texto recopilatorio del evento
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA
9. crit1
    - Significado: Criterio de Motivos (primario)
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA

- Eliminación de valores atípicos: NA
- Discretización de atributos numéricos: NA
- Normalización: NA

10. crit2
    - Significado: Criterio de Motivos (secundario)
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA
11. crit3
    - Significado: Criterio de Motivos (terciario)
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA
12. doubtterr
    - Significado: Dudas de si es un acto de terrorismo
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA

- Discretización de atributos numéricos: NA
- Normalización: NA

13. alternative

    - Significado: Categorización del incidente en caso de no ser terrorismo
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA

14. alternativetxt

    - Significado: Categorización del incidente en caso de no ser terrorismo, textual
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA

15. multiple

    - Significado: Multiples ataques en el atentado
    - Atributo seleccionado: No
    - Imputación de valores perdidos: NA Eliminación de valores atípicos:

- NA
- Discretización de atributos numéricos: NA
- Normalización: NA

16. success

    - Significado: Exito del atentado
    - Atributo seleccionado: No Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA Normalización: NA

17. suicide

    - Significado: Intención del perpetrador de cometer suicidio
    - Atributo seleccionado: Si, indicador de intención del perpetrador
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA

18. attacktype1

    - Significado: Tipo de ataque (principal), textual
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA

- Discretización de atributos numéricos: NA
- Normalización: NA

19. attacktype1txt
    - Significado: Tipo de ataque (principal), textual
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: NA
    - Eliminación de valores atípicos: NA
    - Discretización de atributos numéricos: NA
    - Normalización: NA
20. attacktype2
    - Significado: Tipo de ataque (secundario)
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: moda
    - Eliminación de valores atípicos: criterios IQR
    - Discretización de
    - atributos numéricos: NA
    - Normalización: NA
21. attacktype2txt
    - Significado: Tipo de ataque (secundario), textual
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: Texto usando la moda
    - Eliminación de valores atípicos: criterios IQR

- Discretización de atributos numéricos: NA
- Normalización: NA

22. attacktype3
    - Significado: Tipo de ataque (terciario)
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: moda
    - Eliminación de valores atípicos: criterios IQR
    - Discretización de atributos numéricos: NA
    - Normalización: NA
23. targtype1
    - Significado: numero que indica a que tipo de ataque son relacionadas las víctimas, hay 22 categorías
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: moda
    - Eliminación de valores atípicos: criterios IQR
    - Discretización de atributos numéricos: NA
    - Normalización: NA

    Significado: Atributo seleccionado: Si
    Imputación de valores perdidos: No, el atributo no tiene valores perdidos
    Eliminación de valores atípicos: No, el atributo no tiene valores atípicos
    Discretización de atributos numéricos: No
    Normalización: No

targtype1txt
Significado: Representa el tipo de objetivo del ataque, descripción textual del ataque
Atributo seleccionado: Si
Imputación de valores perdidos: No, el atributo no tiene valores perdidos
Eliminación de valores atípicos: No, el atributo no tiene valores atípicos
Discretización de atributos numéricos: No
Normalización: No

targsubtype1
Significado: valores enteros que expresan una categoría de la columna targsubtype1txt
Atributo seleccionado: Si
Imputación de valores perdidos: Si, usamos la media pues solo había el 6 % de datos perdidos
Eliminación de valores atípicos: No, el atributo no tiene valores atípicos
Discretización de atributos numéricos: No
Normalización: No

targsubtype1txt Significado: Información adicional mas especifica sobre el subtipo del objetivo
Atributo seleccionado: Si
Imputación de valores perdidos: Si, usamos la media pues solo había el 6 % de datos perdidos
Eliminación de valores atípicos: No, el atributo no tiene valores atípicos
Discretización de atributos numéricos: No
Normalización: No

corp1 Significado: Nombre de la corporación o país que fue objetivo
Atributo seleccionado: No
Imputación de valores perdidos:No
Eliminación de valores atípicos: No, se intento agrupar las categorías, pero aun hay categorías poco frecuentes
Discretización de atributos numéricos: No
Normalización: No

target1
Significado: Nombre del edificio,persona, instalación especifica que fue objetivo del ataque
Atributo seleccionado: No, no queremos tanto detalle pues los clasificadores no tendrían una buena precisión al clasificar entre tuplas tan especificas
Imputación de valores perdidos:No, al tener muchas categorías poco frecuentes y al faltar 23.40 % de los datos, no podemos imputar sin afectar los datos
Eliminación de valores atípicos: No, se intento agrupar las categorías, pero aun hay categorías poco frecuentes
Discretización de atributos numéricos: No
Normalización: No

natlty1
Significado:Un entero que se asocia con la nacionalidad de las víctimas, esto puede no ser igual al país en el que se registro el atentado
Atributo seleccionado: Si,es una variable categórica
Imputación de valores perdidos:Si, media
Eliminación de valores atípicos:si

Discretización de atributos numéricos: No
Normalización: No

natlty1txt
Significado:Nacionalidad de las víctimas, esto puede no ser igual al país en el que se registro el atentado
Atributo seleccionado: Si,es una variable categórica
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

targtype2
Significado:Las convenciones en el campo siguen "Tipo de objetivo/víctima"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

targtype2txt
Significado:Las convenciones en el campo siguen "Tipo de objetivo/víctima"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

targsubtype2
Significado:Las convenciones en el campo siguen "subtipo de objetivo/víctima"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

targsubtype2txt
Significado:Las convenciones en el campo siguen "subtipo de objetivo/víctima"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

corp2
Significado:Las convenciones en el campo siguen "nombre de la entidad"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

target2
Significado:Las convenciones en el campo siguen .ºbjetivo especifico/víctima"

Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

natlty2
Significado:Las convenciones en el campo siguen "numero de la nacionalidad del objetivo"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

natlty2_txt
Significado:Las convenciones en el campo siguen "Nnacionalidad del objetivo"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

targtype3
Significado:Las convenciones en el campo siguen "tipo de objetivo/víctima"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No

Normalización: No

targtype3_txt
Significado:Las convenciones en el campo siguen "tipo de objetivo/víctima"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

targsubtype3
Significado:Las convenciones en el campo siguen .ºbjetivo/subtipo de víctima"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

targsubtype3txt
Significado:Las convenciones en el campo siguen .ºbjetivo/subtipo de víctima"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

corp3

Significado:Las convenciones en el campo siguen "nombre de la entidad"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

target3
Significado:Las convenciones en el campo siguen .ºbjetivo especifico/víctima"

Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

natlty3
Significado:Las convenciones en el campo siguen "numero de la nacionalidad del objetivo"
Atributo seleccionado: No, es una columna "vacia"
Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

natlty3txt
Significado:Las convenciones en el campo siguen "Nacionalidad del objetivo"
Atributo seleccionado: No, es una columna "vacía"

Imputación de valores perdidos:no
Eliminación de valores atípicos:si
Discretización de atributos numéricos: No
Normalización: No

24. gname

    - Significado:Nombre del grupo que perpetuo el ataque
    - Atributo seleccionado: si
    - Imputación de valores perdidos:no, el conjunto no tiene valores perdidos
    - Eliminación de valores atípicos:no
    - Discretización de atributos numéricos: No
    - Normalización: No

25. gsubname

    - Significado:contiene calificadores adicionales o detalles sobre el nombre del grupo que llevó
    - fuera del ataque

- Atributo seleccionado: no, es una columna casi vacía
- Imputación de valores perdidos:no
- Eliminación de valores atípicos:no
- Discretización de atributos numéricos: No
- Normalización: No

26. gname2

- Significado:Este campo se utiliza para registrar el nombre del segundo autor cuando la responsabilidad por el
- El ataque se atribuye a mas de un autor.
- Atributo seleccionado: no, es una columna casi vacía
- Imputación de valores perdidos:no
- Eliminación de valores atípicos:no
- Discretización de atributos numéricos: No
- Normalización: No

27. gsubname2

    - Significado:Este campo se utiliza para registrar calificadores adicionales o detalles sobre el segundo grupo de perpetradores.
    - nombre cuando la responsabilidad del ataque se atribuye a mas de un autor
    - Atributo seleccionado: no, es una columna casi vacía
    - Imputación de valores perdidos:no
    - Eliminación de valores atípicos:no
    - Discretización de atributos numéricos: No
    - Normalización: No

28. gname3

    - Significado:Este campo se utiliza para registrar el nombre del tercer autor cuando la responsabilidad del ataque
    - se atribuye a mas de dos autores.
    - Atributo seleccionado: no, es una columna casi vacía

- Imputación de valores perdidos:no
- Eliminación de valores atípicos:no
- Discretización de atributos numéricos: No
- Normalización: No

29. gsubname3

- Significado:Este campo se utiliza para registrar calificadores adicionales de detalles sobre el tercer grupo perpetrador.
- nombre cuando la responsabilidad del ataque se atribuye a mas de dos autores.
- Atributo seleccionado: no, es una columna casi vacía
- Imputación de valores perdidos:no
- Eliminación de valores atípicos:no
- Discretización de atributos numéricos: No
- Normalización: No

30. motive

- Significado:motivo del ataque
- Atributo seleccionado: no, es una columna casi vacía, sin embargo podemos guardar la poca informacion en csv
- Imputación de valores perdidos:no
- Eliminación de valores atípicos:no
- Discretización de atributos numéricos: No
- Normalización: No

31. guncertain1

- Significado:Esta variable indica si la información reportada por las fuentes sobre
- Los nombres del grupo perpetrador se basan en especulaciones o afirmaciones de responsabilidad dudosas.
- Imputación de valores perdidos:si, media
- Eliminación de valores atípicos:no, pero los datos registrados deberían de ser correctos, por lo que no representan un problema para los modelos
- Discretización de atributos numéricos: No

- Normalización: No

32. guncertain2

    - Significado:Esta variable indica si la información reportada por las fuentes sobre
    - El nombre del grupo perpetrador se basa en especulaciones o afirmaciones de responsabilidad dudosas.

    - Atributo seleccionado: no, es una columna casi vacía

    - Imputación de valores perdidos:no

    - Eliminación de valores atípicos:no

    - Discretización de atributos numéricos: No

    - Normalización: No

33. guncertain3

    - Significado:Esta variable indica si la información reportada por las fuentes sobre
    - El nombre del grupo perpetrador se basa en especulaciones o afirmaciones de responsabilidad dudosas.

    - Atributo seleccionado: no, es una columna casi vacía

- Imputación de valores perdidos:no
- Eliminación de valores atípicos:no
- Discretización de atributos numéricos: No
- Normalización: No

34. individual
    - Significado: Numero de individuos involucrados en el ataque terrorista.
    - Atributo seleccionado: Sí
    - Imputación de valores perdidos: Sí, media
    - Eliminación de valores atípicos: Sí
    - Discretización de atributos numéricos: No
    - Normalización: No
35. nperps
    - Significado: Esta variable registra el número de individuos o personas que se identifican como los perpetradores o atacantes de un incidente
    - Atributo seleccionado: No, No, dado que tiene un indice cercano al 50 % de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No

- Normalización: No

36. nperpcap

    - Significado: número de personas que fueron tomadas como cautivas o rehenes durante el incidente o ataque.
    - Atributo seleccionado: No, dado que tiene un indice cercano al 50 % de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

37. claimed

    - Significado: Numero de personas que reclamaron el ataque.
    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

38. claimmode

    - Significado: Modo en el cual fue reclamado el ataque.
    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No

- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No
- Normalización: No

39. claimmodetxt

    - Significado: Modo en el cual fue reclamado el ataque en cadenas.
    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

40. claim2

    - Significado: Numero de personas las cuales reclamaron el ataque.
    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

41. claimmode2

    - Significado: Código en el modo en el cual fue reclamado el ataque.

- Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
- Imputación de valores perdidos: No
- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No
- Normalización: No

42. claimmode2txt

    - Significado: Código en el modo en el cual fue reclamado el ataque en cadenas de caracteres.
    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

43. claim3

    - Significado: Numero de personas las cuales reclamaron el ataque.
    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

44. claimmode3

    - Significado: Código en el modo en el cual fue reclamado el ataque.
    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

45. claimmode3txt

    - Significado: Código en el modo en el cual fue reclamado el ataque en cadenas de caracteres.
    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

46. compclaim

    - Atributo seleccionado: No, dado que tiene un índice muy grande de valores perdidos.
    - Imputación de valores perdidos: No
    - Eliminación de valores atípicos: No
    - Disratización de atributos numéricos: No

- Normalización: No

47. weaptype1

    - Significado: Tipo el código numérico del tipo armas que usaron para el ataque.
    - Atributo seleccionado: No, el atributo txt de este mismo proporciona la misma info.
    - Imputación de valores perdidos: No,
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

48. weaptype1

    txt

    - Significado: Nombre del tipo de las armas que usaron para el ataque.
    - Atributo seleccionado: Sí, contiene información completa y proporciona
    - información sobre la gravedad de los ataques
    - Imputación de valores perdidos: Sí, a los valores perdidos se les aplica el valor mas repetido.
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

49. weapsubtype1

- Significado: Tipo el código numérico del subtipo armas que usaron para el ataque.
- Atributo seleccionado: No, el atributo txt de este mismo proporciona la misma info.
- Imputación de valores perdidos: No,
- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No
- Normalización: No

50. weapsubtype1txt

    - Significado: Nombre del subtipo de las armas que usaron para el ataque.
    - Atributo seleccionado: Sí, contiene información completa y proporciona
    - información sobre la gravedad y el tipo de armas comunes de los ataques
    - Imputación de valores perdidos: Sí, a los valores perdidos se les aplica el valor mas repetido.
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

51. weaptype2

    - Significado: Tipo el código numérico del tipo armas que usaron para el ataque.
    - Atributo seleccionado: No, tiene un gran índice de valores perdidos.

- Imputación de valores perdidos: No,
- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No
- Normalización: No

52. weaptype2txt

    - Significado: Nombre del tipo de las armas que usaron para el ataque.
    - Atributo seleccionado: No, contiene gran índice de valores perdidos y en su versión 1 da mas información.
    - Atributo seleccionado: No.
    - Imputación de valores perdidos: Sí, a los valores perdidos se les aplica el valor mas repetido.
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

53. weapsubtype2

    - Significado: Tipo el código numérico del subtipo2 armas que usaron para el ataque.
    - Atributo seleccionado: No, el atributo contiene muchos valores perdidos y su versión 1 contiene mas información valiosa.
    - Imputación de valores perdidos: No,
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

54. weapsubtype2txt

    - Significado: Nombre del subtipo de las armas que usaron para el ataque.
    - Atributo seleccionado: No, contiene gran porcentaje de valores perdidos, y su versión 3 contiene información mas valiosa y completa.
    - Imputación de valores perdidos: Sí, a los valores perdidos se les aplica el valor mas repetido.
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

55. weaptype3

    - Significado: Tipo el código numérico del tipo armas que usaron para el ataque.
    - Atributo seleccionado: No, tiene un gran índice de valores perdidos.
    - Imputación de valores perdidos: No,
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

56. weaptype3txt

    - Significado: Nombre del tipo de las armas que usaron para el ataque.
    - Atributo seleccionado: No, contiene gran índice de valores perdidos y en su versión 1 da mas información.

- Atributo seleccionado: No, contiene un gran índice de valores perdidos.
- Imputación de valores perdidos: Sí, a los valores perdidos se les aplica el valor mas repetido.
- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No
- Normalización: No

57. weapsubtype3

    - Significado: Tipo el código numérico del subtipo2 armas que usaron para el ataque.
    - Atributo seleccionado: No, el atributo contiene muchos valores perdidos y su versión 1 contiene mas información valiosa.
    - Imputación de valores perdidos: No,
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

58. weapsubtype3txt

    - Significado: Nombre del subtipo de las armas que usaron para el ataque.
    - Atributo seleccionado: No, contiene gran porcentaje de valores perdidos, y su versión 3 contiene información mas valiosa y completa.
    - Imputación de valores perdidos: Sí, a los valores perdidos se les aplica el valor mas repetido.
    - Eliminación de valores atípicos: No

- Discretización de atributos numéricos: No
- Normalización: No

59. weaptype4

    - Significado: Tipo el código numérico del tipo armas que usaron para el ataque.
    - Atributo seleccionado: No, tiene un gran índice de valores perdidos.
    - Imputación de valores perdidos: No,
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

60. weaptype4txt

    - Significado: Nombre del tipo de las armas que usaron para el ataque.
    - Atributo seleccionado: No, contiene gran índice de valores perdidos y en su versión 1 da mas información.
    - Atributo seleccionado: No.
    - Imputación de valores perdidos: Sí, a los valores perdidos se les aplica el valor mas repetido.
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No
    - Normalización: No

61. weapsubtype4

- Significado: Tipo el código numérico del subtipo2 armas que usaron para el ataque.
- Atributo seleccionado: No, el atributo contiene muchos valores perdidos y su versión 1 contiene mas información valiosa.
- Imputación de valores perdidos: No,
- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No
- Normalización: No
- weapsubtype4txt itemize
- Significado: Nombre del subtipo de las armas que usaron para el ataque.
- Atributo seleccionado: No, contiene gran porcentaje de valores perdidos, y su versión 3 contiene información mas valiosa y completa.
- Imputación de valores perdidos: Sí, a los valores perdidos se les aplica el valor mas repetido.
- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No
- Normalización: No

62. weapdetail

- Significado: Detalles de las armas que usaron en el ataque.
- Atributo seleccionado: No
- Imputación de valores perdidos: No.
- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No

- Normalización: No

63. nkill

    - Significado: Numero de muertes en un ataque terrorista.
    - Atributo seleccionado: Sí, contiene una gran cantidad de datos, y posee información valiosa para tareas de predicción
    - Imputación de valores perdidos: Sí, se reemplazan por la mediana.
    - Eliminación de valores atípicos: Sí
    - Discretización de atributos numéricos: No.
    - Normalización: Sí, la normalización se realiza utilizando los métodos çenterz "scale", lo que significa que se centra alrededor de la media y se escalan por la desviación estandar.

64. nkillus

    - Significado: Numero de muertes en un ataque terrorista en EU.
    - Atributo seleccionado: No, contiene una gran cantidad de datos perdidos.
    - Imputación de valores perdidos: No.
    - Eliminación de valores atípicos: No
    - Discretización de atributos numéricos: No.
    - Normalización: No.

65. nkillter

    - Significado: Numero de muertes en un ataque terrorista en EU.

- Atributo seleccionado: No, contiene una gran cantidad de datos perdidos.
- Imputación de valores perdidos: No.
- Eliminación de valores atípicos: No
- Discretización de atributos numéricos: No.
- Normalización: No.

66. nwound

    - Significado: Numero de heridos en un ataque terrorista.
    - Atributo seleccionado: Sí, contiene una gran cantidad de datos, y posee información valiosa para tareas de predicción
    - Imputación de valores perdidos: Sí, se reemplazan por la mediana.
    - Eliminación de valores atípicos: Sí
    - Discretización de atributos numéricos: No.
    - Normalización: Sí, la normalización se realiza utilizando los métodos çenterƶ "scale", lo que significa que se centra alrededor de la media y se escalan por la desviación estandar.

67. nwoundus

    - Significado: Número de heridos ciudadanos de USA
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: Si, media
    - Eliminación de valores atípicos: Si
    - Discretización de atributos numéricos: No
    - Normalización: No

68. nwoundte

    - Significado: Número de heridos ciudadanos de USA
    - Atributo seleccionado: No, pues los heridos suelen ser danos colaterales no previstos.
    - Imputación de valores perdidos: Si, media
    - Eliminación de valores atípicos: Si
    - Discretización de atributos numéricos: No
    - Normalización: No

69. property

    - Significado: Número de heridos ciudadanos de USA
    - Atributo seleccionado: No, puesto la columna propvalue ya da esta información
    - Imputación de valores perdidos: Si, media
    - Eliminación de valores atípicos: Si
    - Discretización de atributos numéricos: No
    - Normalización: No

70. propextent

    - Significado: Número de heridos ciudadanos de USA
    - Atributo seleccionado: No, pues es discretización arbitraria
    - Imputación de valores perdidos: Si, media
    - Eliminación de valores atípicos: Si
    - Discretización de atributos numéricos: No
    - Normalización: No

71. propextent_txt

    - Significado: Número de heridos ciudadanos de USA
    - Atributo seleccionado: No, discrettización arbitraria
    - Imputación de valores perdidos:
    - Eliminación de valores atípicos: Si
    - Discretización de atributos numéricos: No
    - Normalización: No

72. propvalue

    - Significado: Número de heridos ciudadanos de USA
    - Atributo seleccionado: Si
    - Imputación de valores perdidos: Si, media
    - Eliminación de valores atípicos: Si
    - Discretización de atributos numéricos: Si, binnings
    - Normalización: No

73. propcomment

    - Significado: Número de heridos ciudadanos de USA
    - Atributo seleccionado: No, por textual
    - Imputación de valores perdidos:
    - Eliminación de valores atípicos: Si
    - Discretización de atributos numéricos: No
    - Normalización: No

74. ishostkid

- Significado: Número de heridos ciudadanos de USA
- Atributo seleccionado: No, pues solo sirve para nulos abajo
- Imputación de valores perdidos: Si, media
- Eliminación de valores atípicos: Si
- Discretización de atributos numéricos: No
- Normalización: No

75. nhostkid

    Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Si Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

nhostkidus Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No, puesto queremos resultados globales y no solo de USA Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

nhours Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Si Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

ndays Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Si Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

divert Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No, casi todos null Imputación de valores perdidos: Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

kidhijcountry Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Si Imputación de valores perdidos: Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

ransom Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No, puesto solo sirve para nulos abajo Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

ransomamt Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Casi todos nulos Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

ransomamtus Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Casi todos nulos Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

ransompaid Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Casi todos nulos Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

ransompaidus Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Casi todos nulos Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

ransomnote Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Casi todos nulos Imputación de valores perdidos: Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

hostkidoutcome Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No, misma información abajo Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

hostkidoutcome$_{t}xtSignificado$: $NúmerodeheridosciudadanosdeUSAAtributos$
$SiImputacióndevaloresperdidos$: $Eliminacióndevaloresatípicos$: $SiDiscretiz$

nreleased Significado: Número de heridos ciudadanos de USA Atributo seleccionado: Si Imputación de valores perdidos: Si, media Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

addnotes Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No Imputación de valores perdidos: Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

scite1 Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No Imputación de valores perdidos: Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

scite2 Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No Imputación de valores perdidos: Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

scite3 Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No Imputación de valores perdidos: Eliminación de valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

dbsource Significado: Número de heridos ciudadanos de USA Atributo seleccionado: No Imputación de valores perdidos: Eliminación de

valores atípicos: Si Discretización de atributos numéricos: No Normalización: No

$\mathrm{INT}_{L}OGSignificado : Númerodeheridosciudadanosde USAAtributoseleccionado :$
$SiImputacióndevaloresperdidos : Si, mediaEliminacióndevaloresatípicos : SiDiscr$

$\mathrm{INT}_{I}DEOSignificado : Númerodeheridosciudadanosde USAAtributoseleccionado :$
$SiImputacióndevaloresperdidos : Si, mediaEliminacióndevaloresatípicos : SiDiscr$

$\mathrm{INT}_{M}ISCSignificado : Númerodeheridosciudadanosde USAAtributoseleccionado :$
$SiImputacióndevaloresperdidos : Si, mediaEliminacióndevaloresatípicos : SiDiscr$

$\mathrm{INT}_{A}NYSignificado : Númerodeheridosciudadanosde USAAtributoseleccionado :$
$SiImputacióndevaloresperdidos : Si, mediaEliminacióndevaloresatípicos : SiDiscr$

```
       X                eventid              iyear
imonth
 Min.   :     1   Min.   :1.970e+11   Min.   :1970
Min.   : 0.000
 1st Qu.: 57315   1st Qu.:1.994e+11   1st Qu.:1994
1st Qu.: 4.000
 Median :100135   Median :2.011e+11   Median :2011
Median : 6.000
 Mean   : 98059   Mean   :2.005e+11   Mean   :2005
Mean   : 6.488
 3rd Qu.:141098   3rd Qu.:2.014e+11   3rd Qu.:2014
3rd Qu.: 9.000
 Max.   :181691   Max.   :2.017e+11   Max.   :2017
Max.   :12.000

      iday          approxdate          extended
resolution
 Min.   : 0.00   Length:150504       Min.   :0.00000
Length:150504
```

```
 1st Qu.: 8.00   Class :character   1st Qu.:0.00000
Class :character
 Median :15.00   Mode  :character   Median :0.00000
Mode  :character
 Mean   :15.53                      Mean   :0.04031
 3rd Qu.:23.00                      3rd Qu.:0.00000
 Max.   :31.00                      Max.   :1.00000

    country       country$_$txt            region
region$_$txt
 Min.   :  4.0   Length:150504      Min.   : 1.000
Length:150504
 1st Qu.: 83.0   Class :character   1st Qu.: 6.000
Class :character
 Median : 95.0   Mode  :character   Median : 6.000
Mode  :character
 Mean   :115.3                      Mean   : 7.235
 3rd Qu.:160.0                      3rd Qu.:10.000
 Max.   :236.0                      Max.   :12.000

  provstate             city              latitude
longitude
 Length:150504      Length:150504      Min.   :-22.02
Min.   :-458.70
 Class :character   Class :character   1st Qu.: 13.59
1st Qu.:  12.49
 Mode  :character   Mode  :character   Median : 30.51
Median :  44.19
                                      Mean   : 24.43
Mean   :  20.67
                                      3rd Qu.: 34.08
3rd Qu.:  69.83
```

```
                                                    Max.   : 65.68
Max.   : 168.32

    vicinity          location            crit1
crit2
 Min.   :-9.00000   Length:150504      Min.   :0.0000
Min.   :0.0000
 1st Qu.: 0.00000   Class :character   1st Qu.:1.0000
1st Qu.:1.0000
 Median : 0.00000   Mode  :character   Median :1.0000
Median :1.0000
 Mean   : 0.07156                      Mean   :0.9879
Mean   :0.9938
 3rd Qu.: 0.00000                      3rd Qu.:1.0000
3rd Qu.:1.0000
 Max.   : 1.00000                      Max.   :1.0000
Max.   :1.0000

     crit3          doubtterr       alternative$_$txt
multiple
 Min.   :0.0000   Min.   :-9.0000   Length:150504
Min.   :0.0000
 1st Qu.:1.0000   1st Qu.: 0.0000   Class :character
1st Qu.:0.0000
 Median :1.0000   Median : 0.0000   Mode  :character
Median :0.0000
 Mean   :0.8756   Mean   :-0.4432
Mean   :0.1406
 3rd Qu.:1.0000   3rd Qu.: 0.0000
3rd Qu.:0.0000
 Max.   :1.0000   Max.   : 1.0000
Max.   :1.0000
```

```
                    NA's   :1
NA's   :1
    success          suicide         attacktype1
attacktype1$_$txt  attacktype2
 Min.   :0.000   Min.   :0.00000   Min.   :1.000
Min.   :4.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:2.000
1st Qu.:5.000   1st Qu.:2.000
 Median :1.000   Median :0.00000   Median :3.000
Median :5.000   Median :2.000
 Mean   :0.887   Mean   :0.03938   Mean   :3.195
Mean   :5.256   Mean   :2.066
 3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:3.000
3rd Qu.:6.000   3rd Qu.:2.000
 Max.   :1.000   Max.   :1.00000   Max.   :9.000
Max.   :7.000   Max.   :9.000

  attacktype3      targtype1      targtype1$_$txt
targsubtype1
 Min.   :1.000   Min.   : 1.000   Length:150504
Min.   :  1.00
 1st Qu.:7.000   1st Qu.: 3.000   Class :character
1st Qu.: 22.00
 Median :7.000   Median : 4.000   Mode  :character
Median :  35.00
 Mean   :6.996   Mean   : 8.378
Mean   :  46.66
 3rd Qu.:7.000   3rd Qu.:14.000
3rd Qu.:  73.00
 Max.   :8.000   Max.   :22.000
Max.   :113.00
```

```
 targsubtype1$_$txt      corp1          target1
natlty1
 Min.   :  1.00   Min.   :    1   Length:150504
Min.   :  4.0
 1st Qu.: 13.00   1st Qu.:  521   Class :character
1st Qu.: 83.0
 Median : 28.00   Median : 2868   Mode  :character
Median : 95.0
 Mean   : 35.72   Mean   : 8441
Mean   :114.8
 3rd Qu.: 55.00   3rd Qu.:15621
3rd Qu.:160.0
 Max.   :112.00   Max.   :32722
Max.   :238.0

 natlty1$_$txt           gname            guncertain1
individual
 Length:150504      Length:150504      Min.   :0.00000
Min.   :0.000000
 Class :character   Class :character   1st Qu.:0.00000
1st Qu.:0.000000
 Mode  :character   Mode  :character   Median :0.00000
Median :0.000000
                                       Mean   :0.08737
Mean   :0.002817
                                       3rd Qu.:0.00000
3rd Qu.:0.000000
                                       Max.   :1.00000
Max.   :1.000000

 weaptype1$_$txt      weapsubtype1$_$txt       nkill
nwound
```

```
 Length:150504      Length:150504      Min.   : -0.20092
Min.   : -0.08344
 Class :character   Class :character   1st Qu.: -0.20092
1st Qu.: -0.08344
 Mode  :character   Mode  :character   Median : -0.20092
Median : -0.08344
                                       Mean   :  0.01593
Mean   :  0.00535
                                       3rd Qu.: -0.02287
3rd Qu.: -0.02567
                                       Max.   :139.56893
Max.   :236.51161

    nhostkid            nhours            divert
kidhijcountry
 Min.   :0.00e+00   Min.   :  0.00   Length:150504
Length:150504
 1st Qu.:0.00e+00   1st Qu.: 99.37   Class :character
Class :character
 Median :0.00e+00   Median : 99.37   Mode  :character
Mode  :character
 Mean   :9.07e-01   Mean   : 99.37
 3rd Qu.:0.00e+00   3rd Qu.: 99.37
 Max.   :1.70e+04   Max.   :768.00

   ransomamt        hostkidoutcome$_$txt    nreleased
INT$_$LOG
 Min.   : -0.0107   Length:150504       Min.   :-99.00
Min.   :-9.000
 1st Qu.: -0.0107   Class :character    1st Qu.:-36.24
1st Qu.:-9.000
 Median :  0.0000   Mode  :character    Median :-36.24
```

```
Median :-9.000
 Mean   :  0.0000                                  Mean   :-36.24
Mean   :-4.765
 3rd Qu.:  0.0000                                  3rd Qu.:-36.24
3rd Qu.: 0.000
 Max.   :385.1451                                  Max.   :151.00
Max.   :  1.000

    INT$_$IDEO        INT$_$ANY      variable$_$texto$_$agrupad
 Min.   :-9.0    Min.   :-9.00    Length:150504
 1st Qu.:-9.0    1st Qu.:-9.00    Class :character
 Median :-9.0    Median :-9.00    Mode  :character
 Mean   :-4.7    Mean   :-4.69
 3rd Qu.: 0.0    3rd Qu.: 0.00
 Max.   : 1.0    Max.   : 1.00
```

2.1. Código etapa de preprocesamiento:

```r
library(ggplot2)
library(dplyr)
library(caret)
data = read.csv(file = "E:/Descargas/Repos/AMD-2024-1/
   globalterrorismdb_0718dist.csv", sep=",", header = T,
   stringsAsFactors = F)


############################################################
###################Columnas 1-34#####################
############################################################



manejo <- function(tabla){
  ##### SELECCION DE ATRIBUTOS #####
```

```r

seleccion_de_atributos <- tabla

####### VALORES PERDIDOS #######

valores_perdidos <- function(datos) {
  resultados <- matrix(nrow = ncol(datos), ncol = 2)

  for (i in seq_along(datos)) {
    nombre_columna <- names(datos)[i]
    cantidad_perdidos <- sum(is.na(datos[, i]) | datos[,
        i] == "")
    resultados[i, ] <- c(nombre_columna, cantidad_
       perdidos)
  }

  colnames(resultados) <- c("Columna", "Valores_Perdidos
     ")
  return(resultados)
}

valores_perdidos_data <- valores_perdidos(seleccion_de_
   atributos)

# Reemplazar NA en "latitude" por la media
media_latitude <- mean(seleccion_de_atributos$latitude,
   na.rm = TRUE)
seleccion_de_atributos$latitude[is.na(seleccion_de_
   atributos$latitude)] <- media_latitude

# Reemplazar NA en "longitude" por la media
media_longitude <- mean(seleccion_de_atributos$longitude
   , na.rm = TRUE)
seleccion_de_atributos$longitude[is.na(seleccion_de_
   atributos$longitude)] <- media_longitude

# Reemplazar los valores NA en la columna 'specificity'
   con 0
```

```r
seleccion_de_atributos$specificity[is.na(seleccion_de_atributos$specificity)] <- 0

# Reemplazar NA en  por la moda
moda_attacktype2 <- as.numeric(names(sort(table(seleccion_de_atributos$attacktype2), decreasing = TRUE)[1]))
seleccion_de_atributos$attacktype2[is.na(seleccion_de_atributos$attacktype2)] <- moda_attacktype2
moda_attacktype3 <- as.numeric(names(sort(table(seleccion_de_atributos$attacktype3), decreasing = TRUE)[1]))
seleccion_de_atributos$attacktype3[is.na(seleccion_de_atributos$attacktype3)] <- moda_attacktype3

# Reemplazar las cadenas vacias en las columnas "provstate", "city", "attacktype2" y "attacktype3"
seleccion_de_atributos$provstate[seleccion_de_atributos$provstate == ""] <- "Desconocido"
seleccion_de_atributos$city[seleccion_de_atributos$city == ""] <- "Desconocido"
seleccion_de_atributos$attacktype2_txt[seleccion_de_atributos$attacktype2_txt == ""] <- "Armed Assault"
seleccion_de_atributos$attacktype3_txt[seleccion_de_atributos$attacktype3_txt == ""] <- "Facility/Infrastructure Attack"

####### VALORES ATiPICOS #######

seleccion_de_atributos2 <- seleccion_de_atributos
valores_perdidos_data <- valores_perdidos(seleccion_de_atributos2)

attacktype_herarchy = c("Assassination", "Hijacking", "Hostage Taking (Kidnapping)", "Hostage Taking (Barricade Incident)"
                        , "Bombing/Explosion", "Armed Assault", "Unarmed Assault",
```

```
                                "Facility/Infrastructure
                                Attack", "Unknown")
65
66  seleccion_de_atributos2$attacktype1_txt = as.numeric(
       factor(seleccion_de_atributos2$attacktype1_txt,
       levels=attacktype_herarchy))
67  seleccion_de_atributos2$attacktype2_txt = as.numeric(
       factor(seleccion_de_atributos2$attacktype2_txt,
       levels=attacktype_herarchy))
68  seleccion_de_atributos2$attacktype3_txt = as.numeric(
       factor(seleccion_de_atributos2$attacktype3_txt,
       levels=attacktype_herarchy))
69
70  par(mfrow = c(4, 4), mar = c(2, 2, 2, 2))
71  boxplot(seleccion_de_atributos2$country, ylab = "country
       ")
72  boxplot(seleccion_de_atributos2$latitude, ylab = "
       latitude")
73  boxplot(seleccion_de_atributos2$longitude, ylab = "
       longitude")
74  boxplot(seleccion_de_atributos2$specificity, ylab = "
       specificity")
75  boxplot(seleccion_de_atributos2$attacktype1_txt, ylab =
       "attacktype1_txt")
76  boxplot(seleccion_de_atributos2$attacktype2_txt, ylab =
       "attacktype2_txt")
77  boxplot(seleccion_de_atributos2$attacktype3_txt, ylab =
       "attacktype3_txt")
78
79  # Extraccion de valores atipicos basado en criterios IQR
80  outliers_country <- boxplot.stats(seleccion_de_
       atributos2$country)$out
81  outliers_latitude <- boxplot.stats(seleccion_de_
       atributos2$latitude)$out
82  outliers_longitude <- boxplot.stats(seleccion_de_
       atributos2$longitude)$out
83  outliers_specificity <- boxplot.stats(seleccion_de_
       atributos2$specificity)$out
```

```
84  outliers_attacktype1_txt <- boxplot.stats(seleccion_de_
       atributos2$attacktype1_txt)$out
85  outliers_attacktype2_txt <- boxplot.stats(seleccion_de_
       atributos2$attacktype2_txt)$out
86  outliers_attacktype3_txt <- boxplot.stats(seleccion_de_
       atributos2$attacktype3_txt)$out
87
88  # Identificacion de numero de indices de columnas con
       valores atipicos
89  outlier_indices_country <- which(seleccion_de_atributos2
       $country %in% outliers_country)
90  outlier_indices_latitude <- which(seleccion_de_
       atributos2$latitude %in% outliers_latitude)
91  outlier_indices_longitude <- which(seleccion_de_
       atributos2$longitude %in% outliers_longitude)
92  outlier_indices_specificity <- which(seleccion_de_
       atributos2$specificity %in% outliers_specificity)
93  outlier_indices_attacktype1_txt <- which(seleccion_de_
       atributos2$attacktype1_txt %in% outliers_attacktype1_
       txt)
94  outlier_indices_attacktype2_txt <- which(seleccion_de_
       atributos2$attacktype2_txt %in% outliers_attacktype2_
       txt)
95  outlier_indices_attacktype3_txt <- which(seleccion_de_
       atributos2$attacktype3_txt %in% outliers_attacktype3_
       txt)
96
97  valores_atipicos_country <- seleccion_de_atributos2[
       outlier_indices_country, ]$country
98  moda_country <- as.numeric(names(sort(table(seleccion_de
       _atributos2$country), decreasing = TRUE)[1]))
99  seleccion_de_atributos2$country[seleccion_de_atributos2$
       country %in% valores_atipicos_country] <- moda_
       country
100
101 valores_atipicos_latitude <- seleccion_de_atributos2[
       outlier_indices_latitude, ]$latitude
```

```
seleccion_de_atributos2$latitude[seleccion_de_atributos2$latitude %in% valores_atipicos_latitude] <- media_latitude

valores_atipicos_longitude <- seleccion_de_atributos2[outlier_indices_longitude, ]$longitude
seleccion_de_atributos2$longitude[seleccion_de_atributos2$longitude %in% valores_atipicos_longitude] <- media_longitude

valores_atipicos_specificity <- seleccion_de_atributos2[outlier_indices_specificity, ]$specificity
moda_specificity <- as.numeric(names(sort(table(seleccion_de_atributos2$specificity), decreasing = TRUE)[1]))
seleccion_de_atributos2$specificity[seleccion_de_atributos2$specificity %in% valores_atipicos_specificity] <- moda_specificity

valores_atipicos_attacktype1_txt <- seleccion_de_atributos2[outlier_indices_attacktype1_txt, ]$attacktype1_txt
moda_attacktype1_txt <- as.numeric(names(sort(table(seleccion_de_atributos2$attacktype1_txt), decreasing = TRUE)[1]))
seleccion_de_atributos2$attacktype1_txt[seleccion_de_atributos2$attacktype1_txt %in% valores_atipicos_attacktype1_txt] <- moda_attacktype1_txt

valores_atipicos_attacktype2_txt <- seleccion_de_atributos2[outlier_indices_attacktype2_txt, ]$attacktype2_txt
moda_attacktype2_txt <- as.numeric(names(sort(table(seleccion_de_atributos2$attacktype2_txt), decreasing = TRUE)[1]))
seleccion_de_atributos2$attacktype2_txt[seleccion_de_atributos2$attacktype2_txt %in% valores_atipicos_attacktype2_txt] <- moda_attacktype2_txt
```

```
118
119   valores_atipicos_attacktype3_txt <- seleccion_de_
        atributos2[outlier_indices_attacktype3_txt, ]$
        attacktype3_txt
120   moda_attacktype3_txt <- as.numeric(names(sort(table(
        seleccion_de_atributos2$attacktype3_txt), decreasing
        = TRUE)[1]))
121   seleccion_de_atributos2$attacktype3_txt[seleccion_de_
        atributos2$attacktype3_txt %in% valores_atipicos_
        attacktype3_txt] <- moda_attacktype3_txt
122
123   seleccion_de_atributos2$approxdate <- as.factor(
        seleccion_de_atributos2$approxdate)
124   seleccion_de_atributos2$resolution  <- as.factor(
        seleccion_de_atributos2$resolution )
125   seleccion_de_atributos2$country_txt <- as.factor(
        seleccion_de_atributos2$country_txt)
126   seleccion_de_atributos2$region_txt <- as.factor(
        seleccion_de_atributos2$region_txt)
127   seleccion_de_atributos2$provstate <- as.factor(seleccion
        _de_atributos2$provstate)
128   seleccion_de_atributos2$city <- as.factor(seleccion_de_
        atributos2$city)
129   seleccion_de_atributos2$location <- as.factor(seleccion_
        de_atributos2$location)
130   seleccion_de_atributos2$summary <- NULL
131   seleccion_de_atributos2$alternative_txt <- as.factor(
        seleccion_de_atributos2$alternative_txt)
132
133   return(seleccion_de_atributos2)
134 }
135
136
137 ### Manipulacion de tabla
138 data <- manejo(data)
139
140
141 manejo2 <- function(tabla){
```

```r
#hacemos una tabla con el conteo de las categorias de 1 a 22
tabla_frecuencia <- table(tabla$targtype1)
print(tabla_frecuencia)


# Frecuencia de valores en targtype1
value_counts <- table(tabla$targtype1)

# boxplot targtype1 para ver si tiene valores atipicos
boxplot(tabla$targtype1, horizontal = TRUE, main = "Boxplot de targtype1")

# Visualizacion de Frecuencia de targtype1
barplot(value_counts, main = "Frecuencia de valores en targtype1", xlab = "targtype1", ylab = "Frecuencia")



# Calcular el rango intercuartilico (IQR) para ver si targtype tiene valore
Q1 <- quantile(tabla$targtype1, 0.25)
Q3 <- quantile(tabla$targtype1, 0.75)
IQR <- Q3 - Q1

# Definir limites para identificar valores atipicos
lower_limit <- Q1 - 1.5 * IQR
upper_limit <- Q3 + 1.5 * IQR

# Identificar valores atipicos
outliers <- tabla$targtype1 < lower_limit | tabla$targtype1 > upper_limit

# Mostrar valores atipicos
outlier_values <- tabla$targtype1[outliers]
cat("Valores atipicos en targtype1:", unique(outlier_values), "\n")

```

```
174 tabla_frecuencia <- table(tabla$targtype1_txt)
175 print(tabla_frecuencia)
176
177 # Obtiene los niveles unicos en el orden en que aparecen
      los datos
178 unique_levels <- unique(tabla$targtype1_txt)
179
180 # Convierte a factor con niveles manuales
181 tabla$targtype1_txt <- factor(tabla$targtype1_txt,
      levels = unique_levels)
182
183 # Verifica que la columna haya sido convertida a factor
      con los niveles deseados
184 str(tabla$targtype1_txt)
185
186 #impime
187 print(tabla$targtype1_txt)
188
189 #resultado <- factor((Private Citizens & Property,
      Government (Diplomatic),Journalists & Media,Police,
      Utilities,Military,Government (General),Airports &
      Aircraft,Business,Educational Institution,Violent
      Political Party ,Religious Figures/Institutions,
      Unknown,Transportation,Tourists,NGO,Telecommunication
      ,Food or Water Supply,Terrorists/Non-State Militia,
      Other,Maritime, Abortion Related),levels=c("Private
      Citizens & Property","Government (Diplomatic)","
      Journalists & Media","Police","Utilities","Military
      ","Government (General)","Airports & Aircraft","
      Business","Educational Institution","Violent
      Political Party ","Religious Figures/Institutions","
      Unknown","Transportation","Tourists","NGO","
      Telecommunication","Food or Water Supply","Terrorists
      /Non-State Militia","Other","Maritime", "Abortion
      Related"))
190
191
192
```

```r

# Verificar la distribucion de categorias
tabla_frecuencia <- tabla %>%
  group_by(targtype1_txt) %>%
  summarise(frecuencia = n())

# Imprimir la tabla de frecuencias
print(tabla_frecuencia)

# Visualizar la distribucion con un grafico de barras
ggplot(tabla, aes(x = targtype1_txt)) +
  geom_bar() +
  labs(title = "Distribucion de la Variable Categorica",
       x = "Categoria",
       y = "Frecuencia")

# Identificar categorias poco frecuentes
umbral_frecuencia <- 5  # Puedes ajustar este umbral
   segun tus necesidades
categorias_poco_frecuentes <- tabla_frecuencia %>%
  filter(frecuencia < umbral_frecuencia) %>%
  pull(targtype1_txt)

# Imprimir categorias poco frecuentes
if (length(categorias_poco_frecuentes) > 0) {
  cat("Categorias poco frecuentes:", paste(categorias_
     poco_frecuentes, collapse = ", "), "\n")
} else {
  cat("No hay categorias poco frecuentes.\n")
}
#conteo de targsubtype1
tabla_frecuencia <- table(tabla$targsubtype1)
print(tabla_frecuencia)
#valores perdidos

# Contar valores perdidos
valores_perdidos <- sum(is.na(tabla$targsubtype1))

```

```r
# Imprimir la cantidad de valores perdidos
cat("Numero de valores perdidos:", valores_perdidos, "\n")


# Calcular la media de la variable
media_targsubtype1 <- ceiling(mean(tabla$targsubtype1, na.rm = TRUE))

# Imputar la media redondeada hacia arriba
tabla$targsubtype1 <- ifelse(is.na(tabla$targsubtype1), media_targsubtype1, tabla$targsubtype1)


# Frecuencia de valores en targtype1
value_counts <- table(tabla$targsubtype1)

# boxplot targtype1 para ver si tiene valores atipicos
boxplot(tabla$targsubtype1, horizontal = TRUE, main = "Boxplot de targtype1")

# Visualizacion de Frecuencia de targtype1
barplot(value_counts, main = "Frecuencia de valores en targtype1", xlab = "targtype1", ylab = "Frecuencia")

# Rellenar los valores vacios
tabla$targsubtype1_txt <- replace(tabla$targsubtype1_txt, tabla$targsubtype1_txt == "", "International Organization (peacekeeper, aid agency, compound)")


tabla_frecuencia <- table(tabla$targsubtype1_txt)
print(tabla_frecuencia)

#tabla de frecuencias para corp1
tabla_frecuencia <- table(tabla$corp1)
print(tabla_frecuencia)

```

```
260
261
262   # Calcular la frecuencia de cada categoria
263   frecuencia_categorias <- tabla %>%
264     group_by(corp1) %>%
265     summarise(frecuencia = n())
266
267   # Definir un umbral de frecuencia para categorias poco
          frecuentes
268   umbral_frecuencia <- 10  # Puedes ajustar este umbral
          segun tus necesidades
269
270   # Identificar las categorias poco frecuentes
271   categorias_poco_frecuentes <- frecuencia_categorias %>%
272     filter(frecuencia < umbral_frecuencia) %>%
273     pull(corp1)
274
275   # Agrupar las categorias poco frecuentes bajo una
          etiqueta comun
276   tabla <- tabla %>%
277     mutate(variable_texto_agrupada = ifelse(corp1 %in%
          categorias_poco_frecuentes, "Otras", corp1))
278
279
280
281
282   # Verificar la distribucion de categorias
283   tabla_frecuencia <- tabla %>%
284     group_by(corp1) %>%
285     summarise(frecuencia = n())
286
287   # Imprimir la tabla de frecuencias
288   print(tabla_frecuencia)
289
290   # Visualizar la distribucion con un grafico de barras
291   #ggplot(tabla, aes(x = corp1)) +
292   # geom_bar() +
293   #labs(title = "Distribucion de la Variable Categorica",
```

```
294  #     x = "Categoria",
295  #    y = "Frecuencia")
296
297  # Identificar categorias poco frecuentes
298  #umbral_frecuencia <- 5  # Puedes ajustar este umbral
        segun tus necesidades
299  #categorias_poco_frecuentes <- tabla_frecuencia %>%
300  # filter(frecuencia < umbral_frecuencia) %>%
301  #pull(corp1)
302
303  # Imprimir categorias poco frecuentes
304  #if (length(categorias_poco_frecuentes) > 0) {
305  # cat("Categorias poco frecuentes:", paste(categorias_
        poco_frecuentes, collapse = ", "), "\n")
306  #} else {
307  # cat("No hay categorias poco frecuentes.\n")
308  #}
309
310
311  #tabla de frecuencias para corp1
312  tabla_frecuencia <- table(tabla$target1)
313  print(tabla_frecuencia)
314
315
316  # Contar los valores perdidos (cadenas vacias)
317  valores_perdidos_contados <- sum(is.na(tabla$target1) |
        tabla$target1 == "")
318
319  # Imprimir el resultado
320  cat("Numero de valores perdidos (cadenas vacias):",
        valores_perdidos_contados, "\n")
321
322
323  # Calcular la frecuencia de cada categoria
324  frecuencia_categorias <- tabla %>%
325    group_by(target1) %>%
326    summarise(frecuencia = n())
327
```

```r
# Definir un umbral de frecuencia para categorias poco frecuentes
umbral_frecuencia <- 10  # Puedes ajustar este umbral segun tus necesidades

# Identificar las categorias poco frecuentes
categorias_poco_frecuentes <- frecuencia_categorias %>%
  filter(frecuencia < umbral_frecuencia) %>%
  pull(target1)

# Agrupar las categorias poco frecuentes bajo una etiqueta comun
tabla <- tabla %>%
  mutate(variable_texto_agrupada = ifelse(target1 %in% categorias_poco_frecuentes, "Otras", target1))

#hacemos una tabla con el conteo de las categorias de natlty1
tabla_frecuencia <- table(tabla$natlty1)
print(tabla_frecuencia)

table(tabla$natlty1, useNA = "ifany")

# Calcular la media de la variable cuantitativa
media_natlty1 <- mean(tabla$natlty1, na.rm = TRUE)

# Imputar los valores perdidos con el techo de la media
tabla$natlty1 <- ifelse(is.na(tabla$natlty1), ceiling(media_natlty1), tabla$natlty1)
```

```r
# Calcular el rango intercuartilico (IQR) para ver si targtype tiene valore
Q1 <- quantile(tabla$natlty1, 0.25)
Q3 <- quantile(tabla$natlty1, 0.75)
IQR <- Q3 - Q1

# Definir limites para identificar valores atipicos
lower_limit <- Q1 - 1.5 * IQR
upper_limit <- Q3 + 1.5 * IQR

# Identificar valores atipicos
outliers <- tabla$natlty1 < lower_limit | tabla$natlty1 > upper_limit

# Mostrar valores atipicos
outlier_values <- tabla$natlty1[outliers]
cat("Valores atipicos en natlty1:", unique(outlier_values), "\n")

# Valores atipicos que deseas eliminar
valores_atipicos <- c(422, 359, 999, 403, 362, 603, 604, 377, 605, 349, 520, 351, 334, 1001, 347, 1003, 1002, 1004)

# Filtrar el dataframe para excluir filas con valores atipicos
tabla <- tabla[!tabla$natlty1 %in% valores_atipicos, ]









# Obtiene los niveles unicos en el orden en que aparecen los datos
unique_levels <- unique(tabla$natlty1_txt)
```

```r

# Convierte a factor con niveles manuales
tabla$natlty1_txt <- factor(tabla$natlty1_txt, levels = unique_levels)

# Verifica que la columna haya sido convertida a factor con los niveles deseados
str(tabla$natlty1_txt)

# imprime
print(tabla$natlty1_txt)


# Contar los valores perdidos (cadenas vacias)
valores_perdidos_contados <- sum(is.na(tabla$natlty1_txt) | tabla$natlty1_txt == "")

# Imprimir el resultado
cat("Numero de valores perdidos (cadenas vacias):", valores_perdidos_contados, "\n")









#eliminamos targtype2 pues es imposible imputar datos

tabla$targtype2 <- NULL



#eliminamos targtype2_txt pues es imposible imputar datos

tabla$targtype2_txt <- NULL

```

```
423
424
425 #eliminamos targsubtype2 pues es imposible imputar datos
426
427 tabla$targsubtype2 <- NULL
428
429 #eliminamos targsubtype2_txt pues es imposible imputar
      datos
430
431 tabla$targsubtype2_txt <- NULL
432
433
434 #eliminamos corp2 pues es imposible imputar datos
435
436 tabla$corp2 <- NULL
437
438 #eliminamos target2 pues es imposible imputar datos
439
440 tabla$target2 <- NULL
441
442 #eliminamos natlty2 pues es imposible imputar datos
443
444 tabla$natlty2 <- NULL
445
446 #eliminamos natlty2_txt pues es imposible imputar datos
447
448 tabla$natlty2_txt <- NULL
449
450
451 #eliminamos targtype3 pues es imposible imputar datos
452
453 tabla$targtype3 <- NULL
454
455
456
457 #eliminamos targtype3_txt pues es imposible imputar
      datos
458
```

```r
tabla$targtype3_txt <- NULL



#eliminamos targsubtype3 pues es imposible imputar datos

tabla$targsubtype3 <- NULL

#eliminamos targsubtype3_txt pues es imposible imputar
   datos

tabla$targsubtype3_txt <- NULL

#eliminamos corp3 pues es imposible imputar datos

tabla$corp3 <- NULL

#eliminamos target3 pues es imposible imputar datos

tabla$target3 <- NULL

#eliminamos natlty3 pues es imposible imputar datos

tabla$natlty3 <- NULL

#eliminamos natlty2_txt pues es imposible imputar datos

tabla$natlty3_txt <- NULL

# Contar los valores perdidos (cadenas vacias)
valores_perdidos_contados <- sum(is.na(tabla$gname) |
   tabla$gname == "")

# Imprimir el resultado
cat("Numero de valores perdidos (cadenas vacias):",
   valores_perdidos_contados, "\n")

```

```r
# Obtiene los niveles unicos en el orden en que aparecen los datos
unique_levels <- unique(tabla$gname)

# Convierte a factor con niveles manuales
tabla$gname <- factor(tabla$gname, levels = unique_levels)

# Verifica que la columna haya sido convertida a factor con los niveles deseados
str(tabla$gname)


#eliminamos gsubname pues es imposible imputar datos

tabla$gsubname <- NULL

#eliminamos gname2 pues es imposible imputar datos

tabla$gname2 <- NULL

#eliminamos gsubname2 pues es imposible imputar datos

tabla$gsubname2 <- NULL

#eliminamos gname3 pues es imposible imputar datos

tabla$gname3 <- NULL

#eliminamos gsubname3 pues es imposible imputar datos

tabla$gsubname3 <- NULL

#podemos guardar los motivos de algunos grupos antes de borrar la columna con
# write.csv(tabla$motive, file = "ruta/del/archivo.csv", row.names = FALSE)

```

```r
#eliminamos motive pues es imposible imputar datos

tabla$motive <- NULL

# checamos si guncertain1 tiene valores perdidos
valores_perdidos <- sum(is.na(tabla$guncertain1))

cat("Numero de valores perdidos en guncertain1:", valores_perdidos, "\n")



# imputacion de guncertain con el techo de la media
media_guncertain1 <- mean(tabla$guncertain1, na.rm = TRUE)
techo_media <- ceiling(media_guncertain1)

# Imputar valores faltantes con el techo de la media
tabla$guncertain1 <- ifelse(is.na(tabla$guncertain1), techo_media, tabla$guncertain1)


#eliminamos guncertain2 pues es imposible imputar datos

tabla$guncertain2 <- NULL


#eliminamos guncertain3 pues es imposible imputar datos

tabla$guncertain3 <- NULL




# Obtiene los niveles unicos en el orden en que aparecen los datos
unique_levels <- unique(tabla$corp1)
```

```
560
561 # Convierte a factor con niveles manuales
562 tabla$corp1 <- factor(tabla$corp1 , levels = unique_
      levels)
563
564 # Verifica que la columna haya sido convertida a factor
      con los niveles deseados
565 str(tabla$corp1)
566
567 # Suponiendo que tu dataframe se llama "tabla"
568 tabla <- tabla %>%
569   mutate(corp1 = ifelse(corp1 == 1, 924, corp1))
570
571 # Suponiendo que tu dataframe se llama "tabla"
572 tabla$corp1 <- ifelse(is.na(tabla$corp1), 936, tabla$
      corp1)
573
574
575 # Obtiene los niveles unicos en el orden en que aparecen
       los datos
576 unique_levels <- unique(tabla$corp1)
577
578 # Convierte a factor con niveles manuales
579 tabla$corp1 <- factor(tabla$corp1 , levels = unique_
      levels)
580
581 # Verifica que la columna haya sido convertida a factor
      con los niveles deseados
582 str(tabla$corp1)
583
584
585
586 # Obtiene los niveles unicos en el orden en que aparecen
       los datos
587 unique_levels <- unique(tabla$targsubtype1_txt)
588
589 # Convierte a factor con niveles manuales
```

```r
tabla$targsubtype1_txt <- factor(tabla$targsubtype1_txt, levels = unique_levels)

# Verifica que la columna haya sido convertida a factor con los niveles deseados
str(tabla$targsubtype1_txt)


# cambiamos las cadenas vacias a la categoria Unnamed Civilian/Unspecified
tabla <- tabla %>%
  mutate(targsubtype1_txt = ifelse(targsubtype1_txt == 13, 45, targsubtype1_txt))
#regresamos a facotor targetsubtype1_txt
# Obtiene los niveles unicos en el orden en que aparecen los datos
unique_levels <- unique(tabla$targsubtype1_txt)

# Convierte a factor con niveles manuales
tabla$targsubtype1_txt <- factor(tabla$targsubtype1_txt, levels = unique_levels)

# Verifica que la columna haya sido convertida a factor con los niveles deseados
str(tabla$targsubtype1_txt)









# Obtiene los niveles unicos en el orden en que aparecen los datos
unique_levels <- unique(tabla$target1)

# Convierte a factor con niveles manuales
tabla$target1 <- factor(tabla$target1 , levels = unique_levels)
```

```
619
620   # Verifica que la columna haya sido convertida a factor
          con los niveles deseados
621   str(tabla$target1)
622
623
624
625   # Suponiendo que tu dataframe se llama "tabla"
626   tabla <- tabla %>%
627     mutate(target1 = ifelse(is.na(target1), "Civilians",
          target1))
628
629
630   # Obtiene los niveles unicos en el orden en que aparecen
          los datos
631   unique_levels <- unique(tabla$target1)
632
633   # Convierte a factor con niveles manuales
634   tabla$target1 <- factor(tabla$target1 , levels = unique_
          levels)
635
636   # Verifica que la columna haya sido convertida a factor
          con los niveles deseados
637   str(tabla$target1)
638
639   #Convierte a factor correccion
640   tabla$targsubtype1_txt <- as.factor(tabla$targsubtype1_
          txt)
641   tabla$corp1 <- as.factor(tabla$corp1)
642   tabla$target1 <- as.factor(tabla$target1)
643
644   return (tabla)
645 }
646
647 data <- manejo2(data)
648
649
650
```

```r


manejo3 <- function(table){
  # Seleccionar atributos relevantes
  selected_data <- data %>% select(individual, nkill,
     nwound, weaptype1_txt,weapsubtype1_txt)

  # Definir limites de los valores
  calculate_bounds <- function(x) {
    Q1 <- quantile(x, 0.25, na.rm = TRUE)
    Q3 <- quantile(x, 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1
    lower_bound <- Q1 - 1.5 * IQR
    upper_bound <- Q3 + 1.5 * IQR
    return(c(lower = lower_bound, upper = upper_bound))
  }

  # Definir funcion para eliminar o limitar valores
     atipicos
  limit_outliers <- function(x) {
    bounds <- calculate_bounds(x)
    x[x < bounds["lower"]] <- bounds["lower"]
    x[x > bounds["upper"]] <- bounds["upper"]
    return(x)
  }

  # Aplicar funcion de limites de valores atipicos
  selected_data$nkill <- limit_outliers(selected_data$
     nkill)
  selected_data$nwound <- limit_outliers(selected_data$
     nwound)
  selected_data$individual <- limit_outliers(selected_data
     $individual)

  # Imputacion de valores perdidos para variables
     numericas
  selected_data <- selected_data %>% mutate(
```

```
683     nkill = ifelse(is.na(nkill), median(nkill, na.rm =
          TRUE), nkill),
684     nwound = ifelse(is.na(nwound), median(nwound, na.rm =
          TRUE), nwound)
685   )
686
687   # Imputacion para variables categoricas
688   impute_mode <- function(x) {
689     ux <- unique(x)
690     ux[which.max(tabulate(match(x, ux)))]
691   }
692
693   selected_data <- selected_data %>% mutate(
694     weaptype1_txt = ifelse(is.na(weaptype1_txt), impute_
          mode(weaptype1_txt), weaptype1_txt),
695     weapsubtype1_txt = ifelse(is.na(weapsubtype1_txt),
          impute_mode(weapsubtype1_txt), weapsubtype1_txt)
696   )
697
698   # Normalizacion de variables numericas (nkill, nwound)
699   numeric_data <- selected_data %>% select(nkill, nwound)
700   preproc <- preProcess(numeric_data, method = c("center",
        "scale"))
701   normalized_data <- predict(preproc, numeric_data)
702
703   # Combinar datos normalizados con datos no numericos
704   selected_data <- bind_cols(selected_data %>% select(-
        nkill, -nwound), normalized_data)
705
706   # Discretizacion de nkill (aplicada despues de la
        normalizacion)
707   selected_data$nkill_discretizado <- cut(selected_data$
        nkill, breaks=c(-Inf, 0, 10, 50, Inf), labels=c("Muy
        bajo", "Bajo", "Medio", "Alto"))
708
709   # Cambio de Character a factor
710   selected_data[sapply(selected_data, is.character)] <-
        lapply(selected_data[sapply(selected_data, is.
```

```
    character)], factor)
711
712   # Guardar los datos procesados
713   data$nkill <- selected_data$nkill
714   data$nwound <- selected_data$nwound
715   data$individual <- selected_data$individual
716   data$weaptype1_txt <- selected_data$weaptype1_txt
717   data$weapsubtype1_txt <- selected_data$weapsubtype1_txt
718
719   not_selected_data <- c(
720     "nperps", "nperpcap", "claimed", "claimmode", "
        claimmode_txt", "claim2", "claimmode2",
721     "claimmode2_txt", "claim3", "claimmode3", "claimmode3_
        txt", "compclaim", "weaptype1",
722     "weapsubtype1", "weaptype2", "weaptype2_txt", "
        weapsubtype2", "weapsubtype2_txt", "weaptype3",
723     "weaptype3_txt", "weapsubtype3", "weapsubtype3_txt", "
        weaptype4", "weaptype4_txt",
724     "weapsubtype4", "weapsubtype4_txt", "weapdetail", "
        nkillus", "nkillter"
725   )
726
727   # Las columnas no utilizadas se vuelven NULL
728   data <- data %>%
729     mutate(across(all_of(not_selected_data), ~NULL))
730
731   return (data)
732 }
733
734 data <- manejo3(data)
735 summary(data)
736
737
738
739 manejo4 <- function(table){
740   atipicosIQR <- function(data_column){
741     data_column
742     iqr <- IQR(data_column, na.rm = TRUE)
```

```
743
744     limite_superior <- quantile(data_column, 0.75, na.rm =
            TRUE) + 1.5 * iqr
745     limite_inferior <- quantile(data_column, 0.25, na.rm =
            TRUE) - 1.5 * iqr
746
747     valores_atipicos <- data_column[data_column > limite_
           superior | data_column < limite_inferior]
748     #valores_atipicos <- unique(valores_atipicos)
749     valores_atipicos <- valores_atipicos[!is.na(valores_
           atipicos)]
750     return (valores_atipicos)
751
752   }
753
754   #Eliminacion de columnas y valores atipicos
        ============================================================

755   #column nwoundus
756   table <- subset(table, select = -c(nwoundus))
757   #column nwoundte
758   table <- subset(table, select = -c(nwoundte))
759   #column propextent
760   table <- subset(table, select = -c(propextent))
761   #column propextent_txt
762   table <- subset(table, select = -c(propextent_txt))
763   #column propvalue
764   table <- table %>%
765     mutate(propvalue = ifelse((property == 0), 0,
           propvalue))%>%
766     mutate(propvalue = ifelse((property == -9), NA,
           propvalue))%>%
767     mutate(propvalue = ifelse((propvalue == -99), NA,
           propvalue))
768   table <- table[!(table$propvalue %in% atipicosIQR(table$
        propvalue)),]
769   #column property
770   table <- subset(table, select = -c(property))
```

```
771  #column propcomment
772  table <- subset(table, select = -c(propcomment))
773  #column nhostkid
774  table <- table %>%
775    mutate(nhostkid = ifelse((ishostkid == 0), 0, nhostkid
       ))%>%
776    mutate(nhostkid = ifelse((ishostkid == -9), NA,
       nhostkid))%>%
777    mutate(nhostkid = ifelse((nhostkid == -99), NA,
       nhostkid))
778  data_aux2 <- table[!(table$nhostkid %in% atipicosIQR(
     table$nhostkid)),]
779  #column ishostkid
780  table <- subset(table, select = -c(ishostkid))
781  #column nhostkidus
782  table <- subset(table, select = -c(nhostkidus))
783  #column nhours
784  table <- table %>%
785    mutate(ndays = ifelse((ndays == -99), NA, ndays))%>%
786    mutate(ndays = ifelse((ndays == -9), NA, ndays))
787  table <- table %>%
788    mutate(nhours = ifelse((nhours == -99), NA, nhours))
       %>%
789    mutate(nhours = ifelse((nhours == -9), NA, nhours))%>%
790    mutate(nhours = ifelse((is.na(nhours)),  (24*ndays),
       ifelse(is.na(ndays), nhours, (nhours+(24*ndays)))))
791  table <- table[!(table$nhours %in% atipicosIQR(table$
     nhours)),]
792  #column ndays
793  table <- subset(table, select = -c(ndays))
794  #column divert
795  table$divert <- as.factor(table$divert)
796  #column kidhijcountry
797  table$kidhijcountry <- as.factor(table$kidhijcountry)
798  #column ransomamt
799  table <- table %>%
800    mutate(ransomamt = ifelse((ransom == 0), 0, ransomamt)
       )%>%
```

```
801     mutate(ransomamt = ifelse((ransom == -9), NA,
          ransomamt))%>%
802     mutate(ransomamt = ifelse((ransomamt == -99), NA,
          ransomamt))
803   #column ransomamtus
804   table <- subset(table, select = -c(ransomamtus))
805   #column ransompaid
806   table <- table %>%
807     mutate(ransompaid = ifelse((ransom == 0), 0,
          ransompaid))%>%
808     mutate(ransompaid = ifelse((ransom == -9), NA,
          ransompaid))%>%
809     mutate(ransompaid = ifelse((ransom == -99), NA,
          ransompaid))
810
811   table <- table[!(table$ransompaid %in% atipicosIQR(table
        $ransompaid)),]
812   #column ransompaidus
813   table <- subset(table, select = -c(ransompaidus))
814   #column ransomnote
815   table <- subset(table, select = -c(ransomnote))
816   #column ransom
817   table <- subset(table, select = -c(ransom))
818   #column hostkidoutcome
819   table <- subset(table, select = -c(hostkidoutcome))
820   #column hostkidoutcome_txt
821   table$hostkidoutcome_txt <- as.factor(table$
        hostkidoutcome_txt)
822   #column nreleased
823   table <- table[!(table$nreleased %in% atipicosIQR(table$
        nreleased)),]
824   #column addnotes
825   table <- subset(table, select = -c(addnotes))
826   #column scite1
827   table <- subset(table, select = -c(scite1))
828   #column scite2
829   table <- subset(table, select = -c(scite2))
830   #column scite3
```

```
831   table <- subset(table, select = -c(scite3))
832   #column dbsource
833   table <- subset(table, select = -c(dbsource))
834   #column INT_LOG
835   table <- table[!(table$INT_LOG %in% atipicosIQR(table$
          INT_LOG)),]
836   #column INT_IDEO
837   table <- table[!(table$INT_IDEO %in% atipicosIQR(table$
          INT_IDEO)),]
838   #column INT_MISC
839   table <- table[!(table$INT_MISC %in% atipicosIQR(table$
          INT_MISC)),]
840   #column INT_ANY
841   table <- table[!(table$INT_ANY %in% atipicosIQR(table$
          INT_ANY)),]
842   #column related
843   table <- subset(table, select = -c(related))
844
845   #Imputacion de columnas
          ================================================================
846
847   media_aux <- mean(table$propvalue, na.rm = TRUE)
848   table$propvalue <- ifelse(is.na(table$propvalue), media_
          aux, table$propvalue)
849
850   media_aux <- mean(table$nhostkid, na.rm = TRUE)
851   table$nhostkid <- ifelse(is.na(table$nhostkid), media_
          aux, table$nhostkid)
852
853   media_aux <- mean(table$nhours, na.rm = TRUE)
854   table$nhours <- ifelse(is.na(table$nhours), media_aux,
          table$nhours)
855
856   media_aux <- mean(table$ransomamt, na.rm = TRUE)
857   table$ransomamt <- ifelse(is.na(table$ransomamt), media_
          aux, table$ransomamt)
858
```

```
859   media_aux <- mean(table$ransompaid, na.rm = TRUE)
860   table$ransompaid <- ifelse(is.na(table$ransompaid),
          media_aux, table$ransompaid)
861
862   media_aux <- mean(table$nreleased, na.rm = TRUE)
863   table$nreleased <- ifelse(is.na(table$nreleased), media_
          aux, table$nreleased)
864
865   media_aux <- mean(table$INT_LOG, na.rm = TRUE)
866   table$INT_LOG <- ifelse(is.na(table$INT_LOG), media_aux,
          table$INT_LOG)
867
868   media_aux <- mean(table$INT_IDEO, na.rm = TRUE)
869   table$INT_IDEO <- ifelse(is.na(table$INT_IDEO), media_
          aux, table$INT_IDEO)
870
871   media_aux <- mean(table$INT_MISC, na.rm = TRUE)
872   table$INT_MISC <- ifelse(is.na(table$INT_MISC), media_
          aux, table$INT_MISC)
873
874   media_aux <- mean(table$INT_ANY, na.rm = TRUE)
875   table$INT_ANY <- ifelse(is.na(table$INT_ANY), media_aux,
          table$INT_ANY)
876   #Discretizacion de columnas
877
878   #Normalizacion de columnas
879   table$ransomamt <- scale(table$ransomamt)
880
881   return (table)
882 }
883
884
885
886 data <- manejo4(data)
887
888 summary(data)
889 write.csv(data, "E:/Descargas/Repos/AMD-2024-1/
       datosPrepTerrorismo.csv", row.names=TRUE)
```

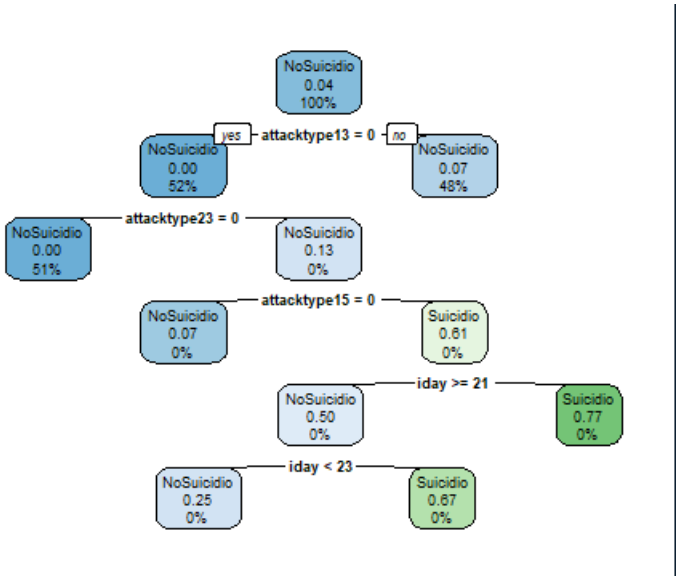
2.2. Conclusiones etapa de preprocesamiento:

Para esta etapa conlcuimos que muchos de los atributos de nuestro conjunto de datos original no nos sería útil para nuestro cometido, otros atributos se veían prometedores, pero al tener un gran porcentaje de valores atípicos u otros parámetros que no aseguraban que nos ayudarían a tener una predicción más concisa y confiable, fueron ignorados y removidos.

- Varible Obejtivo: suicide
  Variables Predictoras: iday, attacktype1, attacktype2

```r
library(caret)
library(rpart.plot)

original <- read.csv("../code/R/proyecto/src/datosPrepTerrorismo.csv")
summary(original)

dataset <- original[,c("iday","attacktype1","attacktype2","suicide")]
dataset$suicide <- as.factor(ifelse(dataset$suicide==0,"Died","Survived"))
dataset$attacktype1 <- as.factor(dataset$attacktype1)
dataset$attacktype2 <- as.factor(dataset$attacktype2)
str(dataset)
summary(dataset)

dataset <- na.omit(dataset)
set.seed(9999)

train <- createDataPartition(dataset[,"suicide"],p=0.8,list=FALSE)
dataset.trn <- dataset[train,]
dataset.tst <- dataset[-train,]

ctrl  <- trainControl(method  = "cv",number  = 10)
fit.cv <- train(suicide ~ ., data = dataset.trn, method = "rpart",
                trControl = ctrl,
                # preProcess = c("center","scale"),
                # tuneGrid =data.frame(cp=0.05))
                tuneLength = 30) # metric="Kappa",

pred <- predict(fit.cv,dataset.tst)
confusionMatrix(table(dataset.tst[,"suicide"],pred))
print(fit.cv)
plot(fit.cv)

rpart.plot(fit.cv$finalModel,fallen.leaves = FALSE)
```

```
> pred <- predict(fit.cv,dataset.tst)
> confusionMatrix(table(dataset.tst[,"suicide"],pred))
Confusion Matrix and Statistics

          pred
            Died Survived
  Died     31981        7
  Survived  1259        7

               Accuracy : 0.9619
                 95% CI : (0.9598, 0.964)
    No Information Rate : 0.9996
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0101

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.962124
            Specificity : 0.500000
         Pos Pred Value : 0.999781
         Neg Pred Value : 0.005529
             Prevalence : 0.999579
         Detection Rate : 0.961719
   Detection Prevalence : 0.961929
      Balanced Accuracy : 0.731062

       'Positive' Class : Died

> print(fit.cv)
```



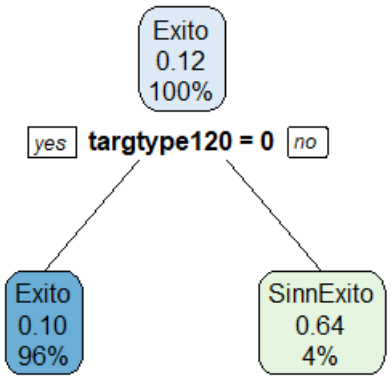
Nuestro árbol se divide en primeramente en attacktype1 dependiendo de si el valor de la fila en dicha columna es 3. nuestra clase positiva es suicidio efectivo y el 52 % de los eventos que fueron con intención de suicidio tuvieron un ataque primario de tipo 3.

Verificamos los porcentajes para filas con attacktype2=3 y del 52 % anterior, solo el 1 % fueron los eventos con intención de suicidio diferente a 3.
El 1 % restante se divide en el resto de hojas de nuestro CART.

- Varible Obejtivo: success
  Variables Predictoras: targtype1, attacktype1

```r
library(caret)
library(rpart.plot)

original <- read.csv("../code/R/proyecto/src/datosPrepTerrorismo.csv")
summary(original)

dataset <- original[,c("targtype1","targsubtype1","success")]
dataset$success <- as.factor(ifelse(dataset$success==0,"SinExito","Exito"))
dataset$targtype1 <- as.factor(dataset$targtype1)
dataset$targsubtype1 <- as.factor(dataset$targsubtype1)
str(dataset)
summary(dataset)

dataset <- na.omit(dataset)
set.seed(9999)

train <- createDataPartition(dataset[,"success"],p=0.8,list=FALSE)
dataset.trn <- dataset[train,]
dataset.tst <- dataset[-train,]

ctrl  <- trainControl(method  = "cv",number  = 10)
fit.cv <- train(success ~ ., data = dataset.trn, method = "rpart",
                trControl = ctrl,
                # preProcess = c("center","scale"),
                # tuneGrid =data.frame(cp=0.05))
                tuneLength = 30) # metric="Kappa",

pred <- predict(fit.cv,dataset.tst)
confusionMatrix(table(dataset.tst[,"success"],pred))
print(fit.cv)
plot(fit.cv)

rpart.plot(fit.cv$finalModel,fallen.leaves = FALSE)
```

```
> pred <- predict(fit.cv,dataset.tst)
> confusionMatrix(table(dataset.tst[,"success"],pred))
Confusion Matrix and Statistics

           pred
            Exito SinnExito
  Exito     28894       416
  SinnExito  3156       788

               Accuracy : 0.8926
                 95% CI : (0.8892, 0.8959)
    No Information Rate : 0.9638
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2654

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9015
            Specificity : 0.6545
         Pos Pred Value : 0.9858
         Neg Pred Value : 0.1998
             Prevalence : 0.9638
         Detection Rate : 0.8689
   Detection Prevalence : 0.8814
      Balanced Accuracy : 0.7780

       'Positive' Class : Exito

> print(fit.cv)
CART

133021 samples
```

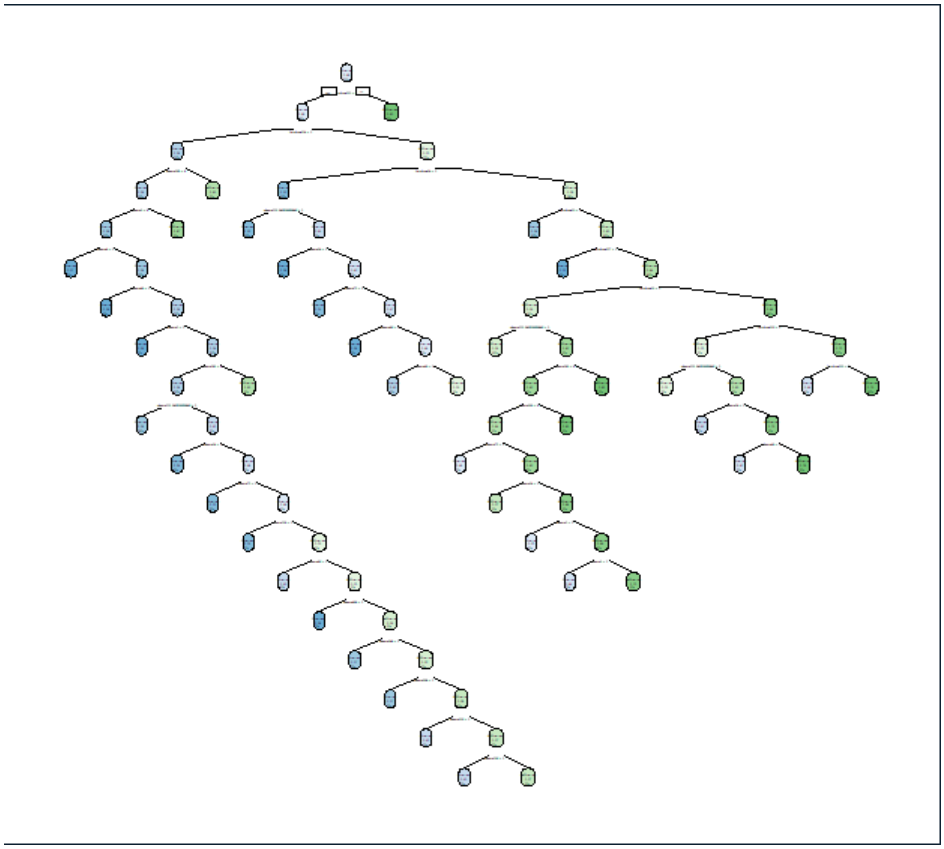


Nuestro árbol solo cuenta con 3 nodos, en el nos interesa saber si el evento fue o no existos y divide nuestro conjunto dependiendo de si el tipo de objetivo es igual a 20 o no, interpretamos fácilmente que la gran mayoría de eventos exitosos (96 %) sucedieron cuando el objetivo fue de tipo 20.

- Varible Obejtivo: nkill
  Variables Predictoras: targtype1, nhours

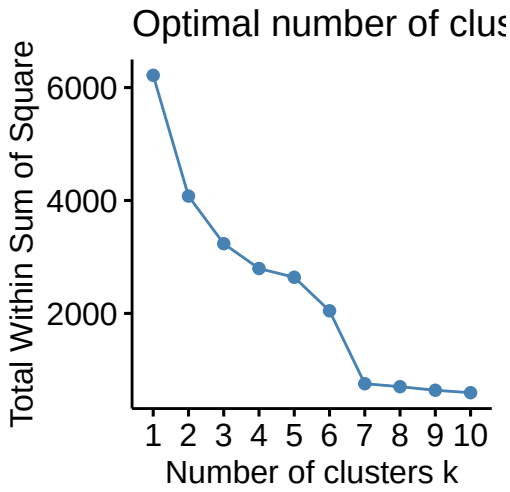
```r
library(dplyr)
library(caret)
library(rpart.plot)

original <- read.csv("../code/R/proyecto/src/datosPrepTerrorismo.csv")
summary(original)

dataset <- mutate(data, nkill = ifelse((nkill>0),1,0))

dataset <- original[,c("nhours","targtype1","nkill")]
dataset$nkill <- as.factor(ifelse(dataset$nkill==0,"NoMuertes","Muertes"))
dataset$targtype1 <- as.factor(dataset$targtype1)
dataset$nhours <- as.factor(dataset$nhours)
str(dataset)
summary(dataset)

dataset <- na.omit(dataset)
set.seed(9999)

train <- createDataPartition(dataset[,"nkill"],p=0.8,list=FALSE)
dataset.trn <- dataset[train,]
dataset.tst <- dataset[-train,]

ctrl  <- trainControl(method  = "cv",number  = 10)
fit.cv <- train(nkill ~ ., data = dataset.trn, method = "rpart",
                trControl = ctrl,
                # preProcess = c("center","scale"),
                tuneLength = 30) # metric="Kappa",

pred <- predict(fit.cv,dataset.tst)
confusionMatrix(table(dataset.tst[,"nkill"],pred))
print(fit.cv)
plot(fit.cv)
rpart.plot(fit.cv$finalModel,fallen.leaves = FALSE)
```

```
 using the largest value.
The final value used for the model was cp = 0.
> plot(fit.cv)
>
> rpart.plot(fit.cv$finalModel,fallen.leaves = FALSE)
> confusionMatrix(table(dataset.tst[,"nkill"],pred))
Confusion Matrix and Statistics

           pred
            Muertes NoMuertes
  Muertes     11324      4635
  NoMuertes    6725      8865

               Accuracy : 0.6399
                 95% CI : (0.6346, 0.6452)
    No Information Rate : 0.5721
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2786

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6274
            Specificity : 0.6567
         Pos Pred Value : 0.7096
         Neg Pred Value : 0.5686
             Prevalence : 0.5721
         Detection Rate : 0.3589
   Detection Prevalence : 0.5058
      Balanced Accuracy : 0.6420

       'Positive' Class : Muertes

>
```

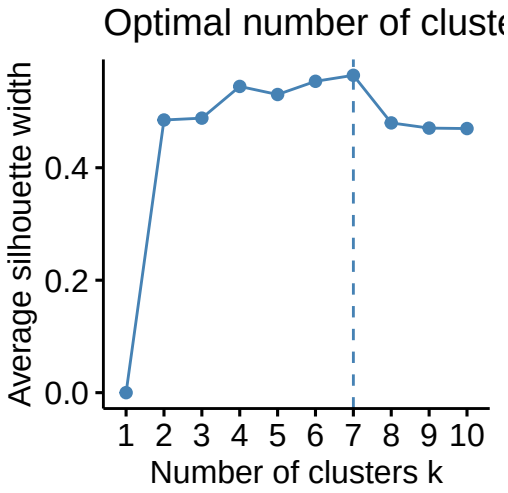


Agrupacion Datos Global Terrrorism Database

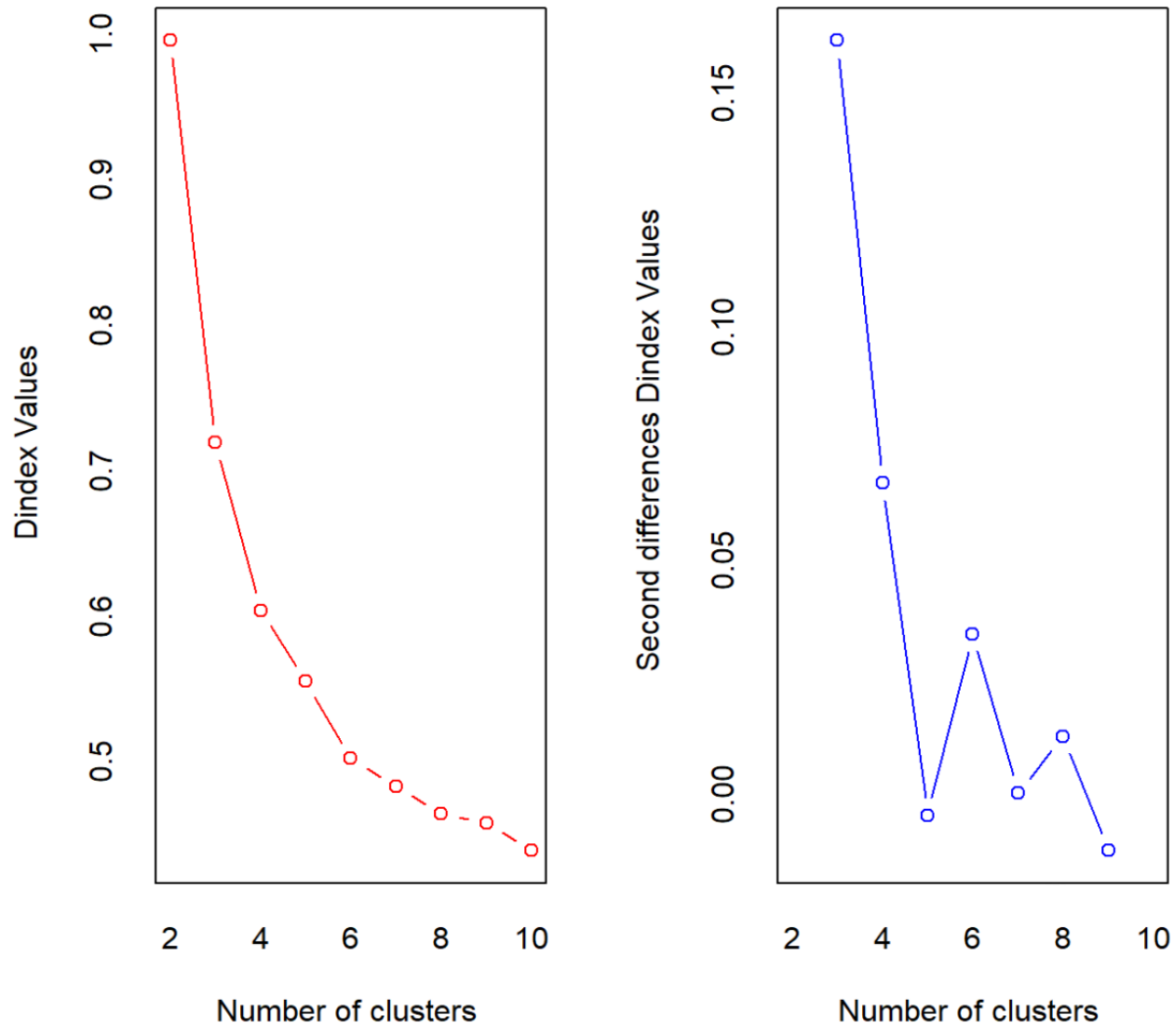
Primero comenzamos por decidir que objetivo buscaran nuestros datos, el cual es encontrar patrones relevantes con los datos de la región, bajas sufridas por los ataques, y los años en que ocurrieron los atentados, de aquí usamos el método Partitioning Around Medoids (pam) y graficamos con dos métodos, wss(With-in-Sum-of-Squares) y silhouette para darnos una idea de cuantos grupos usaremos en el algoritmo k-means, sin embargo, aqui se nos presentaron problemas pues nuestro vector C(ïyear",region", "nkill") tenia una cantidad absurda de información, por lo que con nuestro limitado poder de computo tomamos la decisión de hacer muestreo, sin embargo nuestra muestra no es lo suficientemente grande pues se nos presentaba el mismo error de no poder procesar tanta información, por lo que es muy posible que nuestros datos presenten sesgos, sin embargo estos son los resultados obtenidos:



Del algoritmo pam con la visualización wss podemos observar que deberíamos agrupar en 4 o 5 clusters, ahora bien, veamos con la visualización silhouette



De esta visualización podemos observar que podríamos tener el numero óptimo de clusters con un valor de 7, sin embargo haremos uso de k-means con un numero mínimo de clusters igual a 2 y como maximo 10, para que el mismo algoritmo arroje cual es aglomeramiento óptimo

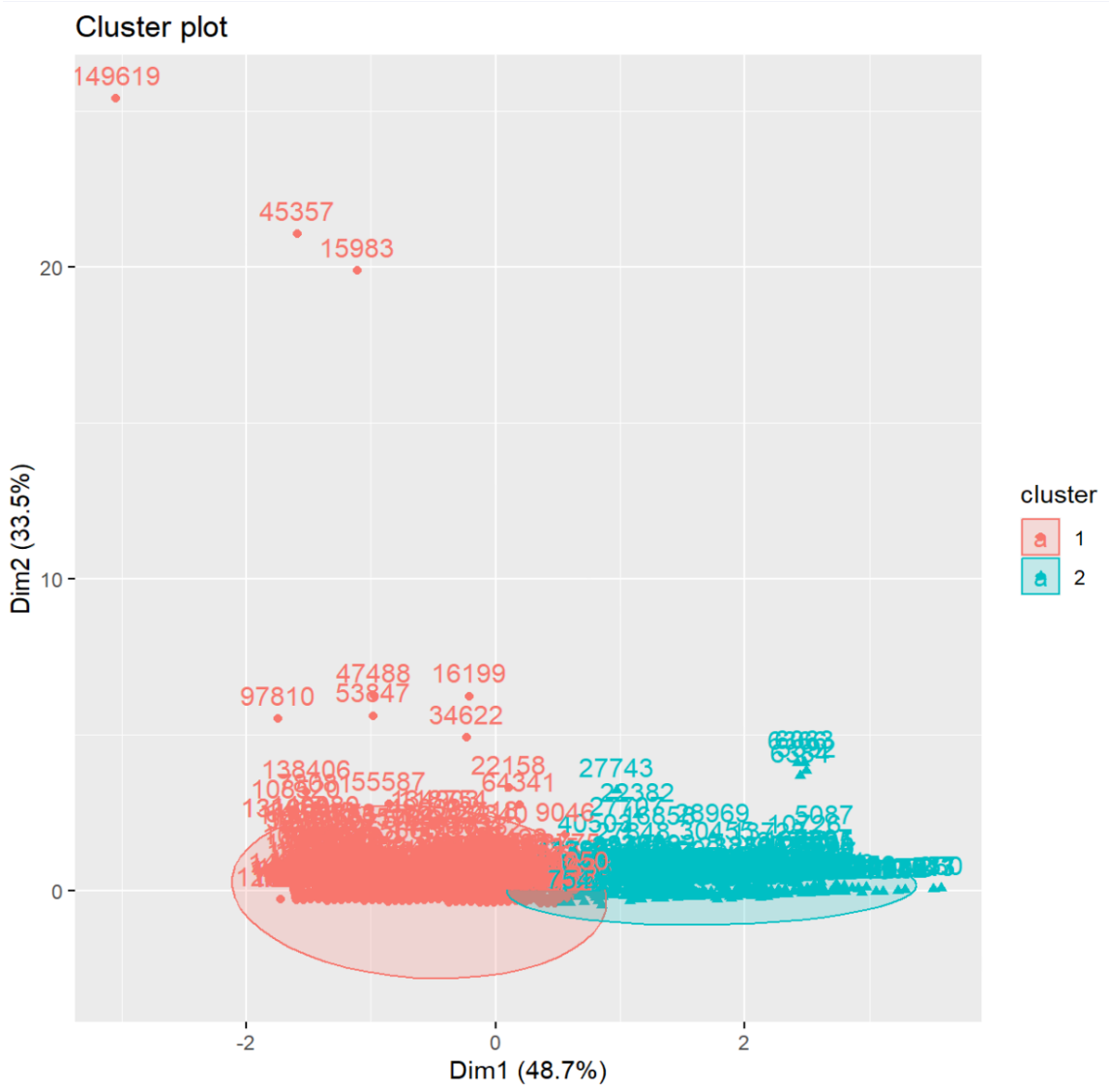


Ademas de obtener el grafo obtenos la siguiente salida de la funcion:

*** : The Hubert index is a graphical method of

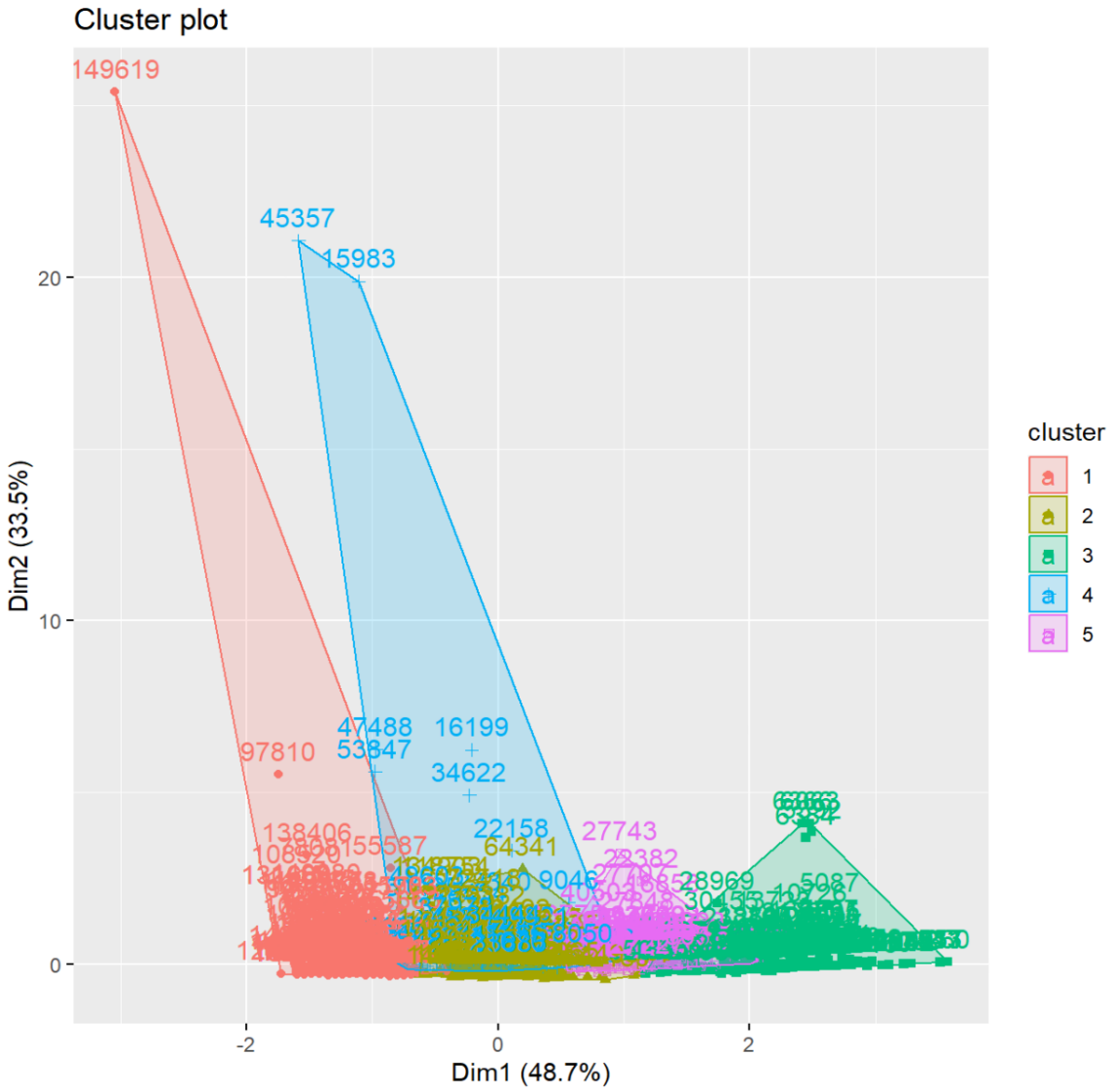
```
     determining the number of clusters.
                    In the plot of Hubert index, we seek
                    a significant knee that corresponds to a
                    significant increase of the value of the
                    measure i.e the significant peak in Hubert
                    index second differences plot.

*** : The D index is a graphical method of determining the
number of clusters.
                    In the plot of D index, we seek a
                    significant knee (the significant peak
                    in Dindex second differences plot) that
                    corresponds
                    to a significant increase of the value of
                    the measure.

*******************************************
* Among all indices:
* 5 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 13 proposed 5 as the best number of clusters
* 2 proposed 6 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                    ***** Conclusion *****

* According to the majority rule, the best number of
clusters is  5


*******************************************
```

Como vimos el numero de grupos oscilaba entre los 4 a los 7, sin embargo con este ultimo grafo tomaremos un aglomeramiento con 5 grupos y uno con 2 para ver si es que efectivamente los datos se dividen de una manera significativa o si se dificulta separar los grupos



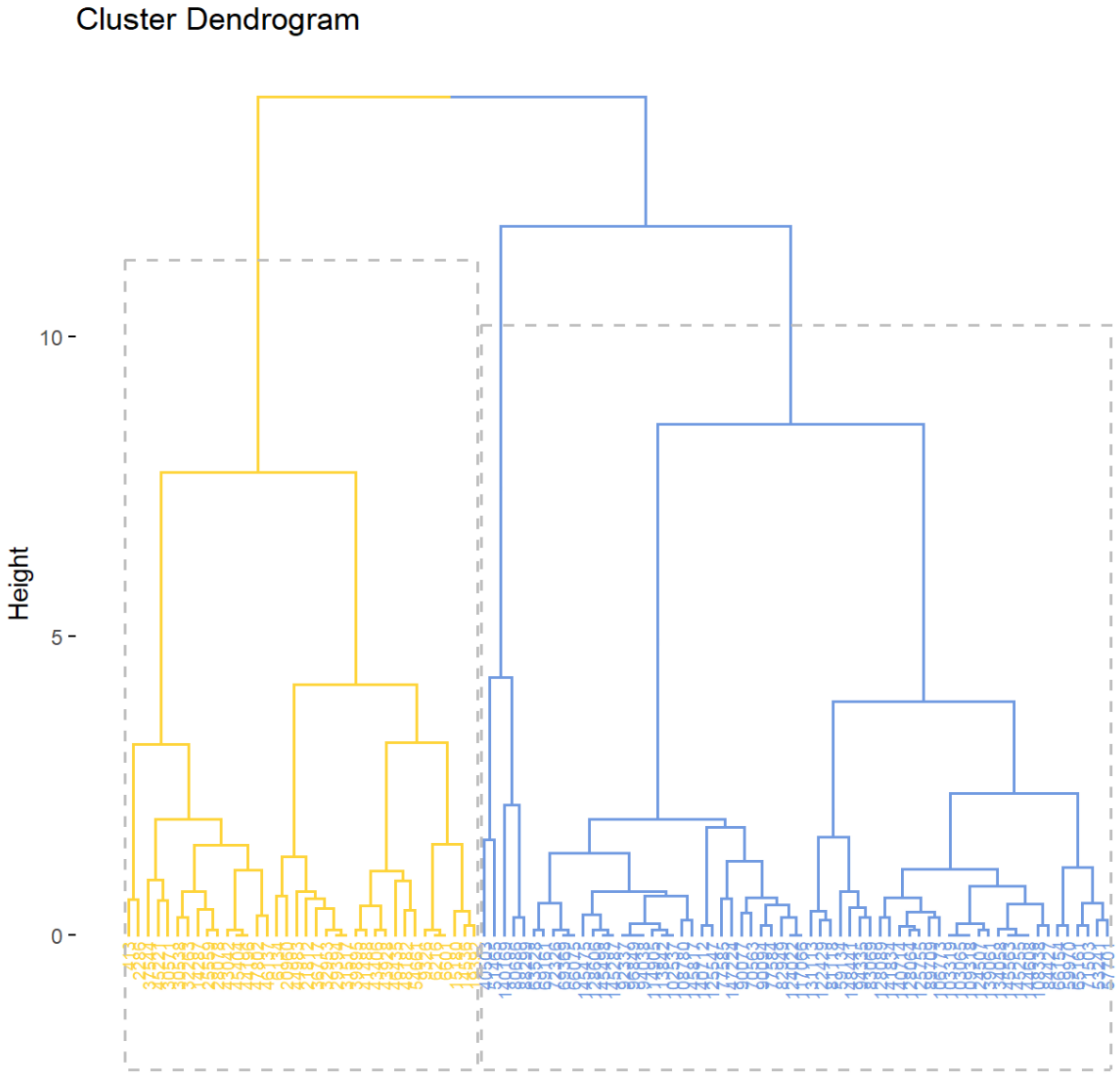
Para 2 clusters es fácil para el algoritmo k-means separar los extremos del eje x, sin embargo conforme nos acercamos vemos que los elementos

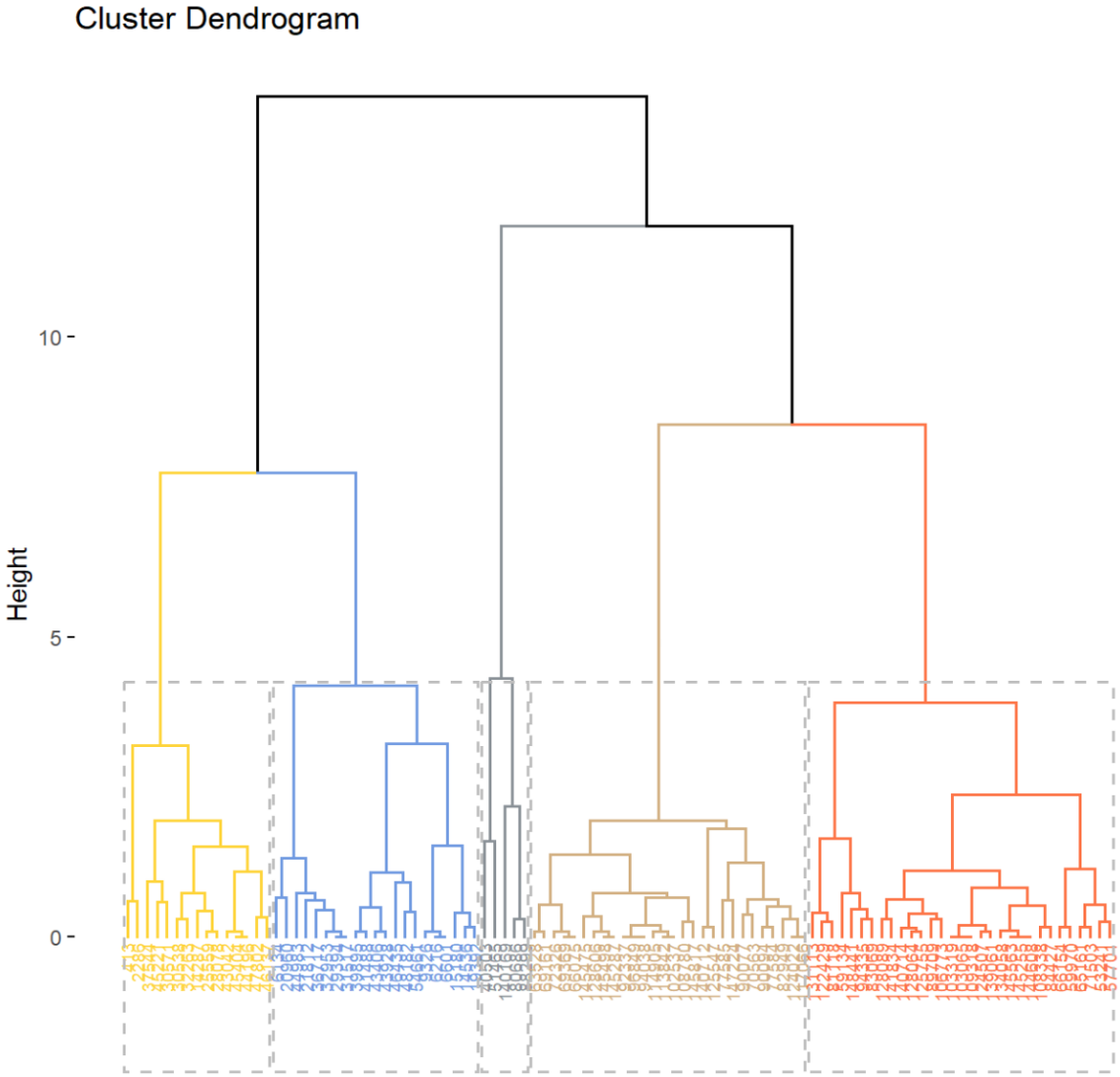
no parecen estar del todo bien divididos, no parecen ser disjuntos entre si



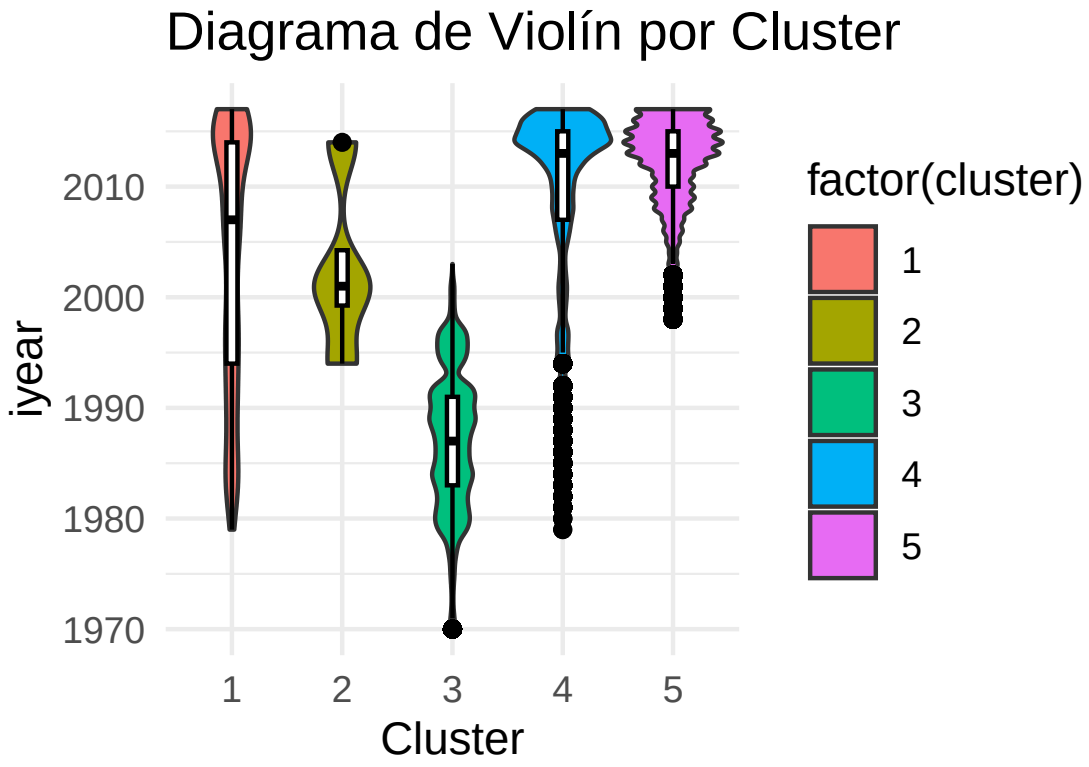
Para nuestro óptimo de cluster vemos que el cluster 1 con el cluster 3 parece no presnetar traslapamiento, sin embargo 2,3,4 siguen pareciendo parte de un grupo.

Por otro lado, la muestra pequeña parece presentarnos muy pocas filas con valores muy altos en el numero de bajas, lo que provoca que veamos esos picos en el grafo.
De igual manera estos datos atipicos provocan que al hacer un dendogrma de la miestra se nos genere un grupo con muy pocos valores dentro del cluster Dendogramas

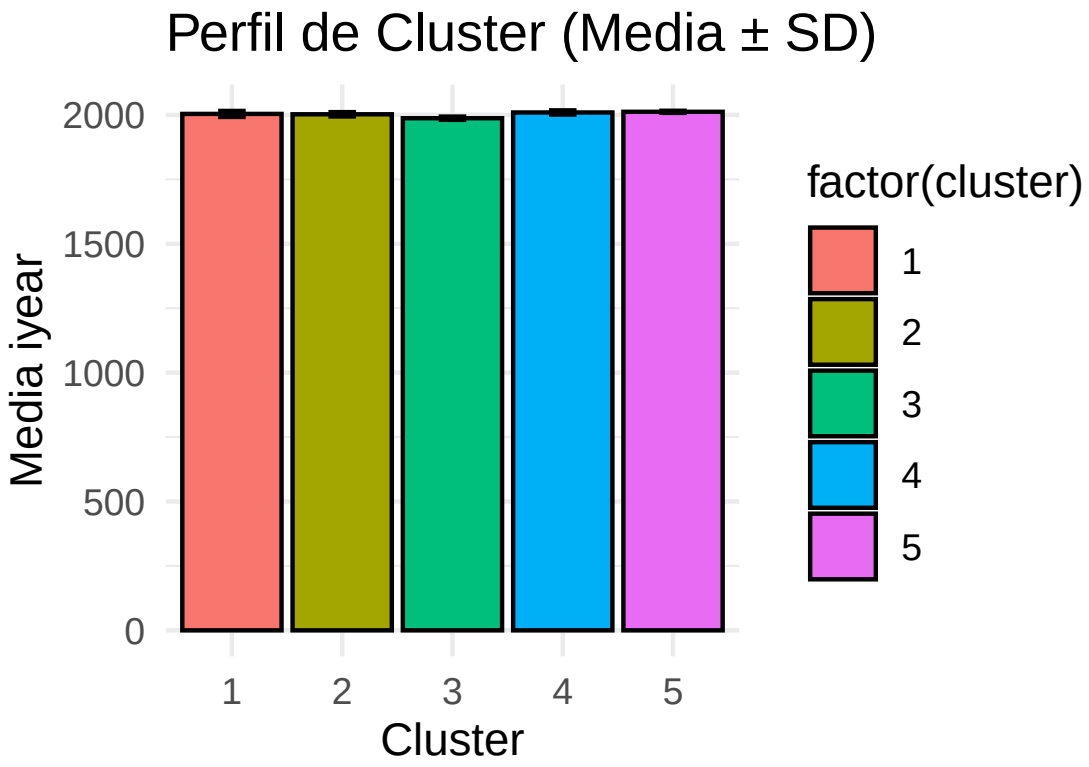



Cluster Dendrogram
Height
10
5
0

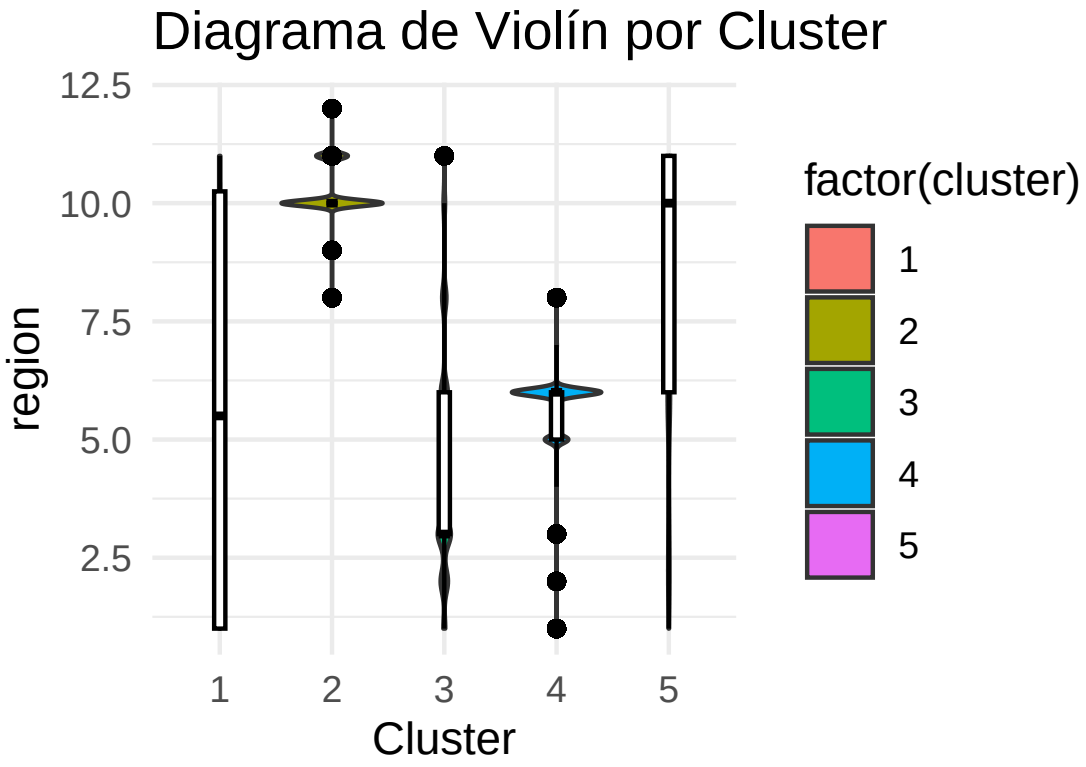
También realizamos graficas de violin que nos ayudan a identificar los cuatriles si es que aplica el caso y la distribución de los datos, vemos que para el año se hace una buena distribución que nos aporta información como que los clusters 2,3,4 y 5 presentan outliers



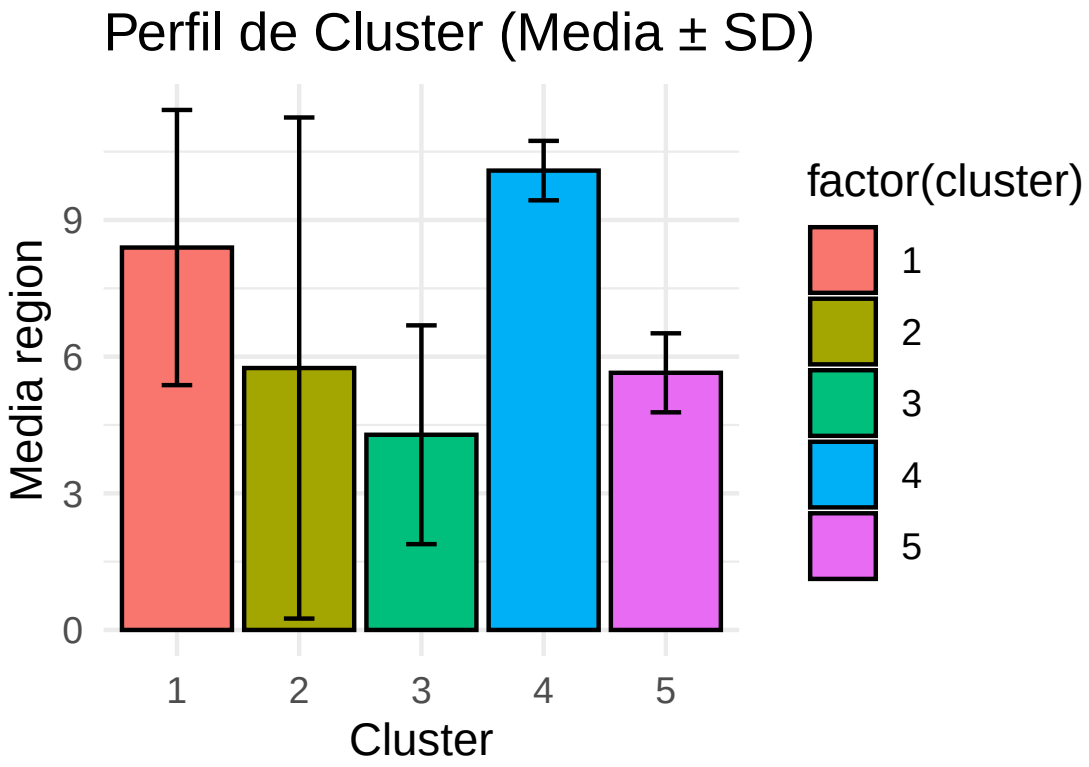
Con el perfil de Cluster vemos de igual manera la distribución de la muestra en los 5 clusters, y en este caso por ser los años en que se presentaron los atentados todos deben de tener valores con los años 2000 hacia arriba



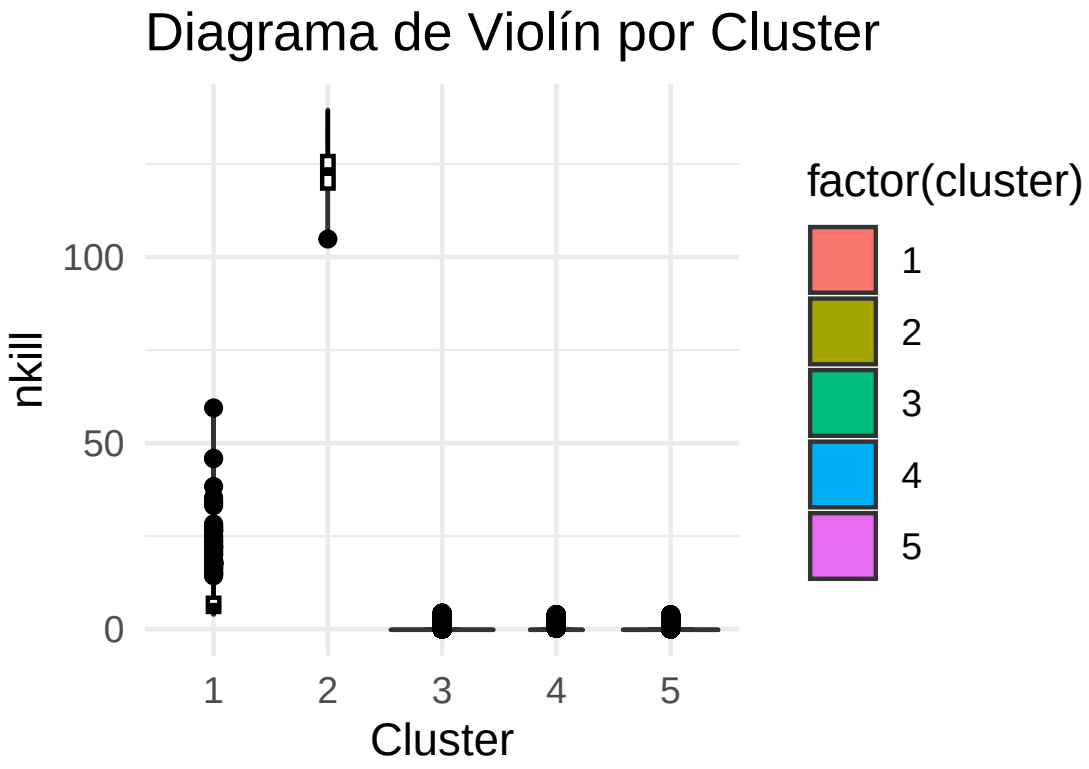
En la gráfica de violín de la región vemos que la distribución de las regiones en los 5 clusters predomina en 2 de ellos mientras que los otros parecieran tener muy pocas observaciones



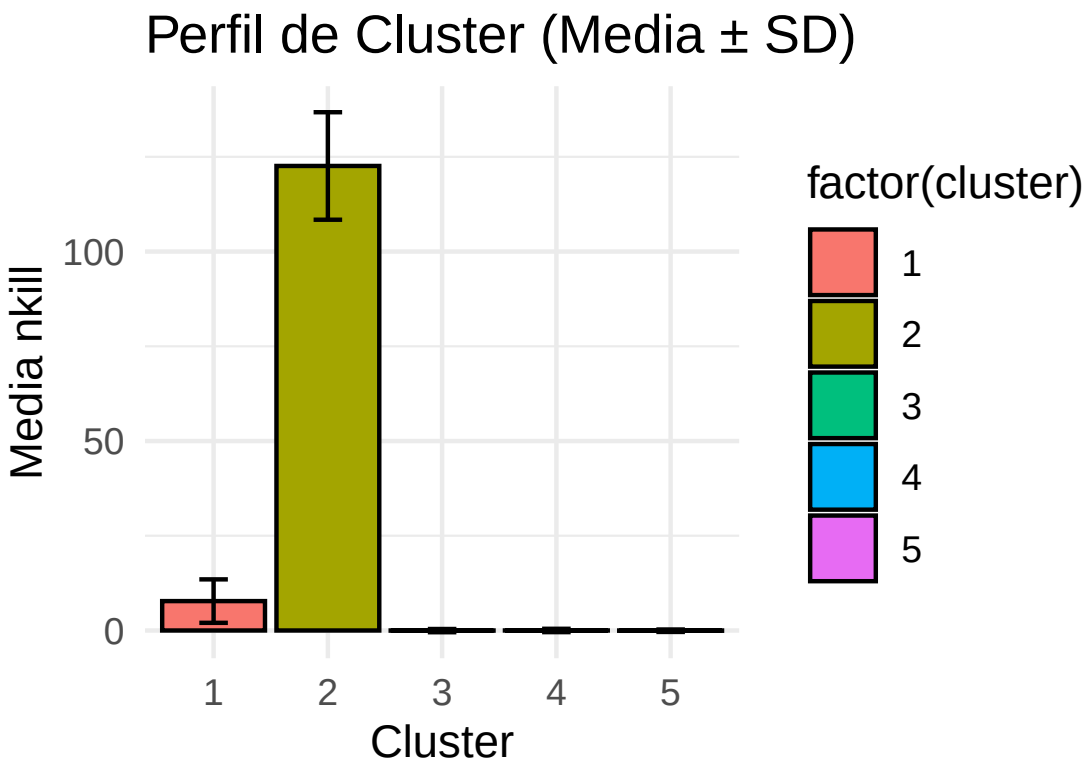
A diferencia de iyear en región tenemos solo 12 regiones y vemos una distribución sesgada a la derecha (cluster 4)



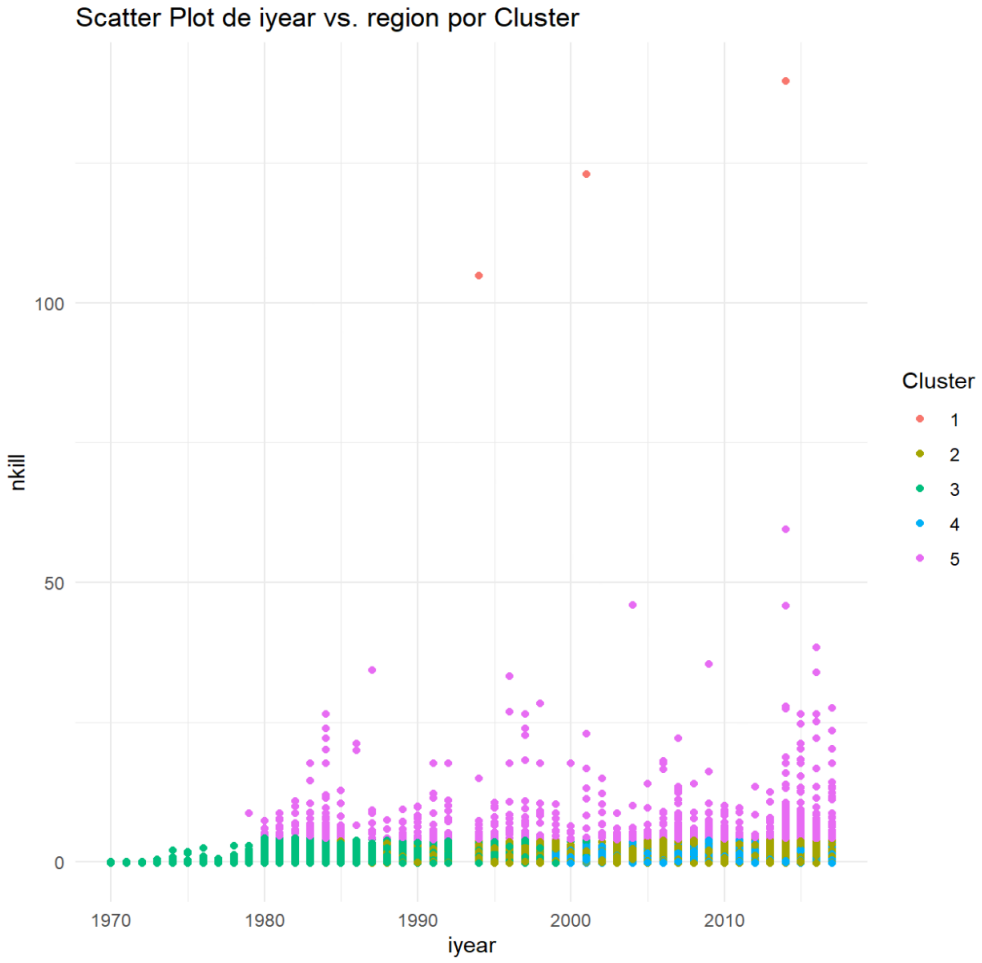
Por la naturaleza tan volatil de nkill es predecible que la dispersión de los datos seria grande pues que exista el mismo numero de bajas en muchos ataques es improbable por lo que todo en la grafica de violin son outliers



vemos una concentración total en el cluster 2



Por último hacemos una gráfica para ver el número de bajas a lo largo de los años y lo que podemos ver es que a pesar de contar con tantos datos de los ataques que se han presentado no se ve que haya ayudado mucho el conocimiento de los mismos, pues si bien, la población en todas las regiones a aumentado con el paso del tiempo, el número de bajas en los ataques mas revientes es en promedio mas grande en comparación con los ataques que se presentaron primero.



El análisis combinado de Árbol CART y Redes Neuronales revela patrones significativos en los datos de eventos terroristas. podemos destacar la importancia crítica del tipo de objetivo, especialmente cuando es igual a 20, en la determinación del éxito del evento, con un notorio 96 % de éxito en estos casos. Este hallazgo sugiere que este atributo específico es altamente predictivo y puede ser clave para comprender y predecir eventos exitosos. También que el ataque primario de tipo 3 no está altamante relacionado con la intención de suicidio. Este conocimiento puede ser crucial para comprender las motivaciones detrás de ciertos eventos terroristas.

Para finalizar nos gustaría decir que a pesar del trabajo de minado de datos no hemos sido capaces de reducir el número de muertes, como notamos en la sección de agrupación: .ª pesar de contar con tantos datos de los ataques que se han presentado no se ve que haya ayudado mucho el conocimiento de los mismos, pues si bien, la población en todas las regiones a aumentado con el paso del tiempo, el número de bajas en los ataques mas revientes es en promedio mas grande en comparación con los ataques que se presentaron primero.