



**BUAP**

**ANÁLISIS DE TENDENCIAS GLOBALES  
SOBRE ATAQUES TERRORISTAS  
UTILIZANDO LA BASE DE DATOS  
“GLOBAL TERRORISM DATABASE” GTD**

**Procesamiento y  
Limpieza de Datos**

Angel Uriel Lopez Vazquez  
Introducción a la Ciencia de Datos  
Jaime Alejandro Romero Sierra  
Martes 08:00 – 10:00

# **ANALISIS INICIAL DE LA BASE DE DATOS**

## Descripcion

La base de datos utilizada para este análisis es una versión modificada de la Global Terrorism Database, que contiene aproximadamente 200,000 entradas distribuidas en 135 columnas. Estas columnas incluyen información sobre la localización del ataque, el tipo de ataque, las tácticas empleadas, los grupos responsables, los objetivos y los resultados, abarcando más de 180,000 incidentes registrados entre 1970 y 2017, con la excepción del año 1993.

## Valores Faltantes

Para calcular el número de valores faltantes en el DataFrame, se sumaron los valores ausentes de cada columna, lo que resultó en un total de **14,999,396**. Este número representa casi **75 veces** el total de entradas en el DataFrame, lo que sugiere la existencia de columnas casi vacías o completamente vacías. Por lo tanto, se decidió revisar cada columna, obteniendo los resultados que se presentan a continuación.

Columna	eventid	year	imonth	iday	approxdate	extended	resolution	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	location	summary	crit1	crit2	crit3	doubtterr
Valores Faltantes	1383	175	143	188	186212	164	193850	152	156	147	176	612	605	5084	5066	158	172	136324	71414	158	1430	177	161
Porcentaje Faltante	0.70037	0.088622	0.072417	0.095206	94.300313	0.083052	98.168302	0.076975	0.079001	0.074443	0.089129	0.309925	0.30638	2.574607	2.565492	0.080013	0.087103	69.036345	36.16503	0.080013	0.724172	0.089635	0.081533

Columna eventid:

Faltante: 1383 | Porcentaje: 0.70%

Columna iyear:

Faltante: 175 | Porcentaje: 0.09%

Columna imonth:

Faltante: 143 | Porcentaje: 0.07%

Columna iday:

Faltante: 188 | Porcentaje: 0.10%

Columna approxdate:

Faltante: 186212 | Porcentaje: 94.30%

Columna extended:

Faltante: 164 | Porcentaje: 0.08%

Columna resolution:

Faltante: 193850 | Porcentaje: 98.17%

Columna country:

Faltante: 152 | Porcentaje: 0.08%

Columna country\_txt:

Faltante: 156 | Porcentaje: 0.08%

Columna region:

Faltante: 147 | Porcentaje: 0.07%

Columna region\_txt:

Faltante: 176 | Porcentaje: 0.09%

Columna provstate:

Faltante: 612 | Porcentaje: 0.31%

Columna city:

Faltante: 605 | Porcentaje: 0.31%

Columna latitude:

Faltante: 5084 | Porcentaje: 2.57%

Columna longitude:

Faltante: 5066 | Porcentaje: 2.57%

Columna specificity:

Faltante: 158 | Porcentaje: 0.08%

Columna vicinity:

Faltante: 172 | Porcentaje: 0.09%

Columna location:

Faltante: 136324 | Porcentaje: 69.04%

## Valores Faltantes

Columna summary:

Faltante: 71414 | Porcentaje: 36.17%

Columna crit1:

Faltante: 158 | Porcentaje: 0.08%

Columna crit2:

Faltante: 1430 | Porcentaje: 0.72%

Columna crit3:

Faltante: 177 | Porcentaje: 0.09%

Columna doubtterr:

Faltante: 161 | Porcentaje: 0.08%

Columna alternative:

Faltante: 166174 | Porcentaje: 84.15%

Columna alternative\_txt:

Faltante: 164897 | Porcentaje: 83.51%

Columna multiple:

Faltante: 159 | Porcentaje: 0.08%

Columna success:

Faltante: 172 | Porcentaje: 0.09%

Columna suicide:

Faltante: 131 | Porcentaje: 0.07%

Columna attacktype1:

Faltante: 172 | Porcentaje: 0.09%

Columna attacktype1\_txt:

Faltante: 154 | Porcentaje: 0.08%

Columna attacktype2:

Faltante: 189318 | Porcentaje: 95.87%

Columna attacktype2\_txt:

Faltante: 189384 | Porcentaje: 95.91%

Columna attacktype3:

Faltante: 195698 | Porcentaje: 99.10%

Columna attacktype3\_txt:

Faltante: 195718 | Porcentaje: 99.11%

Columna targtype1:

Faltante: 1493 | Porcentaje: 0.76%

Columna targtype1\_txt:

Faltante: 169 | Porcentaje: 0.09%

Columna targsubtype1:

Faltante: 11357 | Porcentaje: 5.75%

Columna targsubtype1\_txt:

Faltante: 11330 | Porcentaje: 5.74%

Columna corp1:

Faltante: 46032 | Porcentaje: 23.31%

Columna target1:

Faltante: 849 | Porcentaje: 0.43%

Columna natlty1:

Faltante: 1845 | Porcentaje: 0.93%

Columna natlty1\_txt:

Faltante: 1841 | Porcentaje: 0.93%

Columna targtype2:

Faltante: 184094 | Porcentaje: 93.23%

Columna targtype2\_txt:

Faltante: 184120 | Porcentaje: 93.24%

Columna targsubtype2:

Faltante: 184625 | Porcentaje: 93.50%

Columna targsubtype2\_txt:

Faltante: 184570 | Porcentaje: 93.47%

Columna corp2:

Faltante: 197467 | Porcentaje: 100.00%

Columna target2:

Faltante: 184273 | Porcentaje: 93.32%

Columna natlty2:

Faltante: 184505 | Porcentaje: 93.44%

Columna natlty2\_txt:

Faltante: 184485 | Porcentaje: 93.43%

Columna targtype3:

Faltante: 194954 | Porcentaje: 98.73%

Columna targtype3\_txt:

Faltante: 194915 | Porcentaje: 98.71%

## Valores Faltantes

Columna targsubtype3:

Faltante: 195009 | Porcentaje: 98.76%

Columna targsubtype3\_txt:

Faltante: 195015 | Porcentaje: 98.76%

Columna corp3:

Faltante: 195069 | Porcentaje: 98.79%

Columna target3:

Faltante: 194969 | Porcentaje: 98.73%

Columna natlty3:

Faltante: 195010 | Porcentaje: 98.76%

Columna natlty3\_txt:

Faltante: 194843 | Porcentaje: 98.67%

Columna gname:

Faltante: 151 | Porcentaje: 0.08%

Columna gsubname:

Faltante: 189860 | Porcentaje: 96.15%

Columna gname2:

Faltante: 194023 | Porcentaje: 98.26%

Columna gsubname2:

Faltante: 196076 | Porcentaje: 99.30%

Columna gname3:

Faltante: 195810 | Porcentaje: 99.16%

Columna gsubname3:

Faltante: 196132 | Porcentaje: 99.32%

Columna motive:

Faltante: 141594 | Porcentaje: 71.71%

Columna guncertain1:

Faltante: 564 | Porcentaje: 0.29%

Columna guncertain2:

Faltante: 193992 | Porcentaje: 98.24%

Columna guncertain3:

Faltante: 195851 | Porcentaje: 99.18%

Columna individual:

Faltante: 159 | Porcentaje: 0.08%

Columna nperps:

Faltante: 76922 | Porcentaje: 38.95%

Columna nperpcap:

Faltante: 75066 | Porcentaje: 38.01%

Columna claimed:

Faltante: 71431 | Porcentaje: 36.17%

Columna claimmode:

Faltante: 175583 | Porcentaje: 88.92%

Columna claimmode\_txt:

Faltante: 175562 | Porcentaje: 88.91%

Columna claim2:

Faltante: 194061 | Porcentaje: 98.28%

Columna claimmode2:

Faltante: 195575 | Porcentaje: 99.04%

Columna claimmode2\_txt:

Faltante: 195530 | Porcentaje: 99.02%

Columna claim3:

Faltante: 195817 | Porcentaje: 99.16%

Columna claimmode3:

Faltante: 196040 | Porcentaje: 99.28%

Columna claimmode3\_txt:

Faltante: 196133 | Porcentaje: 99.32%

Columna compclaim:

Faltante: 190998 | Porcentaje: 96.72%

Columna weaptype1:

Faltante: 150 | Porcentaje: 0.08%

Columna weaptype1\_txt:

Faltante: 163 | Porcentaje: 0.08%

Columna weapsubtype1:

Faltante: 22607 | Porcentaje: 11.45%

Columna weapsubtype1\_txt:

Faltante: 22607 | Porcentaje: 11.45%

Columna weaptype2:

Faltante: 182015 | Porcentaje: 92.17%

## Valores Faltantes

Columna weaptype2\_txt:

Faltante: 182037 | Porcentaje: 92.19%

Columna weapsubtype2:

Faltante: 183746 | Porcentaje: 93.05%

Columna weapsubtype2\_txt:

Faltante: 183685 | Porcentaje: 93.02%

Columna weaptype3:

Faltante: 194108 | Porcentaje: 98.30%

Columna weaptype3\_txt:

Faltante: 194141 | Porcentaje: 98.32%

Columna weapsubtype3:

Faltante: 194383 | Porcentaje: 98.44%

Columna weapsubtype3\_txt:

Faltante: 194404 | Porcentaje: 98.45%

Columna weaptype4:

Faltante: 196176 | Porcentaje: 99.35%

Columna weaptype4\_txt:

Faltante: 196154 | Porcentaje: 99.34%

Columna weapsubtype4:

Faltante: 196114 | Porcentaje: 99.31%

Columna weapsubtype4\_txt:

Faltante: 196109 | Porcentaje: 99.31%

Columna weapdetail:

Faltante: 73188 | Porcentaje: 37.06%

Columna nkill:

Faltante: 11261 | Porcentaje: 5.70%

Columna nkillus:

Faltante: 69642 | Porcentaje: 35.27%

Columna nkillter:

Faltante: 72350 | Porcentaje: 36.64%

Columna nwound:

Faltante: 17722 | Porcentaje: 8.97%

Columna nwoundus:

Faltante: 69930 | Porcentaje: 35.41%

Columna nwoundte:

Faltante: 74693 | Porcentaje: 37.83%

Columna property:

Faltante: 141 | Porcentaje: 0.07%

Columna propextent:

Faltante: 128297 | Porcentaje: 64.97%

Columna propextent\_txt:

Faltante: 126977 | Porcentaje: 64.30%

Columna propvalue:

Faltante: 154098 | Porcentaje: 78.04%

Columna propcomment:

Faltante: 133642 | Porcentaje: 67.68%

Columna ishostkid:

Faltante: 344 | Porcentaje: 0.17%

Columna nhostkid:

Faltante: 181556 | Porcentaje: 91.94%

Columna nhostkidus:

Faltante: 181531 | Porcentaje: 91.93%

Columna nhours:

Faltante: 191857 | Porcentaje: 97.16%

Columna ndays:

Faltante: 187438 | Porcentaje: 94.92%

Columna divert:

Faltante: 195810 | Porcentaje: 99.16%

Columna kidhijcountry:

Faltante: 192632 | Porcentaje: 97.55%

Columna ransom:

Faltante: 112779 | Porcentaje: 57.11%

Columna ransomamt:

Faltante: 194696 | Porcentaje: 98.60%

Columna ransomamtus:

Faltante: 195628 | Porcentaje: 99.07%

Columna ransompaid:

Faltante: 195349 | Porcentaje: 98.93%

## Valores Faltantes

Columna ransompaidus:

Faltante: 195616 | Porcentaje: 99.06%

Columna ransomnote:

Faltante: 195558 | Porcentaje: 99.03%

Columna hostkidoutcome:

Faltante: 184379 | Porcentaje: 93.37%

Columna hostkidoutcome\_txt:

Faltante: 197467 | Porcentaje: 100.00%

Columna nreleased:

Faltante: 184922 | Porcentaje: 93.65%

Columna addnotes:

Faltante: 165679 | Porcentaje: 83.90%

Columna scite1:

Faltante: 71535 | Porcentaje: 36.23%

Columna scite2:

Faltante: 113104 | Porcentaje: 57.28%

Columna scite3:

Faltante: 149234 | Porcentaje: 75.57%

Columna dbsource:

Faltante: 155 | Porcentaje: 0.08%

Columna INT\_LOG:

Faltante: 144 | Porcentaje: 0.07%

Columna INT\_IDEO:

Faltante: 155 | Porcentaje: 0.08%

Columna INT\_MISC:

Faltante: 153 | Porcentaje: 0.08%

Columna INT\_ANY:

Faltante: 166 | Porcentaje: 0.08%

Columna related:

Faltante: 169116 | Porcentaje: 85.64%

## Valores Duplicados

Por medio del comando `df.duplicated().sum()`, se encontro que el numero total de filas duplicadas es de **4094 filas**.

## Tipos de Datos

eventid: float64

iyear: object

imonth: object

iday: object

approxdate: object

extended: object

resolution: object

country: object

country\_txt: object

region: object

region\_txt: object

provstate: object

city: object

latitude: object

longitude: object

specificity: object

vicinity: object

location: object

summary: object

crit1: object

crit2: float64

crit3: object

doubtterr: object

alternative: float64

alternative\_txt: object

multiple: object

success: object

suicide: object

attacktype1: object

attacktype1\_txt: object

attacktype2: object

attacktype2\_txt: object

attacktype3: object

attacktype3\_txt: object

targtype1: float64

targtype1\_txt: object

## Tipos de Datos

targsubtype1: object	claimed: object	propextent_txt: object
targsubtype1_txt: object	claimmode: object	propvalue: object
corp1: object	claimmode_txt: object	propcomment: object
target1: object	claim2: object	ishostkid: object
natlty1: object	claimmode2: object	nhostkid: object
natlty1_txt: object	claimmode2_txt: object	nhostkidus: object
targtype2: object	claim3: object	nhours: object
targtype2_txt: object	claimmode3: object	ndays: object
targsubtype2: object	claimmode3_txt: object	divert: object
targsubtype2_txt: object	compclaim: object	kidhijcountry: object
corp2: float64	weaptype1: object	ransom: object
target2: object	weaptype1_txt: object	ransomamt: object
natlty2: object	weapsubtype1: object	ransomamtus: object
natlty2_txt: object	weapsubtype1_txt: object	ransompaid: object
targtype3: object	weaptype2: object	ransompaidus: object
targtype3_txt: object	weaptype2_txt: object	ransomnote: object
targsubtype3: object	weapsubtype2: object	hostkidoutcome: object
targsubtype3_txt: object	weapsubtype2_txt: object	hostkidoutcome_txt:
corp3: object	weaptype3: object	float64
target3: object	weaptype3_txt: object	nreleased: object
natlty3: object	weapsubtype3: object	addnotes: object
natlty3_txt: object	weapsubtype3_txt: object	scite1: object
gname: object	weaptype4: object	scite2: object
gsubname: object	weaptype4_txt: object	scite3: object
gname2: object	weapsubtype4: object	dbsource: object
gsubname2: object	weapsubtype4_txt: object	INT_LOG: object
gname3: object	weapdetail: object	INT_IDEO: object
gsubname3: object	nkill: object	INT_MISC: object
motive: object	nkillus: object	INT_ANY: object
guncertain1: object	nkillter: object	related: object
guncertain2: object	nwound: object	
guncertain3: object	nwoundus: object	
individual: object	nwoundte: object	
nperps: object	property: object	
nperpcap: object	propextent: float64	



### **Problemas Identificados**

La base de datos presenta diversas problemáticas que dificultan su correcta gestión y análisis, entre las cuales se destacan las siguientes:

1. Ausencia significativa de datos: Varias columnas presentan un alto porcentaje de valores faltantes, con rangos que oscilan entre el 40% y el 90%, lo que compromete la integridad de la información.
2. Valores inválidos: Se han detectado datos no válidos, tanto en formato numérico como textual, así como la presencia de números expresados en notación científica, lo cual genera inconsistencias en el conjunto de datos.
3. Formatos inadecuados: La mayoría de las columnas contiene tipos de datos que no corresponden a su contenido, como valores numéricos en campos que deberían contener texto y viceversa.
4. Información irrelevante o excesivamente detallada: Existen columnas con descripciones textuales demasiado específicas o con datos que no resultan pertinentes para los fines del análisis.
5. Columnas innecesarias: Se identificaron columnas con información redundante o irrelevante para los objetivos del análisis, lo que afecta la eficiencia del procesamiento de los datos.

# **PROCESO DE LIMPIEZA DE LA BASE DE DATOS**

## Eliminación de Columnas Innecesarias

En primer lugar, para evitar la pérdida de datos al eliminar entradas inválidas en columnas irrelevantes, se decidió eliminar todas aquellas columnas, tanto numéricas como textuales, que presentaban más del 50% de datos faltantes y/o inválidos, siempre que no fuera viable realizar una imputación adecuada de dichos valores. Además, se consideraron otros criterios, como la especificidad de la información y su utilidad para proporcionar detalles relevantes sobre los ataques.

```
pd.set_option('display.max_columns', None)
df.head()
```

	eventid	year	imonth	may	approxdate	extended	resolution	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	location	summary	crit1	crit2	crit3	doubtterr	alternative	alternative
0	NaN	1970.0	7.0	2.0	NaN	0.0	NaN	58.0	Dominican Republic	2.0	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1.0	0.0	NaN	NaN	1.0	1.0	1.0	0.0	NaN	
1	1.970000e+11	1970.0	0.0	0.0	NaN	0.0	NaN	130.0	Mexico	1.0	North America	Federal	Mexico city	19.371887	-99.066624	1.0	0.0	NaN	NaN	1.0	1.0	1.0	0.0	NaN	
2	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	NaN	160.0	Philippines	5.0	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4.0	0.0	NaN	NaN	1.0	1.0	1.0	0.0	NaN	
3	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	NaN	78.0	Greece	8.0	Western Europe	Attica	Athens	37.99749	23.762728	1.0	0.0	NaN	NaN	1.0	1.0	1.0	0.0	NaN	
4	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	NaN	101.0	Japan	4.0	East Asia	Fukouka	Fukouka	33.580412	130.396361	1.0	0.0	NaN	NaN	1.0	1.0	1.0	-9.0	NaN	

```
columns_to_drop = [
    'resolution', 'location', 'summary', 'alternative', 'alternative_txt',
    'attcktype2', 'attcktype2_txt', 'attcktype3', 'attcktype3_txt',
    'target2', 'target2_txt', 'target2type2', 'target2type2_txt',
    'corp2', 'target2', 'natly2', 'natly2_txt', 'target3', 'target3_txt',
    'target3type3', 'target3type3_txt', 'corp3', 'target3', 'natly3', 'natly3_txt',
    'gsubname', 'gname2', 'gsubname2', 'gname3', 'gsubname3', 'guncertain2',
    'guncertain3', 'claled', 'claim2', 'claimmode', 'claimmode_txt', 'claid2', 'claimmode2',
    'claimmode2_txt', 'claid3', 'claimmode3', 'claimmode3_txt', 'compclaid',
    'weaptype2', 'weaptype2_txt', 'weapsubtype2', 'weapsubtype2_txt',
    'weaptype3', 'weaptype3_txt', 'weapsubtype3', 'weapsubtype3_txt',
    'weaptype4', 'weaptype4_txt', 'weapsubtype4', 'weapsubtype4_txt',
    'propcomment', 'related', 'adnotes', 'scitel', 'scitel2', 'scitel3', 'motive',
    'killus', 'hpercap', 'killiter', 'hounds', 'hounds2', 'propvalue', 'hhostkid',
    'hhostkidus', 'hhours', 'ndays', 'divert', 'kidhcountry', 'ransomamt',
    'ransomamtus', 'ransompaid', 'ransompaidus', 'ransomnote', 'hostkidoutcome',
    'hostkidoutcome_txt', 'released', 'corp1', 'weapdetail', 'target1'
]

df1 = df1.drop(columns=columns_to_drop)
df1.head()
```

	eventid	year	imonth	may	approxdate	extended	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	crit1	crit2	crit3	doubtterr	multiple	success	suicide	attacktype1	attacktype
0	NaN	1970.0	7.0	2.0	NaN	0.0	58.0	Dominican Republic	2.0	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	Assasie
1	1.970000e+11	1970.0	0.0	0.0	NaN	0.0	130.0	Mexico	1.0	North America	Federal	Mexico city	19.371887	-99.066624	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	6.0	Hostage 1 (Kidnap
2	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	160.0	Philippines	5.0	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	Assasie
3	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	78.0	Greece	8.0	Western Europe	Attica	Athens	37.99749	23.762728	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	3.0	Bombing/Exp
4	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	101.0	Japan	4.0	East Asia	Fukouka	Fukouka	33.580412	130.396361	1.0	0.0	1.0	1.0	1.0	-9.0	0.0	1.0	0.0	7.0	Facility/Infestr

Este proceso permitió reducir el número de columnas de 135 a 51, eliminando aquellas que no aportaban valor al análisis y conservando únicamente las que contienen información relevante y de calidad para los objetivos del estudio.

## Eliminación de Columnas Innecesarias

En primer lugar, para evitar la pérdida de datos al eliminar entradas inválidas en columnas irrelevantes, se decidió eliminar todas aquellas columnas, tanto numéricas como textuales, que presentaban más del 50% de datos faltantes y/o inválidos, siempre que no fuera viable realizar una imputación adecuada de dichos valores. Además, se consideraron otros criterios, como la especificidad de la información y su utilidad para proporcionar detalles relevantes sobre los ataques.

```
pd.set_option('display.max_columns', None)
df.head()
```

	eventid	year	imonth	iday	approxdate	extended	resolution	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	location	summary	crit1	crit2	crit3	doubtterr	alternative	alternative
0	NaN	1970.0	7.0	2.0	NaN	0.0	NaN	58.0	Dominican Republic	2.0	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1.0	0.0	NaN	NaN	1.0	1.0	1.0	0.0	NaN	
1	1.970000e+11	1970.0	0.0	0.0	NaN	0.0	NaN	130.0	Mexico	1.0	North America	Federal	Mexico city	19.371887	-99.066624	1.0	0.0	NaN	NaN	1.0	1.0	1.0	0.0	NaN	
2	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	NaN	160.0	Philippines	5.0	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4.0	0.0	NaN	NaN	1.0	1.0	1.0	0.0	NaN	
3	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	NaN	78.0	Greece	8.0	Western Europe	Attica	Athens	37.99749	23.762728	1.0	0.0	NaN	NaN	1.0	1.0	1.0	0.0	NaN	
4	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	NaN	101.0	Japan	4.0	East Asia	Fukouka	Fukouka	33.580412	130.396361	1.0	0.0	NaN	NaN	1.0	1.0	1.0	-9.0	NaN	

```
columns_to_drop = [
    'resolution', 'location', 'summary', 'alternative', 'alternative_txt',
    'attacktype2', 'attacktype2_txt', 'attacktype3', 'attacktype3_txt',
    'targettype2', 'targettype2_txt', 'targettype3', 'targettype3_txt',
    'corp2', 'target2', 'natity2', 'natity2_txt', 'target3', 'target3_txt',
    'targettype3', 'targettype3_txt', 'corp3', 'target3', 'natity3', 'natity3_txt',
    'gsubname', 'gname2', 'gsubname2', 'gname3', 'gsubname3', 'guncertain2',
    'guncertain3', 'claim2', 'claim3', 'claimmode', 'claimmode_txt', 'claim2', 'claimmode2',
    'claimmode2_txt', 'claim3', 'claimmode3', 'claimmode3_txt', 'compclaim',
    'weaptype2', 'weaptype2_txt', 'weapsubtype2', 'weapsubtype2_txt',
    'weaptype3', 'weaptype3_txt', 'weapsubtype3', 'weapsubtype3_txt',
    'weaptype4', 'weaptype4_txt', 'weapsubtype4', 'weapsubtype4_txt',
    'propcomment', 'related', 'adnotes', 'scitel', 'scitel2', 'scitel3', 'motive',
    'killus', 'hopercap', 'killiter', 'hounds', 'hounds2', 'propvalue', 'hhostkid',
    'hhostkidus', 'hhours', 'ndays', 'divert', 'kidhcountry', 'ransomamt',
    'ransomamtus', 'ransompaid', 'ransompaidus', 'ransomnote', 'hostkidoutcome',
    'hostkidoutcome_txt', 'released', 'corp1', 'weapdetail', 'target1'
]

df1 = df1.drop(columns=columns_to_drop)
df1.head()
```

	eventid	year	imonth	iday	approxdate	extended	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	crit1	crit2	crit3	doubtterr	multiple	success	suicide	attacktype1	attacktype
0	NaN	1970.0	7.0	2.0	NaN	0.0	58.0	Dominican Republic	2.0	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	Assasie
1	1.970000e+11	1970.0	0.0	0.0	NaN	0.0	130.0	Mexico	1.0	North America	Federal	Mexico city	19.371887	-99.066624	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	6.0	Hostage 1 (Kidnap
2	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	160.0	Philippines	5.0	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	Assasie
3	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	78.0	Greece	8.0	Western Europe	Attica	Athens	37.99749	23.762728	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	3.0	Bombing/Exp
4	1.970010e+11	1970.0	1.0	0.0	NaN	0.0	101.0	Japan	4.0	East Asia	Fukouka	Fukouka	33.580412	130.396361	1.0	0.0	1.0	1.0	1.0	-9.0	0.0	1.0	0.0	7.0	Facility/Infestr

Este proceso permitió reducir el número de columnas de 135 a 51, eliminando aquellas que no aportaban valor al análisis y conservando únicamente las que contienen información relevante y de calidad para los objetivos del estudio.

## Imputacion de Valores y Remplazo de tipos de datos

En primer lugar, se procedió a modificar los valores de la columna eventid, que estaban en notación científica. Estos fueron reemplazados por números consecutivos, asignando valores desde 1 hasta el número total de registros del DataFrame, garantizando que cada evento tuviera un identificador único y de fácil manejo.

```
df['eventid'] = range(1, len(df) + 1)
df.head()
```

	eventid	year	imonth	iday	approxdate	extended	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	crit1	crit2	crit3	doubtterr	multiple	success	suicide	attacktype1	attacktype1_txt
0	1	1970.0	7.0	2.0	NaN	0.0	58.0	Dominican Republic	2.0	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	Assassination
1	2	1970.0	0.0	0.0	NaN	0.0	130.0	Mexico	1.0	North America	Federal	Mexico city	19.371887	-99.086624	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	6.0	Hostage Taking (Kidnapping)
2	3	1970.0	1.0	0.0	NaN	0.0	160.0	Philippines	5.0	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	Assassination
3	4	1970.0	1.0	0.0	NaN	0.0	78.0	Greece	8.0	Western Europe	Attica	Athens	37.99749	23.762728	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	3.0	Bombing/Explosion
4	5	1970.0	1.0	0.0	NaN	0.0	101.0	Japan	4.0	East Asia	Fukouka	Fukouka	33.580412	130.396361	1.0	0.0	1.0	1.0	1.0	-9.0	0.0	1.0	0.0	7.0	Facility/Infrastructure Attack

Posteriormente, se reemplazaron todos aquellos valores numéricos, asignándoles el tipo de dato correspondiente: se convirtieron a tipo entero o flotante según fuera necesario, en función de los datos que contenían.

```
columns_df = ['eventid', 'year', 'imonth', 'iday', 'approxdate', 'extended', 'country', 'country_txt', 'region', 'region_txt', 'provstate', 'city', 'latitude', 'longitude', 'specificity', 'vicinity', 'crit1', 'crit2', 'crit3', 'doubtterr', 'multiple', 'success', 'suicide', 'attacktype1', 'attacktype1_txt', 'tar']

for col in columns_df:
    df.dropna(subset=[col], inplace=True)
    df[col] = df[col].astype(float)
    df[col] = df[col].astype(int)

df.head()
```

	eventid	year	imonth	iday	approxdate	extended	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	crit1	crit2	crit3	doubtterr	multiple	success	suicide	attacktype1	attacktype1_txt	tar
1	1	1970	7	2	NaN	0	58	Dominican Republic	2	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1	0	1	1	1	0	0	1	0	1	Assassination	
2	2	1970	0	0	NaN	0	130	Mexico	1	North America	Federal	Mexico city	19.371887	-99.086624	1	0	1	1	1	0	0	1	0	6	Hostage Taking (Kidnapping)	
3	3	1970	1	0	NaN	0	160	Philippines	5	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4	0	1	1	1	0	0	1	0	1	Assassination	
4	4	1970	1	0	NaN	0	78	Greece	8	Western Europe	Attica	Athens	37.99749	23.762728	1	0	1	1	1	0	0	1	0	3	Bombing/Explosion	
5	5	1970	1	0	NaN	0	101	Japan	4	East Asia	Fukouka	Fukouka	33.580412	130.396361	1	0	1	1	1	-9	0	1	0	7	Facility/Infrastructure Attack	

En las columnas iday e imonth, los valores "0" fueron reemplazados por la media de los datos correspondientes, lo que permite una mayor coherencia en la información. Esta imputación facilita el manejo y análisis futuro de estas columnas, ya que se eliminan valores nulos o inválidos, haciendo que los datos sean más homogéneos y precisos.

	eventid	year	imonth	iday	approxdate	extended	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	crit1	crit2	crit3	doubtterr	multiple	success	suicide	attacktype1	attacktype1_txt
0	1	1970	7	2	NaN	0	58	Dominican Republic	2	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1	0	1	1	1	0	0	1	0	1	Assassination
1	2	1970	6	15	NaN	0	130	Mexico	1	North America	Federal	Mexico city	19.371887	-99.086624	1	0	1	1	1	0	0	1	0	6	Hostage Taking (Kidnapping)
2	3	1970	1	15	NaN	0	160	Philippines	5	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4	0	1	1	1	0	0	1	0	1	Assassination
3	4	1970	1	15	NaN	0	78	Greece	8	Western Europe	Attica	Athens	37.99749	23.762728	1	0	1	1	1	0	0	1	0	3	Bombing/Explosion
4	5	1970	1	15	NaN	0	101	Japan	4	East Asia	Fukouka	Fukouka	33.580412	130.396361	1	0	1	1	1	-9	0	1	0	7	Facility/Infrastructure Attack

## Imputacion de Valores y Reemplazo de tipos de datos

Después, se creó una fecha combinando los datos de las columnas iday, imonth e iyear, y se insertó en una nueva columna llamada approxdate. Finalmente, los datos de esta columna, originalmente de tipo objeto, fueron convertidos al formato de fecha, lo que facilita su gestión y análisis en el futuro, permitiendo una mejor manipulación de la información temporal en el conjunto de datos.

eventid	year	imonth	iday	approxdate	extended	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	crit1	crit2	crit3	doubtterr	multiple	success	suicide	attacktype1	attacktype1_txt	
0	1	1970	7	2	1970-07-02	0	58	Dominican Republic	2	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1	0	1	1	1	0	0	1	0	1	Assassination
1	2	1970	6	15	1970-06-15	0	130	Mexico	1	North America	Federal	Mexico city	19.371887	-99.086624	1	0	1	1	1	0	0	1	0	6	Hostage Taking (Kidnapping)
2	3	1970	1	15	1970-01-15	0	160	Philippines	5	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4	0	1	1	1	0	0	1	0	1	Assassination
3	4	1970	1	15	1970-01-15	0	78	Greece	8	Western Europe	Attica	Athens	37.99749	23.762728	1	0	1	1	1	0	0	1	0	3	Bombing/Explosion
4	5	1970	1	15	1970-01-15	0	101	Japan	4	East Asia	Fukuoka	Fukuoka	33.580412	130.396361	1	0	1	1	1	-9	0	1	0	7	Facility/Infrastructure Attack

Posteriormente, utilizando la información proporcionada en el Codebook de la Global Terrorism Database, se llevó a cabo la imputación, reemplazo y corrección de variables categóricas, en su mayoría numéricas.

eventid	year	imonth	iday	approxdate	extended	country	country_txt	region	region_txt	provstate	city	latitude	longitude	specificity	vicinity	crit1	crit2	crit3	doubtterr	multiple	success	suicide	attacktype1	attacktype1_txt	
0	1	1970	7	2	1970-07-02	0	58	Dominican Republic	2	Central America & Caribbean	NaN	Santo Domingo	18.456792	-69.951164	1	0	1	1	1	0	0	1	0	1	Assassination
1	2	1970	6	15	1970-06-15	0	130	Mexico	1	North America	Federal	Mexico city	19.371887	-99.086624	1	0	1	1	1	0	0	1	0	6	Hostage Taking (Kidnapping)
2	3	1970	1	15	1970-01-15	0	160	Philippines	5	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4	0	1	1	1	0	0	1	0	1	Assassination
3	4	1970	1	15	1970-01-15	0	78	Greece	8	Western Europe	Attica	Athens	37.997490	23.762728	1	0	1	1	1	0	0	1	0	3	Bombing/Explosion
4	5	1970	1	15	1970-01-15	0	101	Japan	4	East Asia	Fukuoka	Fukuoka	33.580412	130.396361	1	0	1	1	1	-1	0	1	0	7	Facility/Infrastructure Attack

Se utilizó el comando dropna para eliminar todas aquellas entradas vacías en las filas restantes que se consideraban contraproducentes para realizar imputación, así como aquellas que eran de tipo objeto cuya información no podía obtenerse de otra manera. Este paso garantizó que el conjunto de datos estuviera limpio y libre de valores ausentes que pudieran afectar la calidad del análisis posterior.

<pre>df1 = df1.dropna() df1.head()</pre>																									Python
1	2	1970	6	15	1970-06-15	0	130	Mexico	1	North America	Federal	Mexico city	19.371887	-99.086624	1	0	1	1	1	0	0	1	0	6	Hostage Taking (Kidnapping)
2	3	1970	1	15	1970-01-15	0	160	Philippines	5	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	4	0	1	1	1	0	0	1	0	1	Assassination
3	4	1970	1	15	1970-01-15	0	78	Greece	8	Western Europe	Attica	Athens	37.997490	23.762728	1	0	1	1	1	0	0	1	0	3	Bombing/Explosion
4	5	1970	1	15	1970-01-15	0	101	Japan	4	East Asia	Fukuoka	Fukuoka	33.580412	130.396361	1	0	1	1	1	-1	0	1	0	7	Facility/Infrastructure Attack
6	6	1970	1	2	1970-01-02	0	218	Uruguay	3	South America	Montevideo	Montevideo	-34.891151	-56.187214	1	0	1	1	1	0	0	0	0	1	Assassination

## Otros Ajustes

Para agilizar el proceso de manejo y análisis de datos, se realizaron varios ajustes adicionales en la base de datos. Estos incluyen la traducción de columnas y variables del español al inglés, con la excepción de las variables ciudad y grupo perpetrador. Además, se añadió una columna que especifica los datos geográficos en formato de texto, dado que anteriormente solo se disponía de una categoría por identificación. Por último, se modificaron algunos datos categóricos para que coincidieran con los del Codebook, lo que mejora la coherencia y la integridad de la información en el conjunto de datos.

Página 1 de 1																					
	Id	Año	Mes	Día	Fecha	Duracion_Mayor_1	Pais_Id	Pais	Region_Id	Region	Provincia/Estado	Ciudad	Latitud	Longitud	Especificidad_Geografica_Id	Cercania_Ciudad	Criterio_1	Criterio_2	Criterio_3	Dudas_Terrorismo	Ataque_Mul
1	2	1970	6	15	1970-06-15	0	130	México	1	América del Norte	Federal	Mexico city	19.371887	-99.086624	1	0	1	1	1	0	
2	3	1970	1	15	1970-01-15	0	160	Filipinas	5	Asia Sudeste	Tarlac	Desconocido	15.478598	120.599741	4	0	1	1	1	0	
3	4	1970	1	15	1970-01-15	0	78	Grecia	8	Europa Occidental	Attica	Athens	37.997490	23.762728	1	0	1	1	1	0	
4	5	1970	1	15	1970-01-15	0	101	Japón	4	Asia Oriental	Fukuoka	Fukuoka	33.580412	130.396361	1	0	1	1	1	0	
6	6	1970	1	2	1970-01-02	0	218	Uruguay	3	América del Sur	Montevideo	Montevideo	-34.891151	-56.187214	1	0	1	1	1	0	
<div><div></div></div>																					

# **RESULTADOS FINALES BASE DE DATOS**

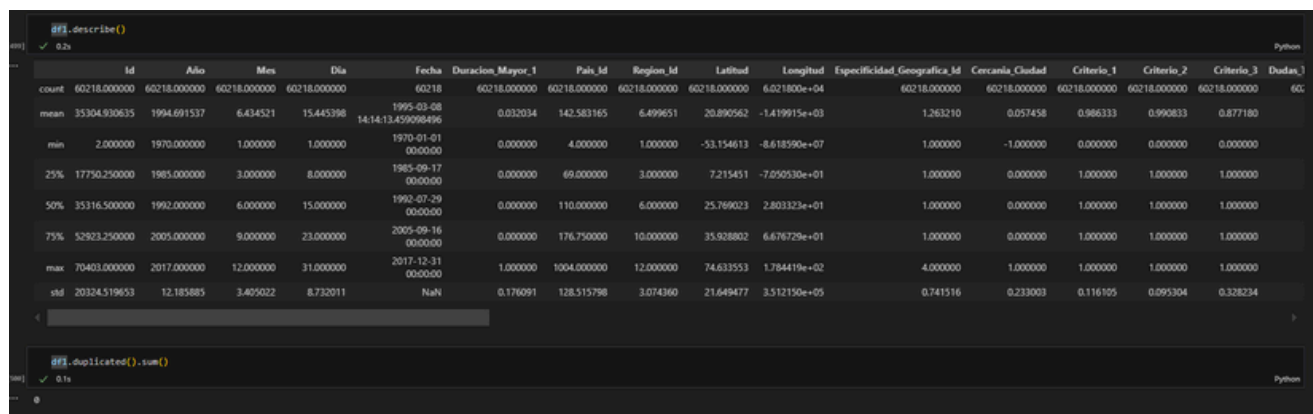


## Conclusión

La base de datos final cuenta con un total de 60,218 filas de información distribuidas en 52 columnas, abarcando diversos ataques terroristas ocurridos entre los años 1970 y 2017, con la excepción del año 1993. Cada columna ha sido estructurada con un tipo de dato adecuado a la naturaleza de la información que contiene, lo que garantiza la integridad y coherencia del conjunto de datos. A continuación se presenta un resumen de las características de la base de datos:

- Dimensiones de la base de datos: 60,218 entradas, 52 columnas.
- Tipos de datos:
  - a. Enteros (int32): 34 columnas.
  - b. Flotantes (float64): 2 columnas.
  - c. Enteros largos (int64): 1 columna.
  - d. Fechas (datetime64[ns]): 1 columna.
  - e. Objetos (object): 14 columnas.

Además, la base de datos ha sido depurada para eliminar filas duplicadas y valores nulos, asegurando así la calidad y fiabilidad de la información para análisis posteriores. Este conjunto de datos proporciona una base sólida para realizar estudios sobre patrones y tendencias en ataques terroristas a nivel global, facilitando el desarrollo de análisis más profundos y significativos en el ámbito de la seguridad y el estudio del terrorismo.



The screenshot displays the output of two pandas operations on a dataset. The first operation, `df.describe()`, provides a summary of the data across 52 columns. The second operation, `df.duplicated().sum()`, shows that there are no duplicate rows in the dataset.

	Id	Año	Mes	Día	Fecha	Duracion_Mayor_1	País_Id	Region_Id	Latitud	Longitud	Especificidad_Geografica_Id	Cercania_Ciudad	Criterio_1	Criterio_2	Criterio_3	Dudas_1
count	60218.000000	60218.000000	60218.000000	60218.000000	60218	60218.000000	60218.000000	60218.000000	60218.000000	6021800e+04	60218.000000	60218.000000	60218.000000	60218.000000	60218.000000	60218.000000
mean	35304.930635	1994.691537	6.434521	15.445398	1995-03-08 14:14:13.459098496	0.032034	142.583165	6.499651	20.890562	-14.19915e+03	1.263210	0.057458	0.986333	0.990833	0.877180	0.000000
min	2.000000	1970.000000	1.000000	1.000000	1970-01-01 00:00:00	0.000000	4.000000	1.000000	-53.154613	-84.18590e+07	1.000000	-1.000000	0.000000	0.000000	0.000000	0.000000
25%	17750.250000	1985.000000	3.000000	8.000000	1985-09-17 00:00:00	0.000000	69.000000	3.000000	7.215451	-7.250530e+01	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000
50%	35316.500000	1992.000000	6.000000	15.000000	1992-07-29 00:00:00	0.000000	110.000000	6.000000	25.789023	2.803323e+01	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000
75%	52923.250000	2005.000000	9.000000	23.000000	2005-09-16 00:00:00	0.000000	176.750000	10.000000	35.928802	6.676729e+01	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000
max	70403.000000	2017.000000	12.000000	31.000000	2017-12-31 00:00:00	1.000000	1004.000000	12.000000	74.633553	1.784419e+02	4.000000	1.000000	1.000000	1.000000	1.000000	1.000000
std	20324.519653	12.185885	3.405022	8.732011	NaN	0.176091	128.515798	3.074360	21.649477	3.512150e+05	0.741516	0.233003	0.116105	0.095304	0.328234	0.000000

The second operation, `df.duplicated().sum()`, returns 0, indicating that there are no duplicate rows in the dataset.

```
df.isnull().sum()
[0]
```

```
Id
Año
Mes
Día
Fecha
Duracion_Mayor_1
País_Id
País
Region_Id
Region
Provincia/Estado
Ciudad
Latitud
Longitud
Especificidad_Geografica_Id
Cercania_Ciudad
Criterio_1
Criterio_2
Criterio_3
Dudas_Terrorismo
Ataque_Multiple
Ataque_Exito
Ataque_Suicida
Tipo_Ataque_Id
Tipo_Ataque
...
Ideologicamente_Internacional
Miscelaneamente_Internacional
Ataque_Internacional
Especificidad_Geografica
dtype: int64
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings.
```

```
totalInvalid = 0
for i in df.columns:
    totalInvalid += (df[i] == 'invalid').sum()
print(f"invalid: {totalInvalid}")
[0]
```

```
invalid: 0
```

```
totalInvalid = 0
for i in df.columns:
    totalInvalid += (df[i] == '-99.0').sum()
print(f"-99.0: {totalInvalid}")
[0]
```

```
-99.0: 0
```

```
nfilas = df.shape[0]
resultados = []
for col in df.columns:
    valores = df[col].isnull().sum()
    porcentaje = (valores / nfilas) * 100
    resultados.append([col, valores, porcentaje])

df_fal = pd.DataFrame(resultados, columns=['Columna', 'Valores Faltantes', 'Porcentaje Faltante'])

df_fal_horiz = df_fal.transpose()
df_fal_horiz.columns = df_fal_horiz.iloc[0]
df_fal_horiz = df_fal_horiz[1:]

df_fal_horiz
[0]
```

Columna	Id	Año	Mes	Día	Fecha	Duracion_Mayor_1	País_Id	País	Region_Id	Region	Provincia/Estado	Ciudad	Latitud	Longitud	Especificidad_Geografica_Id	Cercania_Ciudad	Criterio_1	Criterio_2	Criterio_3	Dudas_Terrorismo	Ataque_Multiple	A
Valores Faltantes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Porcentaje Faltante	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

```
df.info()
[1]
```

```
24 Tipo_Ataque      60218 non-null object
25 Objetivo_Ataque_Id  60218 non-null int32
26 Objetivo_Ataque    60218 non-null object
27 Objetivo_Ataque_Subtipo_Id  60218 non-null int32
28 Objetivo_Ataque_Subtipo  60218 non-null object
29 Nacionalidad_Objeto  60218 non-null int32
30 Nacionalidad_Objeto  60218 non-null object
31 Perpetrador        60218 non-null object
32 Incediombre_Perpetrador  60218 non-null int32
33 Lobo_Solitario      60218 non-null int32
34 Numero_Atacantes    60218 non-null int32
35 Tipo_Arma_Id        60218 non-null int32
36 Tipo_Arma           60218 non-null object
37 Subtipo_Arma_Id     60218 non-null int32
38 Subtipo_Arma        60218 non-null object
39 Muertos             60218 non-null int32
40 Heridos             60218 non-null int32
41 Dato_Propiedad       60218 non-null int32
42 Grado_Daños_Propiedad_Id  60218 non-null int32
43 Grado_Daños_Propiedad  60218 non-null object
44 Secuestro/Rehenes    60218 non-null int32
45 Rescate             60218 non-null int32
46 Base_Datos          60218 non-null object
47 Logísticamente_Internacional  60218 non-null int32
48 Ideologicamente_Internacional  60218 non-null int32
49 Miscelaneamente_Internacional  60218 non-null int32
50 Ataque_Internacional  60218 non-null int32
51 Especificidad_Geografica  60218 non-null object
dtypes: datetime64[ns](1), float64(2), int32(14), int64(1), object(14)
memory usage: 16.5+ MB
```

df1.head()

Python

df1

	Id	Año	Mes	Día	Fecha	Duracion_Mayor_1	País_Id	País	Region_Id	Region	Provincia/Estado	Ciudad	Latitud	Longitud	Especificidad_Geografica_Id	Cercania_Ciudad	Criterio_1	Criterio_2	Criterio_3	Dudas_Terrorismo	Ataque_Mul
1	2	1970	6	15	1970-06-15	0	130	México	1	América del Norte	Federal	Mexico city	19.371887	-99.086624		1	0	1	1	1	0
2	3	1970	1	15	1970-01-15	0	160	Filipinas	5	Asia Sudeste	Tarlac	Desconocido	15.478598	120.599741		4	0	1	1	1	0
3	4	1970	1	15	1970-01-15	0	78	Grecia	8	Europa Occidental	Attica	Athens	37.997490	23.762728		1	0	1	1	1	0
4	5	1970	1	15	1970-01-15	0	101	Japón	4	Asia Oriental	Fukouka	Fukouka	33.580412	130.396361		1	0	1	1	1	0
6	6	1970	1	2	1970-01-02	0	218	Uruguay	3	América del Sur	Montevideo	Montevideo	-34.891151	-56.187214		1	0	1	1	1	0

<div>

df1.to\_csv('Base\_1emple.csv', index=False)

Python

df1

df1.shape

Python

df1

(60218, 52)