Challenge Básico: Introducción a PySpark y Koalas

PÉREZ ROSAS LUIS ALFREDO MCD UDG 2025-B GRUPO 2

Introducción:

Este challenge introduce a los estudiantes a la visualización de grandes bases de datos utilizando PySpark y Koalas, un API que facilita el trabajo con grandes volúmenes de datos en un entorno de Spark pero con sintaxis similar a pandas.

Objetivo general:

Familiarizarse con PySpark y la API de Koalas para la manipulación y visualización de datos a gran escala.

Objetivo específico:

Descargar una base de datos desde Kaggle, importarla a Databricks, y realizar un análisis exploratorio utilizando la API pyspark.pandas.

In [1]:

```
# 1. Importación de Bibliotecas

import findspark
findspark.init()

import pandas as pd
import pyspark

import pandas as pd
import numpy as np
import pyspark.pandas as ps
from pyspark.sql import SparkSession
import matplotlib as plt
import seaborn as sns
```

```
c:\Users\alfre\anaconda3\envs\ProcessBigData25B\lib\site-packages\pyspark\pandas\__init__.py:50: UserWarning: 'PYARRO
W_IGNORE_TIMEZONE' environment variable was not set. It is required to set this environment variable to '1' in both d
river and executor sides if you use pyarrow>=2.0.0. pandas-on-Spark will set it for you but it does not work if there
is a Spark context already launched.
  warnings.warn(
```

In [2]:

```python
from pyspark.sql import SparkSession

spark = SparkSession.builder\
        .master("local[*]")\
        .appName('PySpark_Challenge1')\
        .getOrCreate()
```

Base de datos:

Base de datos: Global Terrorism Database

Plataforma: Kaggle

Instrucciones: Los estudiantes deberán descargar los datos y cargarlos

Apache Spark

In [3]:

```python
gt = spark.read.csv('data/globalterrorismdb_0718dist.csv',
                       inferSchema = True,
                       header = True)
```

In [4]:

```python
gt.show(15)
```

```
+------------+-----+------+----+----------+--------+----------+-------+------------------+------+--------------------
+-----------+-------------+----------+-----------+-----------+--------+--------------------+--------------------+----
-+-----+-----+---------+-----------+--------------------+--------+-------+-------+-----------+--------------------+--
---------+---------------+-----------+---------------+---------+--------------------+------------+-------------------
-+--------------------+--------------------+-------+------------------+---------+-------------+------------+---------
-------+-----+-------+-------+-----------+---------+-------------+------------+----------------+-----+-------+-------
+-----------+--------------------+--------+------+---------+------+---------+--------------------+-----------+-------
----+-----------+----------+------+--------+-------+---------+-------------+------+----------+--------------+------+-
---------+--------------+---------+---------+-------------+------------+--------------------+---------+-------------+
------------+----------------+---------+-------------+------------+----------------+---------+-------------+---------
---+----------------+--------------------+-----+-------+--------+------+--------+--------+--------+----------+-------
-------------+---------+--------------------+---------+--------+----------+------+-----+------+-------------+------+-
--------+-----------+----------+------------+----------+--------------+--------------------+---------+---------------
-----+--------------------+--------------------+--------------------+--------------------+--------------------+------
--------------+--------------------+--------------------+--------------------+
|     eventid|iyear|imonth|iday|approxdate|extended|resolution|country|       country_txt|region|          region_txt
|  provstate|         city|  latitude|  longitude|specificity|vicinity|            location|             summary|crit
1|crit2|crit3|doubtterr|alternative|     alternative_txt|multiple|success|suicide|attacktype1|     attacktype1_txt|at
tacktype2|attacktype2_txt|attacktype3|attacktype3_txt|targtype1|       targtype1_txt|targsubtype1|    targsubtype1_tx
t|               corp1|             target1|natlty1|       natlty1_txt|targtype2|targtype2_txt|targsubtype2|targsubty
pe2_txt|corp2|target2|natlty2|natlty2_txt|targtype3|targtype3_txt|targsubtype3|targsubtype3_txt|corp3|target3|natlty3
|natlty3_txt|               gname|gsubname|gname2|gsubname2|gname3|gsubname3|              motive|guncertain1|guncert
ain2|guncertain3|individual|nperps|nperpcap|claimed|claimmode|claimmode_txt|claim2|claimmode2|claimmode2_txt|claim3|c
laimmode3|claimmode3_txt|compclaim|weaptype1|weaptype1_txt|weapsubtype1|    weapsubtype1_txt|weaptype2|weaptype2_txt|
weapsubtype2|weapsubtype2_txt|weaptype3|weaptype3_txt|weapsubtype3|weapsubtype3_txt|weaptype4|weaptype4_txt|weapsubty
pe4|weapsubtype4_txt|          weapdetail|nkill|nkillus|nkillter|nwound|nwoundus|nwoundte|property|propextent|      p
ropextent_txt|propvalue|         propcomment|ishostkid|nhostkid|nhostkidus|nhours|ndays|divert|kidhijcountry|ransom|r
ansomamt|ransomamtus|ransompaid|ransompaidus|ransomnote|hostkidoutcome|  hostkidoutcome_txt|nreleased|            add
notes|              scite1|              scite2|              scite3|            dbsource|             INT_LOG|
INT_IDEO|            INT_MISC|             INT_ANY|             related|
+------------+-----+------+----+----------+--------+----------+-------+------------------+------+--------------------
+-----------+-------------+----------+-----------+-----------+--------+--------------------+--------------------+----
-+-----+-----+---------+-----------+--------------------+--------+-------+-------+-----------+--------------------+--
---------+---------------+-----------+---------------+---------+--------------------+------------+-------------------
-+--------------------+--------------------+-------+------------------+---------+-------------+------------+---------
-------+-----+-------+-------+-----------+---------+-------------+------------+----------------+-----+-------+-------
+-----------+--------------------+--------+------+---------+------+---------+--------------------+-----------+-------
----+-----------+----------+------+--------+-------+---------+-------------+------+----------+--------------+------+-
---------+--------------+---------+---------+-------------+------------+--------------------+---------+-------------+
------------+----------------+---------+-------------+------------+----------------+---------+-------------+---------
---+----------------+--------------------+-----+-------+--------+------+--------+--------+--------+----------+-------
-------------+---------+--------------------+---------+--------+----------+------+-----+------+-------------+------+-
```

```
--------+-----------+----------+------------+----------+--------------+--------------------+---------+---------------
-----+--------------------+--------------------+--------------------+--------------------+--------------------+------
--------------+--------------------+--------------------+--------------------+
|197000000001| 1970|     7|   2|      NULL|       0|      NULL|     58|Dominican Republic|     2|Central America &...
|       NULL|Santo Domingo| 18.456792| -69.951164|          1|       0|                NULL|                NULL|
1|    1|    1|        0|       NULL|                NULL|       0|      1|      0|          1|       Assassination|
NULL|           NULL|       NULL|           NULL|       14|Private Citizens ...|          68|      Named Civilian|
NULL|        Julio Guzman|     58|Dominican Republic|     NULL|         NULL|        NULL|            NULL| NULL|   N
ULL|   NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|       NULL|
MANO-D|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|          0|       NULL|       NULL|         0
|  NULL|    NULL|   NULL|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|          NULL|
NULL|       13|      Unknown|        NULL|                NULL|     NULL|         NULL|        NULL|            NULL|
NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|
NULL|    1|   NULL|    NULL|     0|    NULL|    NULL|       0|      NULL|                NULL|     NULL|
NULL|        0|    NULL|      NULL|  NULL| NULL|  NULL|         NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|                NULL|                NULL|
NULL|                NULL|                PGIS|                   0|                   0|                   0|
0|                NULL|
|197000000002| 1970|     0|   0|      NULL|       0|      NULL|    130|            Mexico|     1|       North America
|    Federal|  Mexico city| 19.371887| -99.086624|          1|       0|                NULL|                NULL|
1|    1|    1|        0|       NULL|                NULL|       0|      1|      0|          6|Hostage Taking (K...|
NULL|           NULL|       NULL|           NULL|        7|Government (Diplo...|          45|Diplomatic Person...|Bel
gian Ambassado...|Nadine Chaval, da...|     21|           Belgium|     NULL|         NULL|        NULL|            NU
LL| NULL|   NULL|   NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|
NULL|23rd of September...|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|          0|       NULL|
NULL|         0|     7|    NULL|   NULL|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|
NULL|     NULL|       13|      Unknown|        NULL|                NULL|     NULL|         NULL|        NULL|
NULL|     NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|
NULL|    0|   NULL|    NULL|     0|    NULL|    NULL|       0|      NULL|                NULL|     NULL|
NULL|        1|       1|         0|  NULL| NULL|  NULL|       Mexico|     1|   800000|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|                NULL|                NULL|
NULL|                NULL|                PGIS|                   0|                   1|                   1|
1|                NULL|
|197001000001| 1970|     1|   0|      NULL|       0|      NULL|    160|       Philippines|     5|      Southeast Asia
|     Tarlac|      Unknown| 15.478598| 120.599741|          4|       0|                NULL|                NULL|
1|    1|    1|        0|       NULL|                NULL|       0|      1|      0|          1|       Assassination|
NULL|           NULL|       NULL|           NULL|       10| Journalists & Media|          54|Radio Journalist/...|
Voice of America|            Employee|    217|     United States|     NULL|         NULL|        NULL|            NUL
L| NULL|   NULL|   NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|
NULL|             Unknown|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|          0|       NULL|
NULL|         0|  NULL|    NULL|   NULL|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|
NULL|     NULL|       13|      Unknown|        NULL|                NULL|     NULL|         NULL|        NULL|
NULL|     NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|
NULL|    1|   NULL|    NULL|     0|    NULL|    NULL|       0|      NULL|                NULL|     NULL|
NULL|       0|    NULL|      NULL|  NULL| NULL|  NULL|          NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|                NULL|                NULL|
NULL|                NULL|               PGIS|                 -9|                 -9|                  1|
1|                NULL|
|197001000002| 1970|     1|   0|      NULL|       0|      NULL|     78|            Greece|     8|      Western Europe
|     Attica|       Athens|  37.99749|  23.762728|          1|       0|                NULL|                NULL|
1|    1|    1|        0|       NULL|                NULL|       0|      1|      0|          3|   Bombing/Explosion|
NULL|          NULL|       NULL|           NULL|        7|Government (Diplo...|         46|   Embassy/Consulate|
NULL|        U.S. Embassy|    217|     United States|     NULL|          NULL|         NULL|           NULL| NULL|   N
ULL|   NULL|       NULL|     NULL|         NULL|        NULL|             NULL| NULL|   NULL|   NULL|       NULL|
Unknown|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|         0|       NULL|       NULL|
0|  NULL|    NULL|   NULL|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|          NULL|
NULL|       6|   Explosives|          16|Unknown Explosive...|     NULL|          NULL|        NULL|           NULL|
NULL|        NULL|        NULL|           NULL|     NULL|          NULL|         NULL|            NULL|           Exp
losive| NULL|   NULL|    NULL|  NULL|    NULL|    NULL|       1|      NULL|                NULL|     NULL|
NULL|       0|    NULL|      NULL|  NULL| NULL|  NULL|          NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|                NULL|                NULL|
NULL|                NULL|               PGIS|                 -9|                 -9|                  1|
1|                NULL|
|197001000003| 1970|     1|   0|      NULL|       0|      NULL|    101|             Japan|     4|           East Asia
|    Fukouka|      Fukouka| 33.580412| 130.396361|          1|       0|                NULL|                NULL|
1|    1|    1|       -9|       NULL|                NULL|       0|      1|      0|          7|Facility/Infrastr...|
NULL|          NULL|       NULL|           NULL|        7|Government (Diplo...|         46|   Embassy/Consulate|
NULL|      U.S. Consulate|    217|     United States|     NULL|          NULL|         NULL|           NULL| NULL|   N
ULL|   NULL|       NULL|     NULL|         NULL|        NULL|             NULL| NULL|   NULL|   NULL|       NULL|
Unknown|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|         0|       NULL|       NULL|
0|  NULL|    NULL|   NULL|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|          NULL|
NULL|       8|   Incendiary|       NULL|                NULL|     NULL|          NULL|        NULL|           NULL|
NULL|        NULL|        NULL|           NULL|     NULL|          NULL|         NULL|            NULL|           Ince
ndiary| NULL|   NULL|    NULL|  NULL|    NULL|    NULL|       1|      NULL|                NULL|     NULL|
NULL|       0|    NULL|      NULL|  NULL| NULL|  NULL|          NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|                NULL|                NULL|
NULL|                NULL|               PGIS|                 -9|                 -9|                  1|
1|                NULL|
|197001010002| 1970|     1|   1|      NULL|       0|      NULL|    217|     United States|     1|       North America
|   Illinois|        Cairo| 37.005105| -89.176269|          1|       0|                NULL|1/1/1970: Unknown...|
1|    1|    1|        0|       NULL|                NULL|       0|      1|      0|          2|       Armed Assault|
NULL|          NULL|       NULL|           NULL|        3|              Police|         22|Police Building (...|Cai
ro Police Depa...|Cairo Police Head...|    217|     United States|     NULL|          NULL|         NULL|           NU
LL| NULL|    NULL|    NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|
NULL|  Black Nationalists|    NULL|  NULL|     NULL|  NULL|     NULL|To protest the Ca...|          0|       NULL|
NULL|         0|   -99|     -99|      0|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|
NULL|     NULL|        5|     Firearms|           5|    Unknown Gun Type|     NULL|         NULL|        NULL|
NULL|     NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|Seve
ral gunshots ...|    0|      0|       0|     0|       0|       0|       1|         3|Minor (likely < $...|     NULL|
NULL|        0|    NULL|      NULL|  NULL| NULL|  NULL|         NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|The Cairo Chief o...|"""Police Chief Q...|  "" Washingt
on Post|           January 2|              1970."|"""Cairo Police C...|    "" Afro-American|          January 10|
1970."|"Christopher Hewi...|
|197001020001| 1970|     1|   2|      NULL|       0|      NULL|    218|           Uruguay|     3|       South America
| Montevideo|   Montevideo|-34.891151| -56.187214|          1|       0|                NULL|                NULL|
1|    1|    1|        0|       NULL|                NULL|       0|      0|      0|          1|       Assassination|
NULL|           NULL|       NULL|           NULL|        3|              Police|          25|Police Security F...|
Uruguayan Police|Juan Maria de Luc...|    218|           Uruguay|     NULL|         NULL|        NULL|            NUL
L| NULL|   NULL|   NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|
NULL| Tupamaros (Uruguay)|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|          0|       NULL|
NULL|         0|     3|    NULL|   NULL|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|
NULL|     NULL|        5|     Firearms|           2|Automatic or Semi...|     NULL|         NULL|        NULL|
NULL|     NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|   A
utomatic firearm|    0|   NULL|    NULL|     0|    NULL|    NULL|       0|      NULL|                NULL|     NULL|
NULL|        0|    NULL|      NULL|  NULL| NULL|  NULL|         NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|                NULL|                NULL|
NULL|                NULL|                PGIS|                   0|                   0|                   0|
0|                NULL|
|197001020002| 1970|     1|   2|      NULL|       0|      NULL|    217|     United States|     1|       North America
| California|      Oakland| 37.791927|-122.225906|          1|       0|     Edes Substation|1/2/1970: Unknown...|
1|    1|    1|        1|          2|    Other Crime Type|       0|      1|      0|          3|   Bombing/Explosion|
NULL|           NULL|       NULL|           NULL|       21|           Utilities|         107|         Electricity|Pac
ific Gas & Ele...|     Edes Substation|    217|     United States|     NULL|         NULL|        NULL|            NU
LL| NULL|   NULL|   NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|
NULL|             Unknown|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|          0|       NULL|
NULL|         0|   -99|     -99|      0|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|
NULL|     NULL|        6|   Explosives|          16|Unknown Explosive...|     NULL|         NULL|        NULL|
NULL|     NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|
NULL|    0|      0|       0|     0|       0|       0|       1|         3|Minor (likely < $...|    22500|Three transfo
rmer...|        0|    NULL|      NULL|  NULL| NULL|  NULL|         NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|Damages were esti...|"Committee on Gov...|
Civil| and Criminal Dis...|"" U.S. Governmen...|            August 6|              1970."|"Christopher Hewi...|"" Pra
eger Securi...|              2005."|
|197001020003| 1970|     1|   2|      NULL|       0|      NULL|    217|     United States|     1|       North America
|  Wisconsin|      Madison| 43.076592| -89.412488|          1|       0|                NULL|1/2/1970: Karl Ar...|
1|    1|    1|        0|       NULL|                NULL|       0|      1|      0|          7|Facility/Infrastr...|
NULL|           NULL|       NULL|           NULL|        4|           Military|          28|Military Recruiti...|
R.O.T.C.|R.O.T.C. offices ...|    217|     United States|     NULL|         NULL|        NULL|            NULL| NULL|
NULL|   NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|       NULL|
New Year's Gang|    NULL|  NULL|     NULL|  NULL|     NULL|To protest the Wa...|          0|       NULL|       NULL|
0|     1|       1|      1|        1|       Letter|  NULL|      NULL|          NULL|  NULL|      NULL|          NULL|
NULL|        8|   Incendiary|          19|Molotov Cocktail/...|     NULL|         NULL|        NULL|            NULL|
NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|Firebomb consi
sti...|    0|      0|       0|     0|       0|       0|       1|         3|Minor (likely < $...|    60000|Basketball
courts...|        0|    NULL|      NULL|  NULL| NULL|  NULL|         NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|The New Years Gan...|"Tom Bates, ""Rad...|"" HarperColl
insP...|              1992."|"David Newman, Sa...| Heard and Won't ...|     "" Mother Jones|      February-March|
1979."|"The Wisconsin Ca...|
|197001030001| 1970|     1|   3|      NULL|       0|      NULL|    217|     United States|     1|       North America
|  Wisconsin|      Madison|  43.07295| -89.386694|          1|       0|               NULL|1/3/1970: Karl Ar...|
1|    1|    1|        0|       NULL|                NULL|       0|      1|      0|          7|Facility/Infrastr...|
NULL|           NULL|       NULL|           NULL|        2|Government (General)|          21|Government Buildi...|
Selective Service|Selective Service...|    217|     United States|     NULL|         NULL|        NULL|            NU
LL| NULL|   NULL|   NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|
NULL|     New Year's Gang|    NULL|  NULL|     NULL|  NULL|     NULL|To protest the Wa...|          0|       NULL|
NULL|         0|     1|       1|      0|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|
NULL|     NULL|        8|   Incendiary|          20| Gasoline or Alcohol|     NULL|         NULL|        NULL|
NULL|     NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|Pour
ed gasoline o...|    0|      0|       0|     0|       0|       0|       1|         3|Minor (likely < $...|     NULL|
Slight damage|        0|    NULL|      NULL|  NULL| NULL|  NULL|         NULL|     0|     NULL|       NULL|      NULL
|        NULL|      NULL|          NULL|                NULL|     NULL|Karl Armstrong's ...|"Committee on Gov...|
Civil| and Criminal Dis...|"" U.S. Governmen...|            August 6|              1970."|"Tom Bates, ""Rad...|"" Har
perCollinsP...|              1992."|
|197001050001| 1970|     1|   1|      NULL|       0|      NULL|    217|     United States|     1|       North America
|  Wisconsin|      Baraboo|   43.4685| -89.744299|          1|       0|               NULL|               NULL|
1|    1|    0|        1|          1|Insurgency/Gueril...|       0|      0|      0|          3|   Bombing/Explosion|
NULL|           NULL|       NULL|           NULL|        4|           Military|          27|Military Barracks...|
NULL|Badger Army ammo ...|    217|     United States|     NULL|         NULL|        NULL|            NULL| NULL|   N
ULL|   NULL|       NULL|     NULL|         NULL|        NULL|            NULL| NULL|   NULL|   NULL|       NULL|Weath
er Undergrou...|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|          0|       NULL|       NULL|
0|  NULL|    NULL|   NULL|     NULL|         NULL|  NULL|      NULL|          NULL|  NULL|      NULL|          NULL|
NULL|        6|   Explosives|          16|Unknown Explosive...|     NULL|         NULL|        NULL|            NULL|
NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|           Exp
losive|    0|   NULL|    NULL|     0|    NULL|    NULL|       0|         3|Minor (likely < $...|        0|
NULL|        0|    NULL|      NULL|  NULL| NULL|  NULL|         NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|          NULL|                NULL|     NULL|                NULL|                NULL|
NULL|                NULL|                PGIS|                   0|                   0|                   0|
0|                NULL|
|197001060001| 1970|     1|   6|      NULL|       0|      NULL|    217|     United States|     1|       North America
|   Colorado|       Denver| 39.758968|-104.876305|          1|       0|                NULL|1/6/1970: Unknown...|
1|    1|    1|        1|          2|    Other Crime Type|       0|      1|      0|          7|Facility/Infrastr...|
NULL|           NULL|       NULL|           NULL|        4|            Military|          28|Military Recruiti...|Arm
y Recruiting S...|Army Recruiting S...|    217|     United States|     NULL|          NULL|        NULL|            NU
LL| NULL|   NULL|   NULL|       NULL|     NULL|         NULL|         NULL|                NULL| NULL|   NULL|   NULL|
NULL| Left-Wing Militants|    NULL|  NULL|     NULL|  NULL|     NULL|Protest the draft...|          0|       NULL|
NULL|         0|   -99|     -99|      0|     NULL|         NULL|  NULL|      NULL|           NULL|  NULL|      NULL|
NULL|     NULL|        8|   Incendiary|          19|Molotov Cocktail/...|     NULL|          NULL|        NULL|
NULL|     NULL|         NULL|        NULL|            NULL|     NULL|          NULL|        NULL|              NULL|
Molotov cocktail|    0|      0|       0|     0|       0|       0|       1|          3|Minor (likely < $...|      305|
NULL|        0|    NULL|      NULL|  NULL| NULL|  NULL|          NULL|     0|     NULL|       NULL|      NULL|
NULL|      NULL|           NULL|                 NULL|     NULL|                NULL|"Committee on Gov...|
Civil| and Criminal Dis...|"" U.S. Governmen...|           August 6|             1970."|"Christopher Hewi...|"" Pra
eger Securi...|              2005."|
|197001080001| 1970|     1|   8|      NULL|       0|      NULL|     98|             Italy|     8|      Western Europe
|      Lazio|         Rome| 41.890961|  12.490069|          1|       0|                NULL|                    NULL|
1|    1|    1|       -9|       NULL|                NULL|       0|      1|      0|          4|           Hijacking|
NULL|           NULL|       NULL|           NULL|        6| Airports & Aircraft|          42|Aircraft (not at ...| Tr
ans World Airline|Flight 802 Boeing...|    217|     United States|     NULL|          NULL|        NULL|            NU
LL| NULL|   NULL|   NULL|       NULL|     NULL|         NULL|         NULL|                NULL| NULL|   NULL|   NULL|
NULL|             Unknown|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|          0|       NULL|
NULL|         0|     1|    NULL|   NULL|     NULL|         NULL|  NULL|      NULL|           NULL|  NULL|      NULL|
NULL|     NULL|        5|     Firearms|           4|Rifle/Shotgun (no...|        5|     Firearms|           3|
Handgun|     NULL|         NULL|        NULL|            NULL|     NULL|         NULL|        NULL|            NULL|R
ifle - carbine; ...|    0|      0|       0|     0|       0|       0|       0|      NULL|                NULL|     NUL
L|                NULL|        0|    NULL|      NULL|  NULL| NULL|Beirut|       Beirut|     0|     NULL|        NULL|
NULL|        NULL|      NULL|            2|Hostage(s) releas...|     NULL|               NULL|                NULL|
NULL|                NULL|       Hijacking DB|                  -9|                  -9|                   1|
1|                NULL|
|197001090001| 1970|     1|   9|      NULL|       0|      NULL|    217|     United States|     1|       North America
|   Michigan|      Detroit| 42.331685| -83.047924|          1|       0|                NULL|1/9/1970: Unknown...|
1|    1|    1|        0|       NULL|                NULL|       0|      1|      0|          7|Facility/Infrastr...|
NULL|           NULL|       NULL|           NULL|        2|Government (General)|          21|Government Buildi...|U.
S. Government h...|Packard Propertie...|    217|     United States|     NULL|          NULL|        NULL|            N
ULL| NULL|   NULL|   NULL|       NULL|     NULL|         NULL|         NULL|                NULL| NULL|   NULL|   NULL|
NULL| Left-Wing Militants|    NULL|  NULL|     NULL|  NULL|     NULL|                NULL|          0|       NULL|
NULL|         0|   -99|     -99|      0|     NULL|         NULL|  NULL|      NULL|           NULL|  NULL|      NULL|
NULL|     NULL|        8|   Incendiary|          19|Molotov Cocktail/...|     NULL|          NULL|        NULL|
NULL|     NULL|         NULL|        NULL|            NULL|     NULL|          NULL|        NULL|              NULL|
Firebomb|    0|       0|        0|     0|       0|       0|       1|          3|Minor (likely < $...|     NULL|Building
was damaged|        0|    NULL|      NULL|  NULL| NULL|  NULL|          NULL|     0|     NULL|        NULL|      NULL|
NULL|      NULL|          NULL|                 NULL|     NULL|                NULL|"Committee on Gov...|
Civil| and Criminal Dis...|"" U.S. Governmen...|            August 6|               1970."|"Christopher Hewi...|"" Pra
eger Securi...|              2005."|
|197001090002| 1970|     1|   9|      NULL|       0|      NULL|    217|     United States|     1|        North America
|Puerto Rico|  Rio Piedras| 18.386932| -66.061127|          1|        0|Caparra Shopping ...|1/9/1970:  The Ar...|
1|    1|    1|        1|          2|    Other Crime Type|       0|      1|      0|         7|Facility/Infrastr...|
NULL|          NULL|       NULL|            NULL|       1|            Business|          7|Retail/Grocery/Ba...|Ame
rican owned bu...|       Baker's Store|    217|     United States|     NULL|          NULL|         NULL|              NU
LL| NULL|     NULL|    NULL|        NULL|      NULL|           NULL|          NULL|            NULL| NULL|     NULL|     NULL|
NULL|Armed Commandos o...|     NULL|  NULL|      NULL|  NULL|      NULL|To protest United...|          1|        NULL|
NULL|        0|   -99|     -99|      1|       10|      Unknown|  NULL|       NULL|          NULL|  NULL|       NULL|
NULL|     NULL|        8|   Incendiary|          18|          Arson/Fire|     NULL|          NULL|         NULL|
NULL|     NULL|        NULL|        NULL|             NULL|      NULL|          NULL|         NULL|              NULL|Fire
set in back ...|    0|       0|       0|     0|       0|       0|       1|          2|Major (likely >= ...|  2000000|
Store destroyed|        0|    NULL|       NULL|  NULL| NULL|  NULL|         NULL|     0|      NULL|         NULL|       NU
LL|       NULL|       NULL|           NULL|                    NULL|     NULL|The fire began at...|"Committee on the...|""
U.S. Governmen...|           August 6|              1970."|"""No Evidence Of...|"" The Virgin Isl...|          Janua
ry 13|             1970."|"""Toward People'...|
+------------+-----+------+----+----------+--------+----------+-------+------------------+------+---------------------
+-----------+-------------+----------+-----------+-----------+--------+--------------------+--------------------+-----
-+-----+-----+---------+-----------+--------------------+--------+-------+-------+-----------+--------------------+---
---------+---------------+-----------+---------------+---------+--------------------+------------+--------------------
-+--------------------+--------------------+-------+------------------+---------+-------------+------------+----------
-------+-----+-------+-------+-----------+-----------------------+------------+----------------+-----+-------+--------
+-----------+--------------------+--------+------+---------+------+---------+--------------------+-----------+--------
----+-----------+----------+-------+-------+---------+-------+-------------+------+--------------+----------+------+--
---------+--------------+---------+-----------------------+------------+--------------------+---------+--------------+
------------+----------------+---------+-------------+------------+----------------+---------+-------------+----------
---+----------------+--------------------+-----+-------+--------+------+--------+--------+--------+----------+--------
-------------+---------+--------------------+---------+--------+----------+------+-----+------+-------------+------+--
--------+-----------+----------+-----------------------+--------------+--------------------+---------+----------------
-----+--------------------+--------------------+--------------------+--------------------+--------------------+-------
--------------+--------------------+--------------------+--------------------+
only showing top 15 rows
```

In [5]:

```
# Convertir a pandas-on-Spark DataFrame

df = gt.toPandas()
```

In [6]:

```
# Mostrar las primeras filas del DataFrame
```

```
df.head()
```

Out[6]:

	eventid	iyear	imonth	iday	approxdate	extended	resolution	country	country_txt	region	...	addnotes	scite1	sc
0	197000000001	1970	7	2	None	0	None	58	Dominican Republic	2	...	None	None	N
1	197000000002	1970	0	0	None	0	None	130	Mexico	1	...	None	None	N
2	197001000001	1970	1	0	None	0	None	160	Philippines	5	...	None	None	N
3	197001000002	1970	1	0	None	0	None	78	Greece	8	...	None	None	N
4	197001000003	1970	1	0	None	0	None	101	Japan	4	...	None	None	N

5 rows × 135 columns

In [7]:

```
# Obtener los nombres de las columnas como una lista

column_headers_list = df.columns.tolist()
print(column_headers_list)
```

```
['eventid', 'iyear', 'imonth', 'iday', 'approxdate', 'extended', 'resolution', 'country', 'country_txt', 'region', 'r
egion_txt', 'provstate', 'city', 'latitude', 'longitude', 'specificity', 'vicinity', 'location', 'summary', 'crit1',
'crit2', 'crit3', 'doubtterr', 'alternative', 'alternative_txt', 'multiple', 'success', 'suicide', 'attacktype1', 'at
tacktype1_txt', 'attacktype2', 'attacktype2_txt', 'attacktype3', 'attacktype3_txt', 'targtype1', 'targtype1_txt', 'ta
rgsubtype1', 'targsubtype1_txt', 'corp1', 'target1', 'natlty1', 'natlty1_txt', 'targtype2', 'targtype2_txt', 'targsub
type2', 'targsubtype2_txt', 'corp2', 'target2', 'natlty2', 'natlty2_txt', 'targtype3', 'targtype3_txt', 'targsubtype
3', 'targsubtype3_txt', 'corp3', 'target3', 'natlty3', 'natlty3_txt', 'gname', 'gsubname', 'gname2', 'gsubname2', 'gn
ame3', 'gsubname3', 'motive', 'guncertain1', 'guncertain2', 'guncertain3', 'individual', 'nperps', 'nperpcap', 'claim
ed', 'claimmode', 'claimmode_txt', 'claim2', 'claimmode2', 'claimmode2_txt', 'claim3', 'claimmode3', 'claimmode3_tx
t', 'compclaim', 'weaptype1', 'weaptype1_txt', 'weapsubtype1', 'weapsubtype1_txt', 'weaptype2', 'weaptype2_txt', 'wea
psubtype2', 'weapsubtype2_txt', 'weaptype3', 'weaptype3_txt', 'weapsubtype3', 'weapsubtype3_txt', 'weaptype4', 'weapt
ype4_txt', 'weapsubtype4', 'weapsubtype4_txt', 'weapdetail', 'nkill', 'nkillus', 'nkillter', 'nwound', 'nwoundus', 'n
woundte', 'property', 'propextent', 'propextent_txt', 'propvalue', 'propcomment', 'ishostkid', 'nhostkid', 'nhostkidu
s', 'nhours', 'ndays', 'divert', 'kidhijcountry', 'ransom', 'ransomamt', 'ransomamtus', 'ransompaid', 'ransompaidus',
'ransomnote', 'hostkidoutcome', 'hostkidoutcome_txt', 'nreleased', 'addnotes', 'scite1', 'scite2', 'scite3', 'dbsourc
e', 'INT_LOG', 'INT_IDEO', 'INT_MISC', 'INT_ANY', 'related']
```

Limpieza de Datos

In [8]:

```
# Conversión de Columnas Numéricas
# Convertimos la columna 'nkill' (y otras numéricas si es necesario) a tipo numérico, manejando posibles errores de c

# Conversión de columnas a tipo numérico el prefijo "n" (nkill, nkillus, nkillter, nwound, nwoundus, nwoundte)

df['nkill'] = ps.to_numeric(df['nkill'], errors='coerce')
df['nkillus'] = ps.to_numeric(df['nkillus'], errors='coerce')
df['nkillter'] = ps.to_numeric(df['nkillter'], errors='coerce')
df['nwound'] = ps.to_numeric(df['nwound'], errors='coerce')
df['nwoundus'] = ps.to_numeric(df['nwoundus'], errors='coerce')
df['nwoundte'] = ps.to_numeric(df['nwoundte'], errors='coerce')
```

In [9]:

```
# Verificación de Valores Nulos
# Identificamos columnas con valores nulos y calculamos su porcentaje.

null_counts = df.isnull().sum()
null_percentage = (null_counts / df.shape[0]) * 100
nulls = ps.DataFrame({'nulos': null_counts, 'porcentaje': null_percentage})
nulls[nulls['nulos'] > 0]
```

Out[9]:

	nulos	porcentaje
approxdate	172452	94.914993
resolution	179471	98.778145
provstate	421	0.231712
city	434	0.238867
latitude	4556	2.507554
longitude	4557	2.508104
specificity	6	0.003302
location	126196	69.456385
summary	66129	36.396409
doubtterr	1	0.000550
alternative	152241	83.791162
alternative_txt	152414	83.886379
multiple	297	0.163464
success	178	0.097969
suicide	85	0.046783
attacktype1	54	0.029721
attacktype1_txt	35	0.019263
attacktype2	174947	96.288204
attacktype2_txt	175111	96.378467
attacktype3	181084	99.665916
attacktype3_txt	181154	99.704443
targtype1	419	0.230611

	nulos	porcentaje
targtype1_txt	263	0.144751
targsubtype1	10503	5.780694
targsubtype1_txt	10451	5.752074
corp1	42517	23.400719
target1	682	0.375363
natlty1	1600	0.880616
natlty1_txt	1584	0.871810
targtype2	170117	93.629844
targtype2_txt	170289	93.724510
targsubtype2	170853	94.034927
targsubtype2_txt	170911	94.066850
corp2	171508	94.395430
target2	170633	93.913843
natlty2	170845	94.030524
natlty2_txt	170853	94.034927
targtype3	180453	99.318623
targtype3_txt	180472	99.329081
targsubtype3	180563	99.379166
targsubtype3_txt	180570	99.383018
corp3	180653	99.428700
target3	180507	99.348344
natlty3	180539	99.365956

	nulos	porcentaje
natlty3_txt	180540	99.366507
gname	487	0.268037
gsubname	175640	96.669620
gname2	179575	98.835385
gsubname2	181441	99.862404
gname3	181302	99.785900
gsubname3	181637	99.970279
motive	131567	72.412503
guncertain1	687	0.378114
guncertain2	179215	98.637247
guncertain3	180967	99.601521
individual	517	0.284549
nperps	71368	39.279876
nperpcap	69737	38.382198
claimed	66283	36.481169
claimmode	162383	89.373167
claimmode_txt	162472	89.422151
claim2	179356	98.714851
claimmode2	180731	99.471630
claimmode2_txt	180873	99.549785
claim3	181275	99.771040
claimmode3	181497	99.893225

	nulos	porcentaje
claimmode3_txt	181521	99.906435
compclaim	176959	97.395578
weaptype1	692	0.380866
weaptype1_txt	436	0.239968
weapsubtype1	20958	11.534969
weapsubtype1_txt	20846	11.473326
weaptype2	167993	92.460826
weaptype2_txt	168199	92.574206
weapsubtype2	169926	93.524721
weapsubtype2_txt	170047	93.591317
weaptype3	179726	98.918493
weaptype3_txt	179767	98.941059
weapsubtype3	179963	99.048935
weapsubtype3_txt	179980	99.058291
weaptype4	181594	99.946613
weaptype4_txt	181598	99.948814
weapsubtype4	181607	99.953768
weapsubtype4_txt	181609	99.954868
weapdetail	68106	37.484520
nkill	11074	6.094963
nkillus	64922	35.732095
nkillter	67239	37.007337

	nulos	porcentaje
nwound	16440	9.048329
nwoundus	64793	35.661095
nwoundte	69183	38.077285
property	50	0.027519
propextent	117295	64.557408
propextent_txt	117429	64.631159
propvalue	142422	78.386932
propcomment	123759	68.115097
ishostkid	542	0.298309
nhostkid	167657	92.275897
nhostkidus	167853	92.383772
nhours	177390	97.632794
ndays	173472	95.476386
divert	181268	99.767187
kidhijcountry	178334	98.152358
ransom	104420	57.471201
ransomamt	180254	99.209097
ransomamtus	181080	99.663715
ransompaid	180886	99.556940
ransompaidus	181117	99.684079
ransomnote	181165	99.710497
hostkidoutcome	170754	93.980439

	nulos	porcentaje
hostkidoutcome_txt	170728	93.966129
nreleased	171300	94.280950
addnotes	153695	84.591422
scite1	66796	36.763516
scite2	70989	39.071280
scite3	74436	40.968457
dbsource	877	0.482688
INT_LOG	26853	14.779488
INT_IDEO	33185	18.264526
INT_MISC	8094	4.454816
INT_ANY	1851	1.018763
related	84106	46.290680

4. Análisis Estadístico Descriptivo

4.1 Estadísticas Generales

In [10]:

```
# Estadísticas Descriptivas
# Calculamos estadísticas descriptivas para entender la distribución de las variables numéricas.

df.describe()
```

Out[10]:

	eventid	iyear	imonth	iday	extended	country	region	latitude
count	1.816910e+05	181691.000000	181691.000000	181691.000000	181691.000000	181691.000000	181691.000000	177135.000000
mean	2.002705e+11	2002.638997	6.467277	15.505644	0.045346	131.968501	7.160938	23.498343
std	1.325957e+09	13.259430	3.388303	8.814045	0.208063	112.414535	2.933408	18.569242
min	1.970000e+11	1970.000000	0.000000	0.000000	0.000000	4.000000	1.000000	-53.154613
25%	1.991021e+11	1991.000000	4.000000	8.000000	0.000000	78.000000	5.000000	11.510046
50%	2.009022e+11	2009.000000	6.000000	15.000000	0.000000	98.000000	6.000000	31.467463
75%	2.014081e+11	2014.000000	9.000000	23.000000	0.000000	160.000000	10.000000	34.685087
max	2.017123e+11	2017.000000	12.000000	31.000000	1.000000	1004.000000	12.000000	74.633553

In [11]:

```python
# Análisis estadístico descriptivo
print(df[['nkill', 'nkillus', 'nkillter', 'nwound', 'nwoundus', 'nwoundte']].describe())
```

```
               nkill        nkillus       nkillter         nwound  \
count  170617.000000  116769.000000  114452.000000  165251.000000
mean        2.403823       0.053336       0.509497       3.152641
std        11.554776       5.728405       4.196603      35.939637
min        -9.000000       0.000000       0.000000       0.000000
25%         0.000000       0.000000       0.000000       0.000000
50%         0.000000       0.000000       0.000000       0.000000
75%         2.000000       0.000000       0.000000       2.000000
max      1570.000000    1360.000000     500.000000    8191.000000

            nwoundus       nwoundte
count  116898.000000  112508.000000
mean        0.049556       0.117467
std         3.145368       1.969353
min         0.000000       0.000000
25%         0.000000       0.000000
50%         0.000000       0.000000
75%         0.000000       0.000000
max       751.000000     405.000000
```

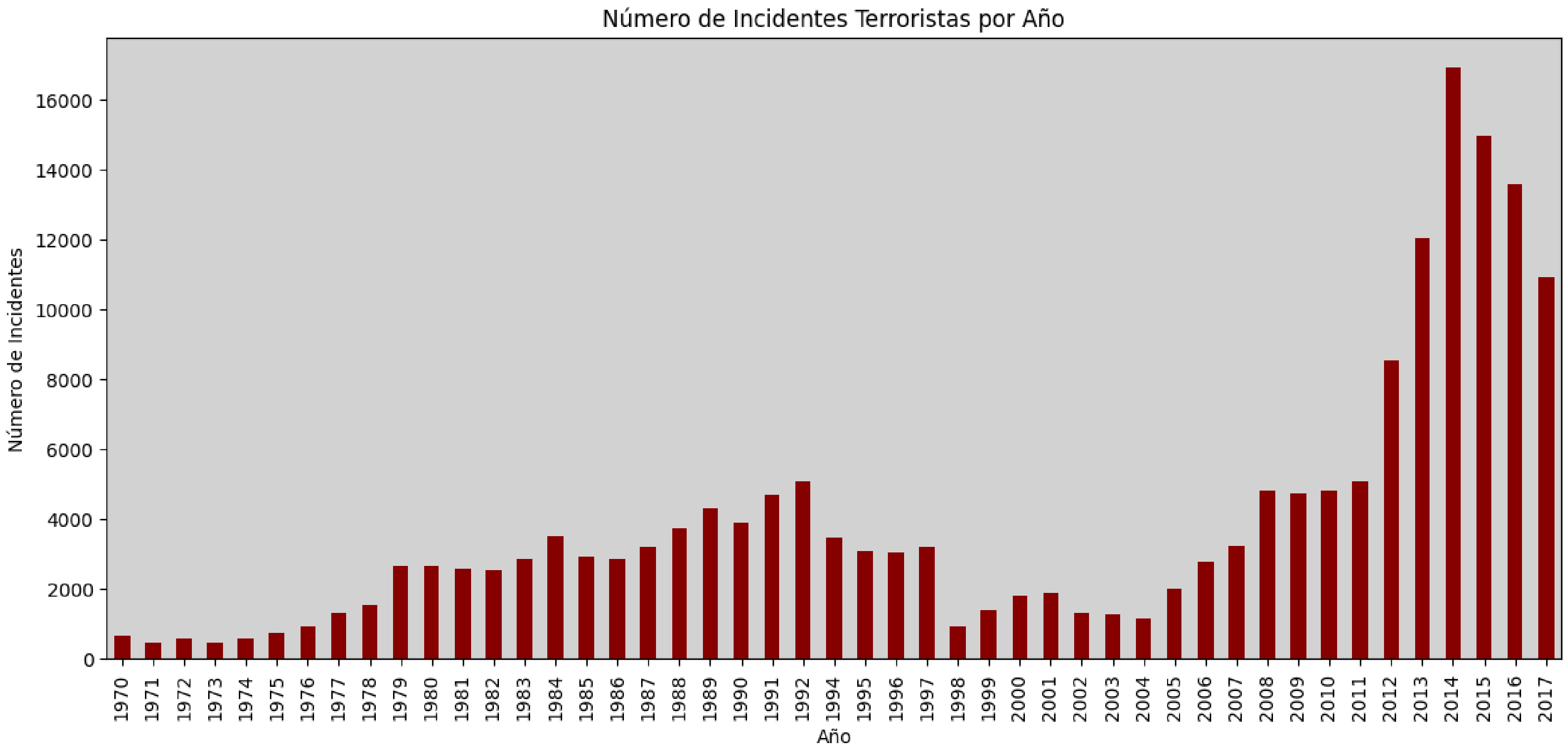
In [12]:

```
# Análisis de Correlación
# Calculamos la correlación entre variables numéricas para observar posibles relaciones.

correlation_matrix = df[['nkill', 'nwound']].corr()
print("Matriz de Correlación entre 'nkill' y 'nwound':\n", correlation_matrix)
```

```
Matriz de Correlación entre 'nkill' y 'nwound':
             nkill    nwound
nkill   1.000000  0.534501
nwound  0.534501  1.000000
```

In [18]:

```
# Análisis Exploratorio de Datos (EDA)
# Incidentes por Año
# Graficamos la cantidad de incidentes terroristas reportados por año.

ax = df['iyear'].value_counts().sort_index().plot(kind='bar', figsize=(14, 6), color='darkred')
ax.set_xlabel("Año")
ax.set_ylabel("Número de Incidentes")
ax.set_title("Número de Incidentes Terroristas por Año")
ax.set_facecolor('lightgrey')
```

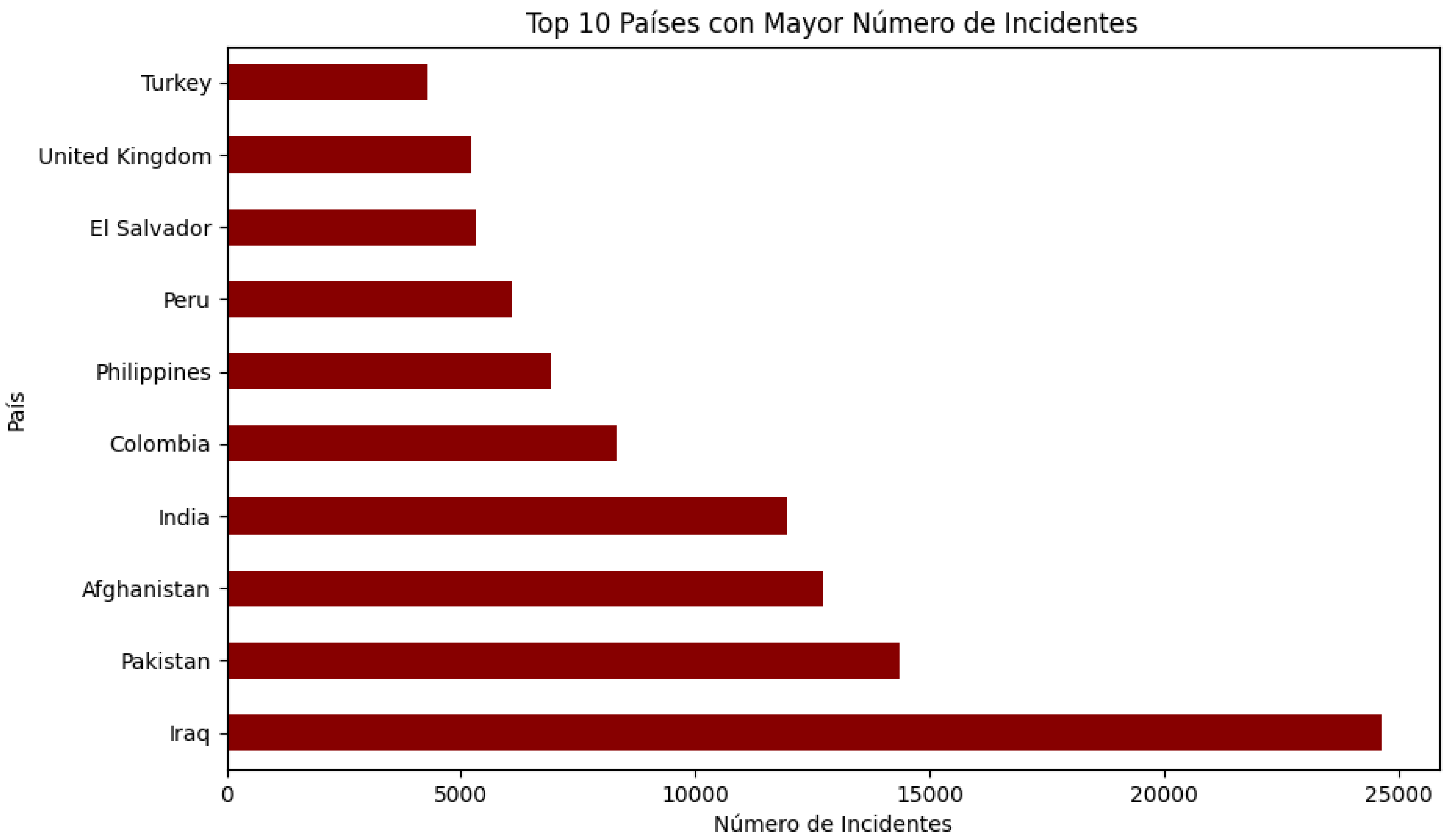


In [ ]:

```
# Países con Mayor Número de Incidentes
# Mostramos los 10 países con mayor número de incidentes.

ax = df['country_txt'].value_counts().head(10).plot(kind='barh', figsize=(10, 6), color='darkred')
ax.set_xlabel("Número de Incidentes")
ax.set_ylabel("País")
ax.set_title("Top 10 Países con Mayor Número de Incidentes")
```

Out[ ]:

```
Text(0.5, 1.0, 'Top 10 Países con Mayor Número de Incidentes')
```

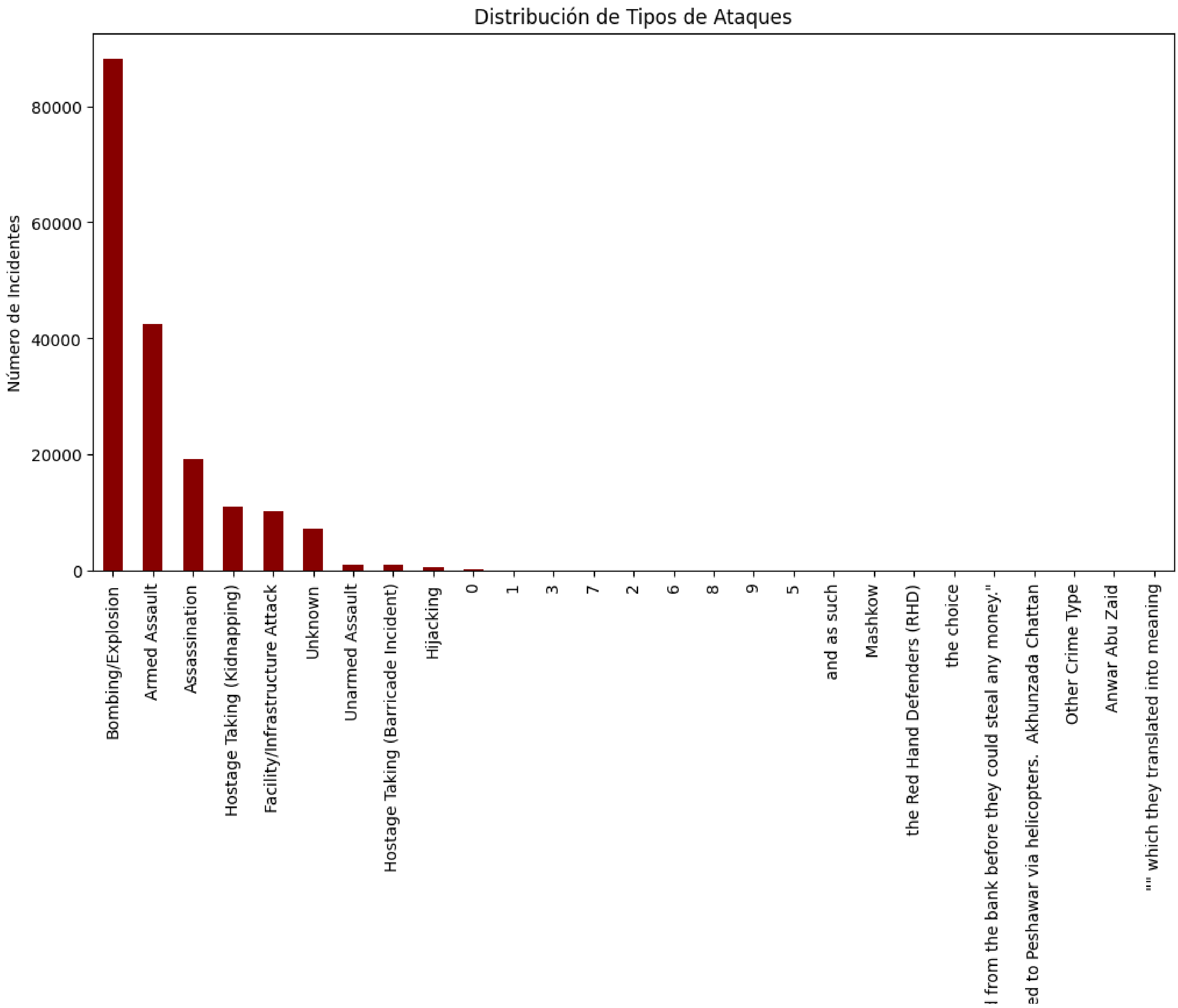


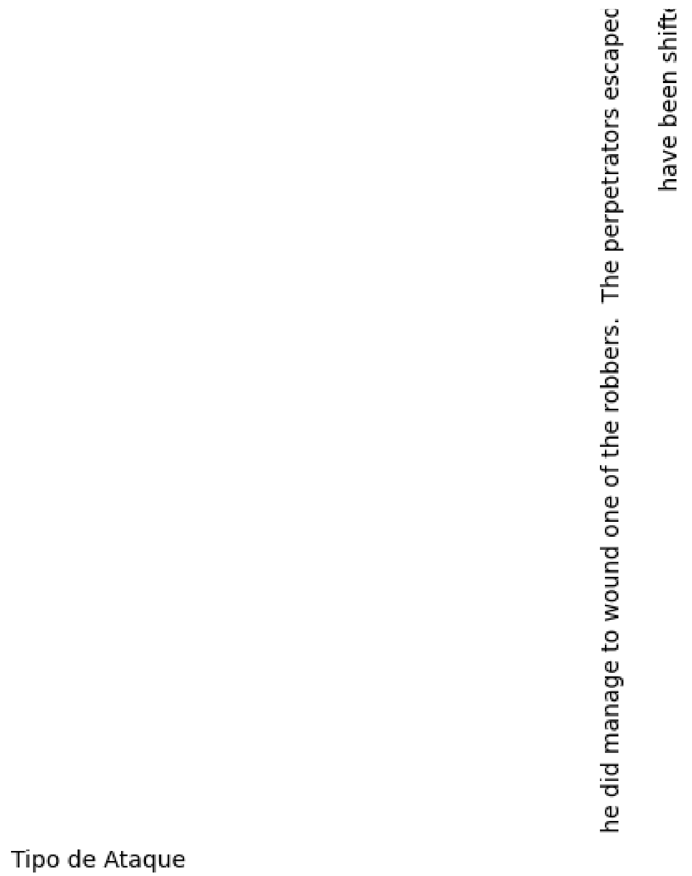
In [20]:

```
# Tipos de Ataques
# Analizamos la distribución de tipos de ataques.

ax = df['attacktype1_txt'].value_counts().plot(kind='bar', figsize=(12, 6), color='darkred')
ax.set_xlabel("Tipo de Ataque")
ax.set_ylabel("Número de Incidentes")
ax.set_title("Distribución de Tipos de Ataques")
```

Out[20]:

```
Text(0.5, 1.0, 'Distribución de Tipos de Ataques')
```


Distribución de Tipos de Ataques
Número de Incidentes
80000
60000
40000
20000
0
Bombing/Explosion
Armed Assault
Assassination
Hostage Taking (Kidnapping)
Facility/Infrastructure Attack
Unknown
Unarmed Assault
Hostage Taking (Barricade Incident)
Hijacking
0
1
3
7
2
6
8
9
5
and as such
Mashkow
the Red Hand Defenders (RHD)
the choice
l from the bank before they could steal any money."
ed to Peshawar via helicopters. Akhunzada Chattan
Other Crime Type
Anwar Abu Zaid
"" which they translated into meaning

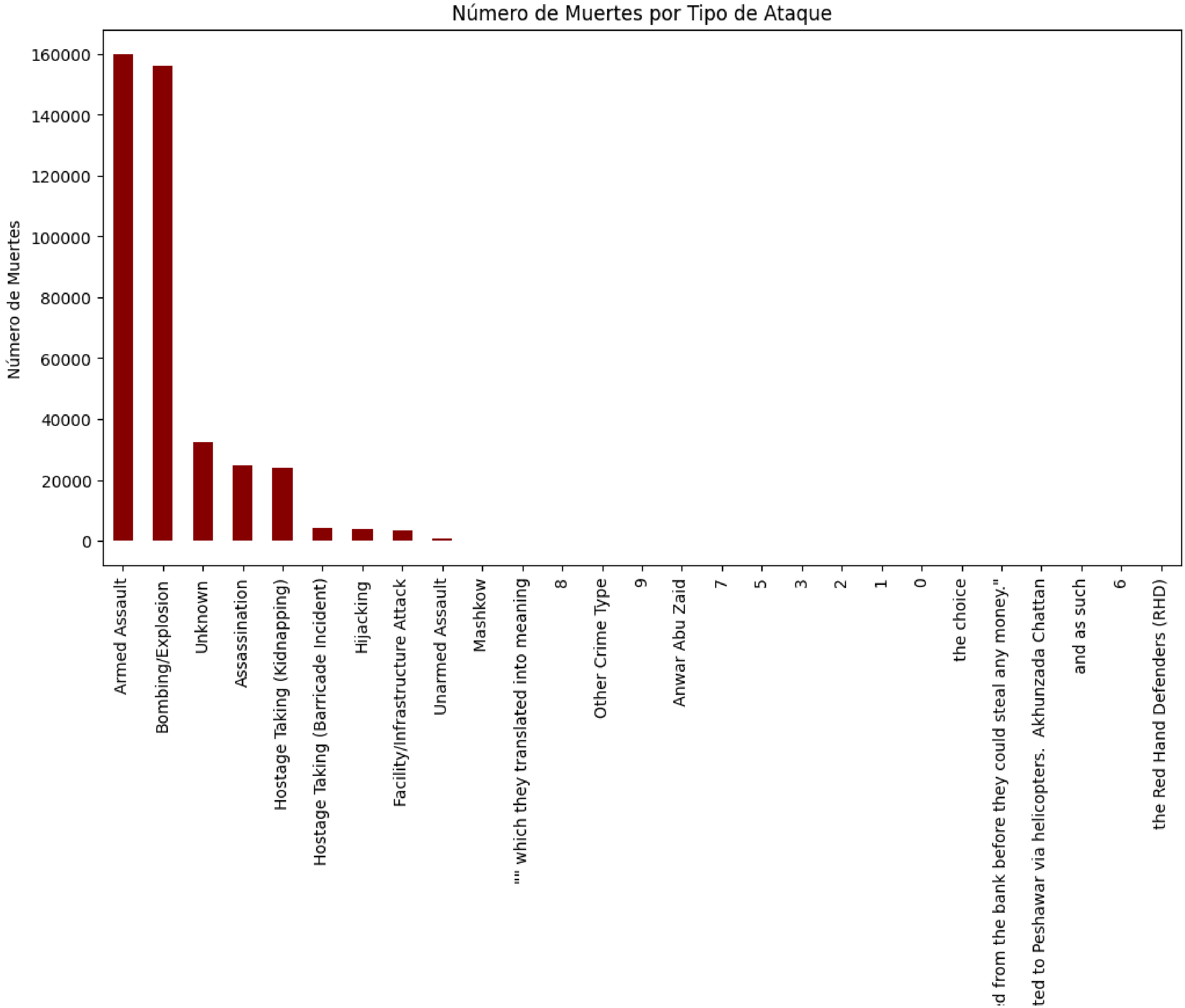


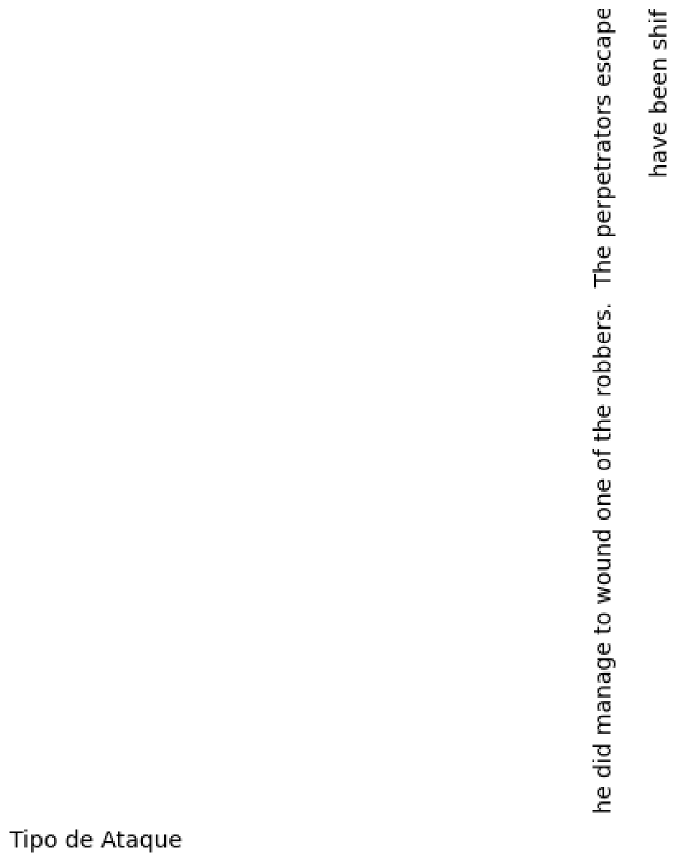
In [21]:

```
# Muertes por Tipo de Ataque
# Exploramos la cantidad total de muertes por cada tipo de ataque.

ax = df.groupby('attacktype1_txt')['nkill'].sum().sort_values(ascending=False).plot(kind='bar', figsize=(12, 6), col
ax.set_xlabel("Tipo de Ataque")
ax.set_ylabel("Número de Muertes")
ax.set_title("Número de Muertes por Tipo de Ataque")
```

Out[21]:

```
Text(0.5, 1.0, 'Número de Muertes por Tipo de Ataque')
```


Número de Muertes por Tipo de Ataque
Número de Muertes
160000
140000
120000
100000
80000
60000
40000
20000
0
Armed Assault
Bombing/Explosion
Unknown
Assassination
Hostage Taking (Kidnapping)
Hostage Taking (Barricade Incident)
Hijacking
Facility/Infrastructure Attack
Unarmed Assault
Mashkow
"" which they translated into meaning
8
Other Crime Type
9
Anwar Abu Zaid
7
5
3
2
1
0
the choice
d from the bank before they could steal any money."
ted to Peshawar via helicopters. Akhunzada Chattan
and as such
6
the Red Hand Defenders (RHD)

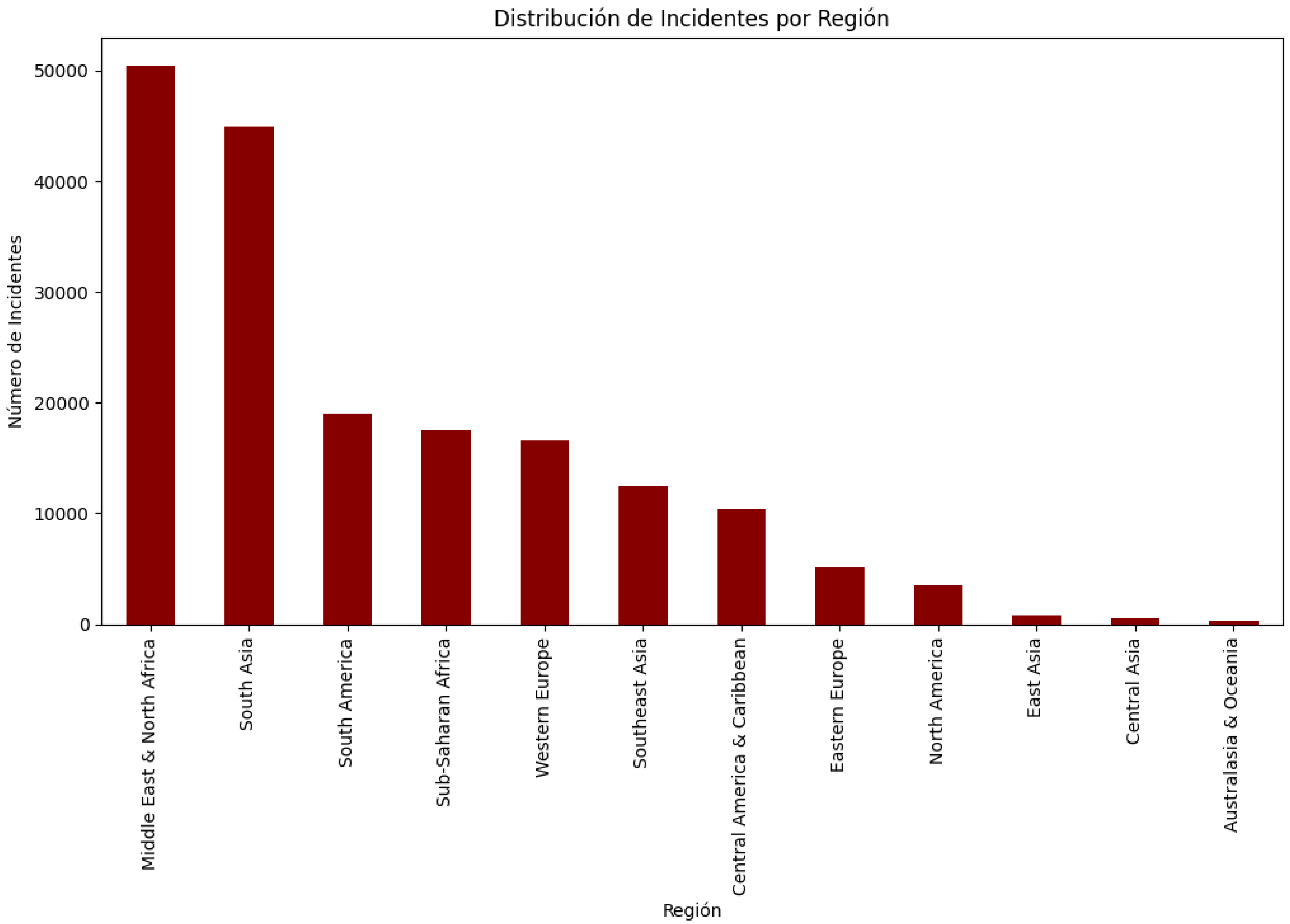


In [22]:

```
# Análisis Geográfico
# Incidentes por Región
# Observamos la cantidad de incidentes en cada región para detectar patrones geográficos.

ax = df['region_txt'].value_counts().plot(kind='bar', figsize=(12, 6), color='darkred')
ax.set_xlabel("Región")
ax.set_ylabel("Número de Incidentes")
ax.set_title("Distribución de Incidentes por Región")
```

Out[22]:

```
Text(0.5, 1.0, 'Distribución de Incidentes por Región')
```



In [23]:
```
# Parar la sesión de Spark
spark.stop()
```