Tipología y ciclo de vida de los datos

Practica 2: Limpieza y validación de los datos

Jorge Gonzalez
Estudiante Master Data Science UOC

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Contexto:
La Global Terrorism Database (GTD) es una base de datos de código abierto que incluye información sobre ataques terroristas en todo el mundo desde 1970 hasta 2016. El GTD incluye datos sistemáticos sobre incidentes que han ocurrido durante este período de tiempo y ahora incluye más de 170,000 casos. La base de datos es mantenida por investigadores del Consorcio Nacional para el Estudio del Terrorismo y Respuestas al Terrorismo (START), con sede en la Universidad de Maryland.

Contenido:
- Geografía: en todo el mundo
- Período de tiempo: 1970-2016, excepto 1993
- Unidad de análisis: ataque
- Variables:> 100 variables en el lugar, tácticas, perpetradores, objetivos y resultados
- Fuentes: artículos de medios no clasificados

Que preguntas pretende responder:
El conjunto de datos pretende responder para cada incidente, la información sobre la fecha y la ubicación, las armas utilizadas y la naturaleza del objetivo, el número de víctimas y, cuando sea identificable, el grupo o la persona responsable.

Las preguntas que se plantean como parte del problema que se quiere resolver son las siguientes:

¿Cuál es el número de ataques terroristas realizados por año?

¿Cuáles son los países más afectados por el terrorismo?

¿Cuáles son las regiones más afectados por el terrorismo?

¿Qué tipos de ataques han causado más muertes?

¿Cuáles son los grupos terroristas que perpetran mayor número de ataques terroristas?

¿Quiénes son los principales objetivos de los ataques terroristas?

2. Limpieza de los datos.

2.1 Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

Se explora el dataset:

str(Terrorismo) #Proporciona la estructura del conjunto de datos

```
'data.frame':   170350 obs. of  135 variables:
 $ eventid           : num  1.97e+11 1.97e+11 1.97e+11 1.97e+11 1.97e+11 ...
 $ iyear             : int  1970 1970 1970 1970 1970 1970 1970 1970 1970 1970 ...
 $ imonth            : int  7 0 1 1 1 1 1 1 1 1 ...
 $ iday              : int  2 0 0 0 0 1 2 2 2 3 ...
 $ approxdate        : Factor w/ 1834 levels "","01/04/2000",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ extended          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ resolution        : Factor w/ 1860 levels "","1/1/1978",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ country           : int  58 130 160 78 101 217 218 217 217 217 ...
 $ country_txt       : Factor w/ 205 levels "Afghanistan",..: 48 117 141 69 91 190 191 190 190
 $ region            : int  2 1 5 8 4 1 3 1 1 1 ...
 $ region_txt        : Factor w/ 12 levels "Australasia & Oceania",..: 2 7 10 12 4 7 8 7 7 7 .
 $ provstate         : Factor w/ 2495 levels "","(Region) of Republican Subordination (Provinc
 $ city              : Factor w/ 33958 levels "","15 September Dam",..: 27696 20531 32069 2312
 $ latitude          : num  18.5 19.4 15.5 38 33.6 ...
 $ longitude         : num  -70 -99.1 120.6 23.7 130.4 ...
 $ specificity       : int  1 1 4 1 1 1 1 1 1 1 ...
 $ vicinity          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ location          : Factor w/ 39817 levels "","\"Colony 39\" settlement in uncleared areas
 $ summary           : Factor w/ 101539 levels "","00/00/2012: Sometime between July 29 and Au
 $ crit1             : int  1 1 1 1 1 1 1 1 1 1 ...
 $ crit2             : int  1 1 1 1 1 1 1 1 1 1 ...
 $ crit3             : int  1 1 1 1 1 1 1 1 1 1 ...
```

names(Terrorismo) #Lista variables en el conjunto de datos

```
  [1] "eventid"           "iyear"             "imonth"            "iday"              "approxdate"
  [6] "extended"          "resolution"        "country"           "country_txt"       "region"
 [11] "region_txt"        "provstate"         "city"              "latitude"          "longitude"
 [16] "specificity"       "vicinity"          "location"          "summary"           "crit1"
 [21] "crit2"             "crit3"             "doubtterr"         "alternative"       "alternative_txt"
 [26] "multiple"          "success"           "suicide"           "attacktype1"       "attacktype1_txt"
 [31] "attacktype2"       "attacktype2_txt"   "attacktype3"       "attacktype3_txt"   "targtype1"
 [36] "targtype1_txt"     "targsubtype1"      "targsubtype1_txt"  "corp1"             "target1"
 [41] "natlty1"           "natlty1_txt"       "targtype2"         "targtype2_txt"     "targsubtype2"
 [46] "targsubtype2_txt"  "corp2"             "target2"           "natlty2"           "natlty2_txt"
 [51] "targtype3"         "targtype3_txt"     "targsubtype3"      "targsubtype3_txt"  "corp3"
 [56] "target3"           "natlty3"           "natlty3_txt"       "gname"             "gsubname"
 [61] "gname2"            "gsubname2"         "gname3"            "gsubname3"         "motive"
 [66] "guncertain1"       "guncertain2"       "guncertain3"       "individual"        "nperps"
 [71] "nperpcap"          "claimed"           "claimmode"         "claimmode_txt"     "claim2"
 [76] "claimmode2"        "claimmode2_txt"    "claim3"            "claimmode3"        "claimmode3_txt"
 [81] "compclaim"         "weaptype1"         "weaptype1_txt"     "weapsubtype1"      "weapsubtype1_txt"
 [86] "weaptype2"         "weaptype2_txt"     "weapsubtype2"      "weapsubtype2_txt"  "weaptype3"
 [91] "weaptype3_txt"     "weapsubtype3"      "weapsubtype3_txt"  "weaptype4"         "weaptype4_txt"
 [96] "weapsubtype4"      "weapsubtype4_txt"  "weapdetail"        "nkill"             "nkillus"
[101] "nkillter"          "nwound"            "nwoundus"          "nwoundte"          "property"
[106] "propextent"        "propextent_txt"    "propvalue"         "propcomment"       "ishostkid"
[111] "nhostkid"          "nhostkidus"        "nhours"            "ndays"             "divert"
[116] "kidhijcountry"     "ransom"            "ransomamt"         "ransomamtus"       "ransompaid"
[121] "ransompaidus"      "ransomnote"        "hostkidoutcome"    "hostkidoutcome_txt" "nreleased"
[126] "addnotes"          "scite1"            "scite2"            "scite3"            "dbsource"
[131] "INT_LOG"           "INT_IDEO"          "INT_MISC"          "INT_ANY"           "related"
```

```
head(Terrorismo) # Muesra las primeras 6 filas de conjunto de datos
```

```
      eventid iyear imonth iday approxdate extended resolution country        country_txt region
1 1.97000e+11  1970      7    2                  0                 58 Dominican Republic      2
2 1.97000e+11  1970      0    0                  0                130             Mexico      1
3 1.97001e+11  1970      1    0                  0                160        Philippines      5
4 1.97001e+11  1970      1    0                  0                 78             Greece      8
5 1.97001e+11  1970      1    0                  0                101              Japan      4
6 1.97001e+11  1970      1    1                  0                217      United States      1
                  region_txt provstate          city latitude longitude specificity vicinity location
1 Central America & Caribbean           Santo Domingo 18.45679 -69.95116           1        0
2              North America             Mexico city 19.43261 -99.13321           1        0
3              Southeast Asia    Tarlac       Unknown 15.47860 120.59974           4        0
4              Western Europe    Attica        Athens 37.98377  23.72816           1        0
5                   East Asia                Fukouka 33.58041 130.39636           1        0
6              North America  Illinois         Cairo 37.00511 -89.17627           1        0

1
2
3
4
5
6 1/1/1970: Unknown African American assailants fired several bullets at police headquarters in Cairo,
  crit1 crit2 crit3 doubtterr alternative alternative_txt multiple success suicide attacktype1
1     1     1     1         0          NA                        0       1       0           1
2     1     1     1         0          NA                        0       1       0           6
3     1     1     1         0          NA                        0       1       0           1
4     1     1     1         0          NA                        0       1       0           3
5     1     1     1        -9          NA                        0       1       0           7
6     1     1     1         0          NA                        0       1       0           2
               attacktype1_txt attacktype2 attacktype2_txt attacktype3 attacktype3_txt targtype1
1                Assassination          NA                          NA                          14
2    Hostage Taking (Kidnapping)          NA                          NA                           7
```

Seleccionar las variables de interés y renombrarlas

1. iyear
2. imonth
3. iday
4. country_txt
5. region_txt
6. latitude
7. longitude
8. attacktype1_txt
9. targtype1_txt
10. weaptype1_txt
11. nkill - confirmed fatalities of event
12. nwound
13. gname
14. motive
15. success

```
#Identificar y seleccionar los datos de interes a analizar
Var_Terrorismo=subset(Terrorismo,select=c(iyear,imonth,iday,country_txt,region_txt,latitude,longitude,attacktype1_txt,targtype1_txt,weaptype1_txt,nkill,nwound,gname,motive,success))
```

```
#Renombrar las columnas
Var_Terrorismo <- rename(Var_Terrorismo, c(iyear="Año"))
Var_Terrorismo <- rename(Var_Terrorismo, c(imonth="Mes"))
Var_Terrorismo <- rename(Var_Terrorismo, c(iday="Dia"))
Var_Terrorismo <- rename(Var_Terrorismo, c(country_txt="Pais"))
Var_Terrorismo <- rename(Var_Terrorismo, c(region_txt="Region"))
Var_Terrorismo <- rename(Var_Terrorismo, c(latitude="Latitud"))
Var_Terrorismo <- rename(Var_Terrorismo, c(longitude="Longitud"))
Var_Terrorismo <- rename(Var_Terrorismo, c(attacktype1_txt="Tipo_Ataque"))
Var_Terrorismo <- rename(Var_Terrorismo, c(targtype1_txt="Objetivo"))
Var_Terrorismo <- rename(Var_Terrorismo, c(weaptype1_txt="Tipo_Arma"))
Var_Terrorismo <- rename(Var_Terrorismo, c(nkill="Muertos"))
Var_Terrorismo <- rename(Var_Terrorismo, c(nwound="Heridos"))
Var_Terrorismo <- rename(Var_Terrorismo, c(gname="Grupo"))
Var_Terrorismo <- rename(Var_Terrorismo, c(motive="Motivo"))
Var_Terrorismo <- rename(Var_Terrorismo, c(success="Exitosos"))
```

2.2 ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

Identificar elementos vacíos

```
colSums(is.na(Var_Terrorismo))
```

```
     Año      Mes      Dia     Pais   Region  Latitud Longitud Tipo_Ataque 
       0        0        0        0        0     4606     4606           0 
Objetivo Tipo_Arma  Muertos  Heridos   Grupo   Motivo Exitosos 
       0        0     9682    15325        0        0        0 
```

Se ve evidencia que las columnas Latitud, Longitud, Muertos y Heridos tienen elementos vacíos. Así mismo es posible identificar el número de vacíos para cada una. Para el caso puntual de este dataset, el vacío representa información que no se tiene, estos valores no serán tenidos en cuenta durante el análisis.

Identificar valores extremos

```
chisq.out.test(Var_Terrorismo$Muertos)
```

```
        chi-squared test for outlier

data:  Var_Terrorismo$Muertos
X-squared = 17479, p-value < 2.2e-16
alternative hypothesis: highest value 1500 is an outlier
```

Se ve evidencia un outlier en el número de víctimas, el cual corresponde al año 2014. Se remueve el outlier del dataset.

```
Var_Terrorismo <- Var_Terrorismo[Var_Terrorismo$Muertos != 1500,]
```

3. Análisis de los datos.

3.1 Selección de los grupos de datos que se quieren analizar/comparar.

Se selecciona el grupo de datos que se quiere analizar, que para este caso son el número de víctimas y los años

```
MxA= Var_Terrorismo[!is.na(Var_Terrorismo$Muertos),]
MxA= subset(MxA,select=c(Muertos, Año))
```

3.2 Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

Para comprobar la normalidad y homogeneidad de la varianza se utilizan 3 pruebas diferentes:

Test de Bartlett:

```
bartlett.test(MxA$Muertos~MxA$Año, data=MxA)
```

```
        Bartlett test of homogeneity of variances

data:  MxA$Muertos by MxA$Año
Bartlett's K-squared = 84331, df = 45, p-value < 2.2e-16
```

Test de Fligner:

```
fligner.test(MxA$Muertos~MxA$Año, data=MxA)
```

```
        Fligner-Killeen test of homogeneity of variances

data:  MxA$Muertos by MxA$Año
Fligner-Killeen:med chi-squared = 7359.8, df = 45, p-value < 2.2e-16
```

Test de Levene:

levene.test(MxA$Muertos,MxA$Año)

```
        modified robust Brown-Forsythe Levene-type test

data:  MxA$Muertos
Test Statistic = 19.154, p-value < 2.2e-16
```

3.3 Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

Shapiro: Prueba de normalidad de los datos

shapiro.test(MxA$Muertos)

```
> shapiro.test(MxA$Muertos)
Error in shapiro.test(MxA$Muertos) :
  sample size must be between 3 and 5000
```

Debido a la cantidad de datos del dataset, no es posible aplicar la prueba de Shapiro

Wilcoxon:

wilcox.test(MxA$Muertos,MxA$Año, data=MxA)

```
        Wilcoxon rank sum test with continuity correction

data:  MxA$Muertos and MxA$Año
W = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

kruskal:

kruskal.test(MxA$Muertos,MxA$Año, data=MxA)

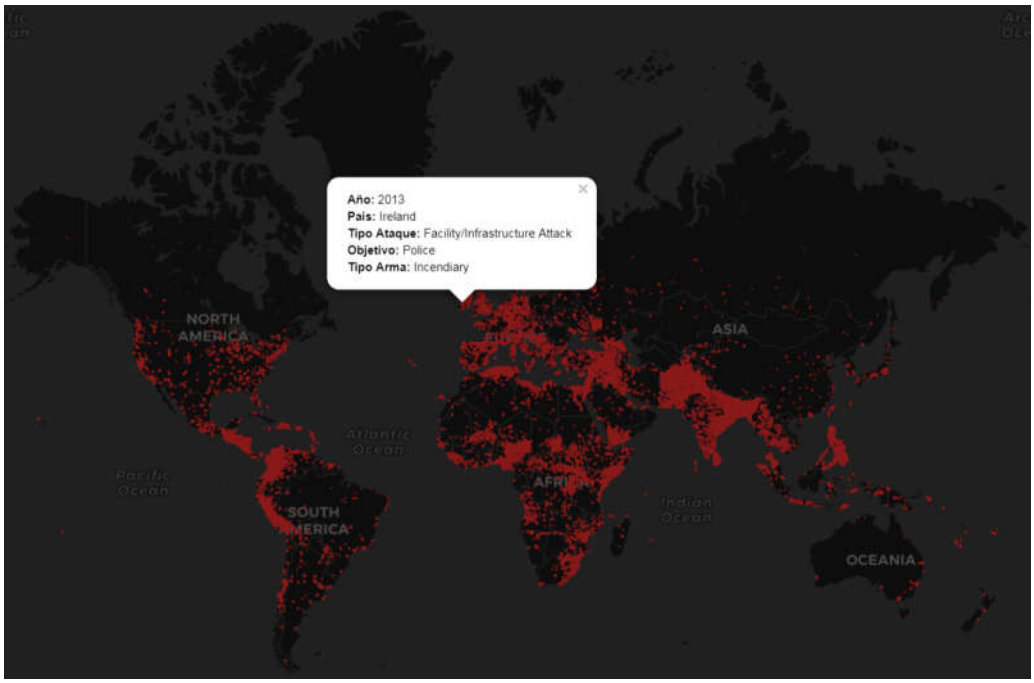
```
        Kruskal-Wallis rank sum test

data:  MxA$Muertos and MxA$Año
Kruskal-Wallis chi-squared = 4041.3, df = 45, p-value < 2.2e-16
```

4. Representación de los resultados a partir de tablas y gráficas.

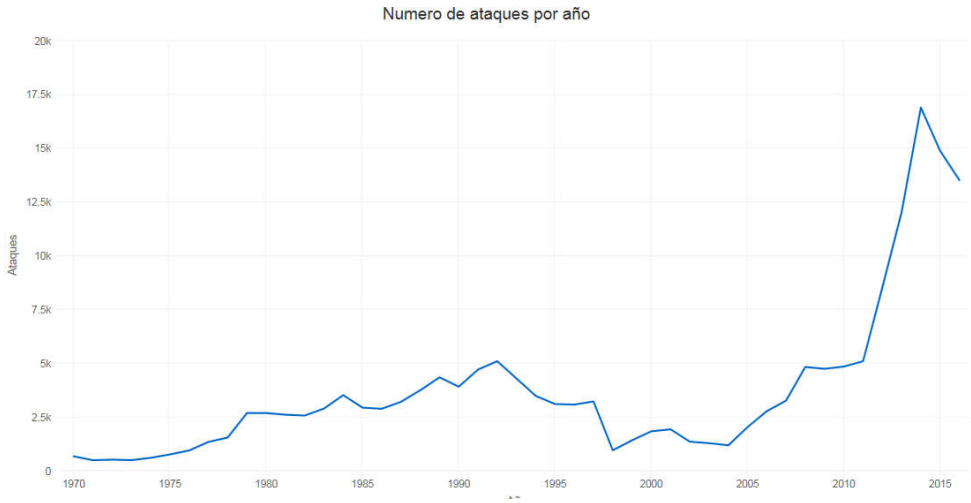
Se elaboran diferentes visualizaciones de datos que facilitan la interpretación de los resultados.

- Mapa interactivo de ataques en el mundo



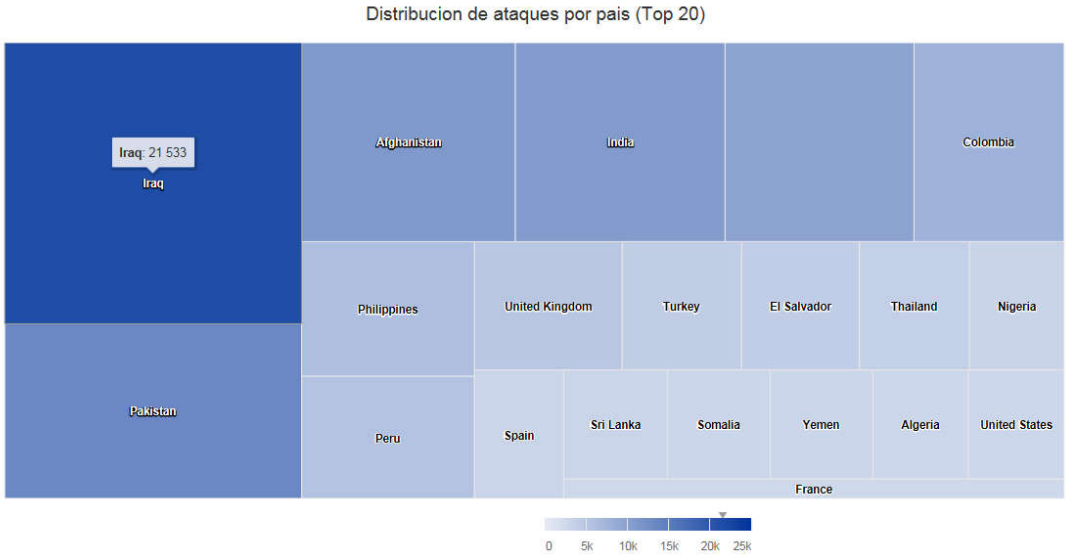
Grafica 1. Mapa interactivo de los ataques perpetrados alrededor del mundo

- Número de ataques terroristas por año



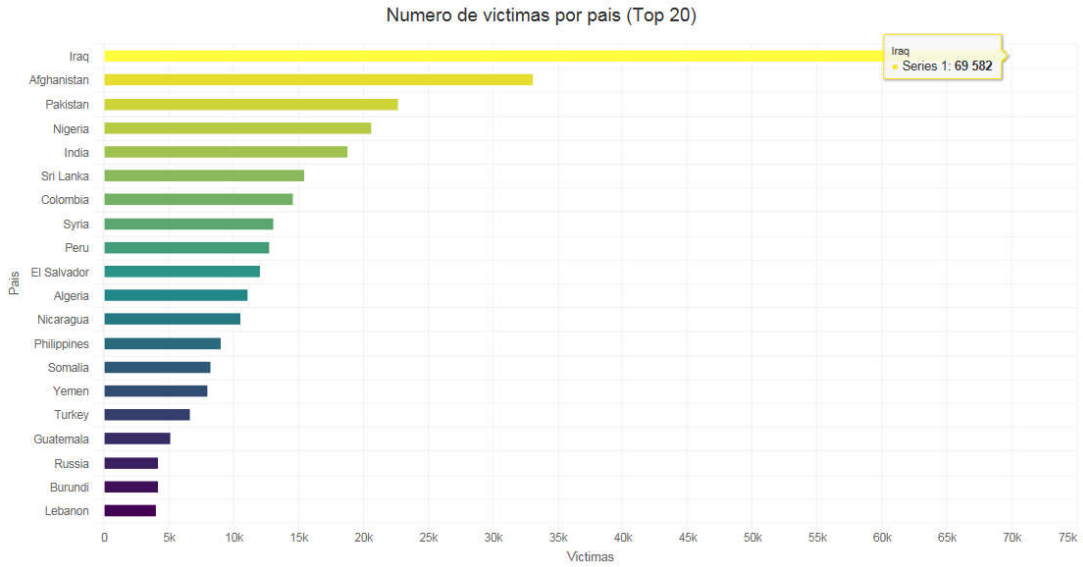
Grafica 2. Número de ataques terroristas por cada año

- Número de ataques terroristas por país



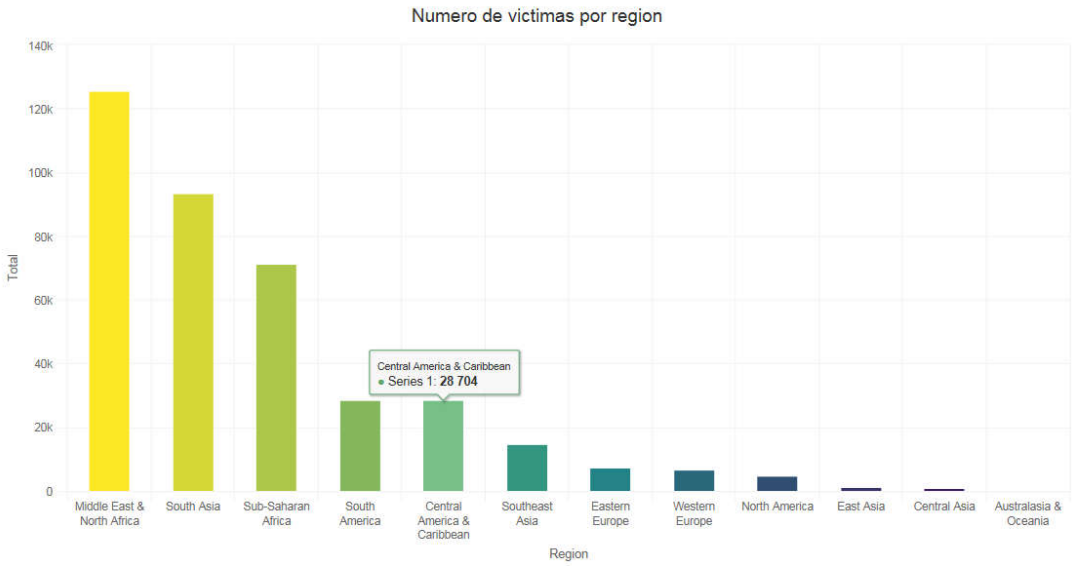
Grafica 3. Distribución de ataques terroristas por país (Top 20)

- Número de víctimas por país



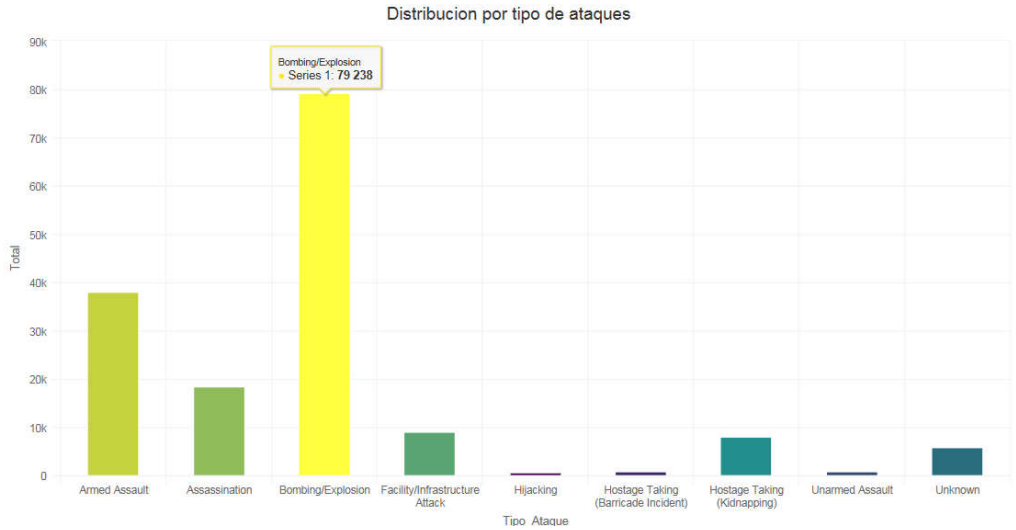
Grafica 4. Número de víctimas fatales por país (Top 20)

- Número de víctimas por región



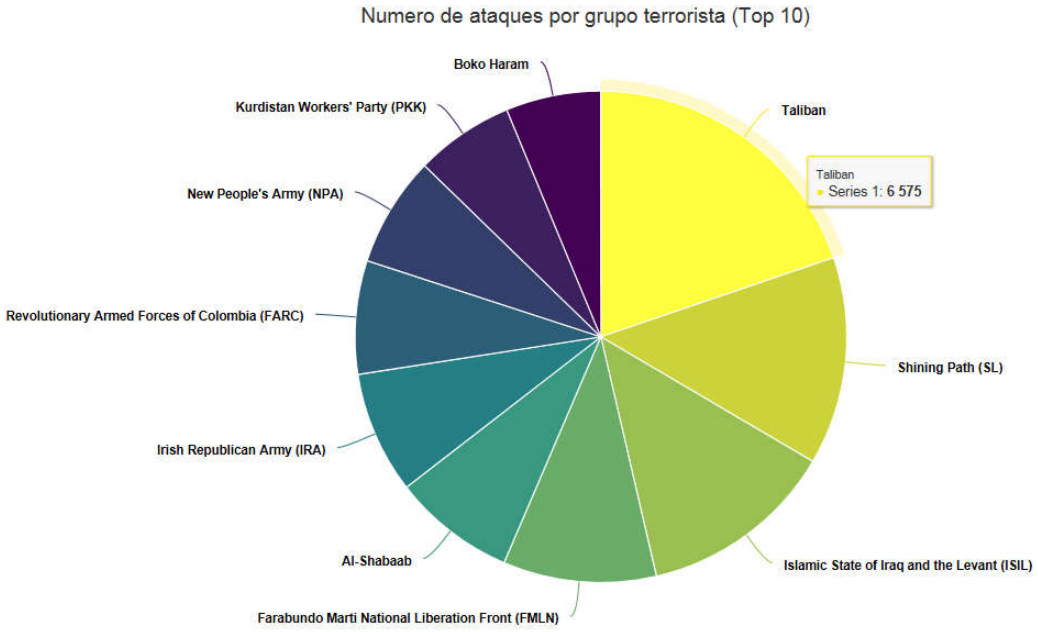
Grafica 5. Número de víctimas fatales por región

- Número de ataques terroristas por tipo



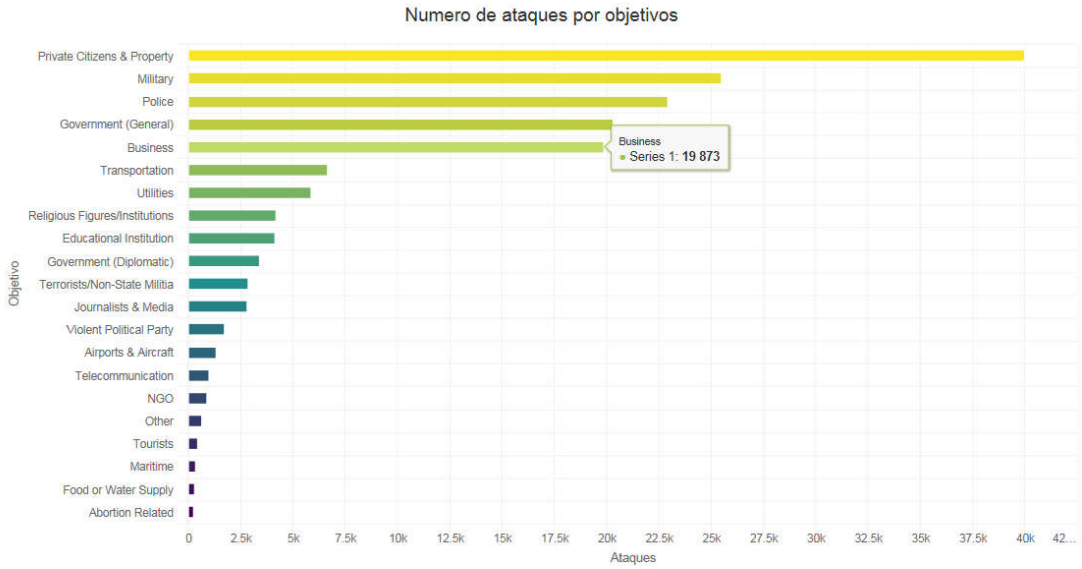
Grafica 6. Distribución de ataques terroristas por tipo

- Número de ataques terroristas por grupo



Grafica 7. Número de ataques por grupo terrorista (Top 20)

- Número de ataques terroristas por objetivo



Grafica 8. Número de ataques por objetivos

5. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Una vez se han ejecutado los pasos anteriores, se pueden resolver las preguntas planteadas.

¿Cuál es el número de ataques terroristas realizados por año?. En la gráfica 1 se pueden identificar el número de ataques terrorista perpetrados por cada año, identificando un incremento significativo en la última década, especialmente del año 2011 en adelante, donde han pasado de aproximadamente 5000 por año a 15000 en el 2016.

Año	Numero Ataques
2011	5071
2012	8500
2013	11996
2014	16860
2015	14852
2016	13488

¿Cuáles son los países más afectados por el terrorismo?. En la gráfica 2 es posible ver que los 5 países que han sufrido un mayor número de ataques terroristas en el tiempo son:

País	Numero Ataques
Iraq	22130
Pakistan	13634
Afghanistan	11306
India	10978
Colombia	8163

Sin embargo, no todos los países con el mayor número de ataques terrorista son precisamente los que mayor cantidad de víctimas fatales han tenido. En la gráfica 3 se evidencian los países con más víctimas:

País	Numero Victimas
Iraq	71082
Afghanistan	33146
Pakistan	22734
Nigeria	20665
India	18842

¿Cuáles son las regiones más afectados por el terrorismo?. La grafica 4 muestra que las regiones con mayor número de víctimas son:

Región	Numero Victimas
Middle East & North Africa	125676
South Asia	93434
PakistanSub-Saharan Africa	71244
South America	28730
Central America & Caribean	28704

¿Qué tipos de ataques han causado más muertes?. En la gráfica 5 se puede ver que los 3 tipos de ataques que han causado mayor número de muertes son:

Tipo Ataque	Numero Victimas
Bombing / Explosion	87073
Armed Assault	40223
Assassination	18402

¿Cuáles son los grupos terroristas que perpetran mayor número de ataques terroristas? La grafica 6 permite ver los principales grupos terroristas y el número de ataques cometido por cada uno de estos:

Grupo Terrorista	Numero Ataques
Taliban	6575
Shining Path (SL)	4551
Islamic State of Iraq and the Levant	4287
Farabundo Marti National Liberation Front (FMLN)	3351
Al-Shabaab	2683

¿Quiénes son los principales objetivos de los ataques terroristas?. Los principales objetivos del terrorismo, como se observa en la gráfica 7, son:

Objetivo	Numero Ataques
Private Citizens & Property	39994
Military	25508
Police	22938
Goverment	20314
Business	19873

6. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código en R utilizado para la realización de esta práctica se encuentra en Github.