



BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA



Reporte Fase 2: Procesamiento y Limpieza de Datos INTRODUCCIÓN A LA CIENCIA DE DATOS

Limpieza de una Base de Datos Ensuciada

Estimación Probabilística de Terremotos Basada en Datos Históricos

Introducción a la Ciencia de Datos
Jaime Alejandro Romero Sierra

Samuel Naranjo Madrigal
202470537

21.10.24

Índice

Análisis Inicial de la Base de Datos	3
Limpieza de Datos	6
1. Eliminación de columna ‘Tipo_dato’	7
2. Limpieza de duplicados	8
3. Limpieza de columna ‘Estatus’	9
4. Eliminación de columna ‘Tiempo’	12
5. Limpieza de columna ‘Relevancia’	13
6. Limpieza de columna ‘Magnitud’	14
7. Limpieza de columna ‘Profundidad’	15
8. Convertir ‘Lugar’ y ‘Estado’ a String	16
9. Limpieza de columna ‘Tsunami’	17
10. Limpieza de columna ‘Fecha’	19
11. Limpieza de columnas restantes (‘Lugar’, ‘Estado’, ‘Latitud’, ‘Longitud’)	21
12. Eliminación de datos negativos en ‘Magnitud’ y ‘Profundidad’	24
13. Correcciones Finales	25
Documentación y Reporte	26

Análisis Inicial de la Base de Datos

3,563,492 filas × 12 columnas

Resumen Estadístico de los Datos

	Tsunami	Relevancia	Tipo_dato	Magnitud	Longitud	Latitud
count	3.456588e+06	3.387515e+06	0.0	3.456588e+06	3.456588e+06	3.387402e+06
mean	4.452367e-04	7.401895e+01	NaN	1.774327e+00	-1.012856e+02	3.746734e+01
std	2.109594e-02	1.016261e+02	NaN	1.291145e+00	7.696839e+01	2.041068e+01
min	0.000000e+00	0.000000e+00	NaN	-9.990000e+00	-1.799997e+02	-8.442200e+01
25%	0.000000e+00	1.300000e+01	NaN	9.100000e-01	-1.464272e+02	3.406433e+01
50%	0.000000e+00	3.300000e+01	NaN	1.460000e+00	-1.189532e+02	3.793400e+01
75%	0.000000e+00	8.100000e+01	NaN	2.300000e+00	-1.159240e+02	4.785329e+01
max	1.000000e+00	2.910000e+03	NaN	9.100000e+00	1.800000e+02	8.738600e+01

	Tsunami	Relevancia	Tipo_dato	Magnitud	Longitud	Latitud
Cantidad	3,456,588	3,387,515	0	3,456,588	3,456,588	3,387,402
Promedio	0.0004452	74.01895	NaN	1.774327	-101.2856	37.46734
Desviación std	0.02109594	101.6261	NaN	1.291145	76.96839	20.41068
mínimo	0	0	NaN	-9.99	-179.9997	-84.422
25%	0	13	NaN	0.91	-146.4272	34.06433
50%	0	33	NaN	1.46	-118.9532	37.934
75%	0	81	NaN	2.3	-115.924	47.85329
máximo	1	2910	NaN	9.1	180	87.386

Valores Faltantes por Columna

Tiempo	106904
Lugar	106904
Estatus	106904
Tsunami	106904
Relevancia	175977
Tipo_dato	3563492
Magnitud	106904
Estado	106904
Longitud	106904
Latitud	176090
Profundidad	106904
Fecha	106904

Columna	Val. Total	Val. NaN	% Faltante
Tiempo	3,563,492	106,904	2.99%
Lugar	3,563,492	106,904	2.99%
Estatus	3,563,492	106,904	2.99%
Tsunami	3,563,492	106,904	2.99%
Relevancia	3,563,492	175,977	4.93%
Tipo_dato	3,563,492	3,563,492	100%
Magnitud	3,563,492	106,904	2.99%
Estado	3,563,492	106,904	2.99%
Longitud	3,563,492	106,904	2.99%
Latitud	3,563,492	176,090	4.94%
Profundidad	3,563,492	106,904	2.99%
Fecha	3,563,492	106,904	2.99%

Filas duplicadas

57,346

np.int64(57346)

Tipos de Datos

Columna	Tipo de Dato	Tipo de Dato Esperado
Tiempo	object	int
Lugar	object	string
Estatus	object	string
Tsunami	float64	boolean
Relevancia	float64	float64
Tipo_dato	float64	string
Magnitud	float64	float64
Estado	object	string
Longitud	float64	float64
Latitud	float64	float64
Profundidad	object	float64
Fecha	object	datetime

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3563492 entries, 0 to 3563491
Data columns (total 12 columns):
 #   Column      Dtype  
 0   Tiempo       object 
 1   Lugar        object 
 2   Estatus      object 
 3   Tsunami      float64
 4   Relevancia   float64
 5   Tipo_dato    float64
 6   Magnitud     float64
 7   Estado       object 
 8   Longitud     float64
 9   Latitud      float64
 10  Profundidad  object 
 11  Fecha        object 
dtypes: float64(6), object(6)
memory usage: 326.2+ MB
```

Limpieza de Datos

Antes de todo el proceso de limpieza, es necesario definir las librerías que se necesitan. En este caso se utilizó “*Pandas*” para el uso de data frames, indispensable para cargar, visualizar y manipular la base de datos. Y a su vez, se utilizó “*Numpy*” para un paso en específico en el transcurso de la limpieza. El uso de Numpy será abordado a detalle en el ese respectivo paso.

```
#librerías
import pandas as pd
import numpy as np
✓ 0.0s
```

También cabe mencionar que de antemano se modificó el nombre de las columnas con su traducción al español, esto con el fin de que el análisis inicial de la base de datos fuera aún más claro. Por esta razón, no se considera este cambio en la limpieza de datos, ya que se contempló en el apartado anterior.

```
#nombre de columnas en español
df = df.rename(columns={
    'time':'Tiempo',
    'place':'Lugar',
    'status':'Estatus',
    'tsunami':'Tsunami',
    'significance':'Relevancia',
    'data_type':'Tipo_dato',
    'magnitude':'Magnitud',
    'state':'Estado',
    'longitude':'Longitud',
    'latitude':'Latitud',
    'depth':'Profundidad',
    'date':'Fecha'
})
df
✓ 0.1s
```

Sin nada más que añadir, a continuación, el proceso de limpieza de los datos:

1. Eliminación de columna ‘Tipo_dato’

		Tiempo	Lugar	Estatus	Tsunami	Relevancia	Tipo_dato	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	bbb	12 km NNW of Meadow Lakes, Alaska	reviewed	0.0	96.0	NaN	2.50	Alaska	-149.669200	61.730200	30.1	1990-01-01 00:22:33.990000+00:00	
1	631153491210	14 km S of Volcano, Hawaii	reviewed	0.0	31.0	NaN	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00	
2	631154083450	7 km W of Cobb, California	reviewed	0.0	19.0	NaN	1.11	California	-122.806167	38.821000	3.22	1990-01-01 00:34:43.450000+00:00	
3	631155512130	11 km E of Mammoth Lakes, California	reviewed	0.0	15.0	NaN	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00	
4	631155824490	16km N of Fillmore, CA	reviewed	0.0	134.0	NaN	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00	
...
3563487	1504955275771	33 km SE of Denali National Park, Alaska	reviewed	0.0	1.0	NaN	0.30	Alaska	-151.346800	63.292600	18.4	2017-09-09 11:07:55.771000+00:00	
3563488	1157452084151	85 km WSW of Nanwalek, Alaska	reviewed	0.0	26.0	NaN	1.30	Alaska	-153.209400	58.968500	70.7	2006-09-05 10:28:04.151000+00:00	
3563489	1132870985109	16 km N of Sutcliffe, Nevada	reviewed	0.0	44.0	NaN	1.70	Nevada	NaN	40.094600	12.8	2005-11-24 22:23:05.109000+00:00	
3563490	1225142766830	Komandorskiye Ostrova, Russia region	reviewed	0.0	449.0	NaN	5.40	Russia region	169.048000	54.345000	12.2	2008-10-27 21:26:06.830000+00:00	
3563491	697353582860	8 km SW of Coleville, California	reviewed	0.0	NaN	NaN	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00	

La columna de ‘Tipo_dato’ estaba destinada a definir si el evento se trataba de un terremoto o no. Sin embargo al buscar los NaNs se puede observar que el 100% de la columna está compuesta por ellos.

Al tratarse de una base de datos cuyo objetivo es registrar terremotos, y viendo que la columna está compuesta de NaNs en su totalidad, se tomó la decisión de eliminarla por completo. Asumiendo que todos los datos tratan de terremotos y, sabiendo que, en caso de no ser así, la exactitud de base no cambiaría significativamente por una cantidad minúscula o prácticamente inexistente de discrepancias.

Python												
		Tiempo	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	bbb	12 km NNW of Meadow Lakes, Alaska	reviewed	0.0	96.0	2.50	Alaska	-149.669200	61.730200	30.1	1990-01-01 00:22:33.990000+00:00	
1	631153491210	14 km S of Volcano, Hawaii	reviewed	0.0	31.0	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00	
2	631154083450	7 km W of Cobb, California	reviewed	0.0	19.0	1.11	California	-122.806167	38.821000	3.22	1990-01-01 00:34:43.450000+00:00	
3	631155512130	11 km E of Mammoth Lakes, California	reviewed	0.0	15.0	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00	
4	631155824490	16km N of Fillmore, CA	reviewed	0.0	134.0	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00	
...
3563487	1504955275771	33 km SE of Denali National Park, Alaska	reviewed	0.0	1.0	0.30	Alaska	-151.346800	63.292600	18.4	2017-09-09 11:07:55.771000+00:00	
3563488	1157452084151	85 km WSW of Nanwalek, Alaska	reviewed	0.0	26.0	1.30	Alaska	-153.209400	58.968500	70.7	2006-09-05 10:28:04.151000+00:00	
3563489	1132870985109	16 km N of Sutcliffe, Nevada	reviewed	0.0	44.0	1.70	Nevada	NaN	40.094600	12.8	2005-11-24 22:23:05.109000+00:00	
3563490	1225142766830	Komandorskiye Ostrova, Russia region	reviewed	0.0	449.0	5.40	Russia region	169.048000	54.345000	12.2	2008-10-27 21:26:06.830000+00:00	
3563491	697353582860	8 km SW of Coleville, California	reviewed	0.0	NaN	NaN	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00

2. Limpieza de duplicados

En el análisis inicial se calculó un total de 57,346 filas duplicadas. Estas son en su totalidad una copia de una fila ya establecida, y no se trata de datos duplicados por separado o exclusivos a una columna.

Con esto en mente, se eliminaron todas estas filas duplicadas. Dejando un total de 3,506,088 filas.

```
#Eliminación de filas duplicadas
df2 = df2.drop_duplicates()
df2
✓ 9.7s
```

	Tiempo	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	bbb	12 km NNW of Meadow Lakes, Alaska	reviewed	0.0	96.0	2.50	Alaska	-149.669200	61.730200	30.1	1990-01-01 00:22:33.990000+00:00
1	631153491210	14 km S of Volcano, Hawaii	reviewed	0.0	31.0	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00
2	631154083450	7 km W of Cobb, California	reviewed	0.0	19.0	1.11	California	-122.806167	38.821000	3.22	1990-01-01 00:34:43.450000+00:00
3	631155512130	11 km E of Mammoth Lakes, California	reviewed	0.0	15.0	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00
4	631155824490	16km N of Fillmore, CA	reviewed	0.0	134.0	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00
...
3563483	939668724560	13 km SW of Aspen Springs, California	reviewed	0.0	28.0	1.36	California	-118.842500	37.483167	-1.271	1999-10-11 19:05:24.560000+00:00
3563484	1373781400820	16km ESE of Anza, CA	reviewed	0.0	NaN	0.17	California	-116.503833	33.515500	14.44	2013-07-14 05:56:40.820000+00:00
3563485	1030993456100	NaN	reviewed	0.0	284.0	4.30	New Zealand	175.050000	-40.810000	36.0	2002-09-02 19:04:16.100000+00:00
3563489	1132870985109	16 km N of Sutcliffe, Nevada	reviewed	0.0	44.0	1.70	Nevada	NaN	40.094600	12.8	2005-11-24 22:23:05.109000+00:00
3563491	697353582860	8 km SW of Coleville, California	reviewed	0.0	NaN	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00

3506088 rows × 11 columns

```
#Verificar datos duplicados
df2.duplicated().sum()
✓ 8.9s
np.int64(0)
```

3. Limpieza de columna ‘Estatus’

Al observar los valores únicos del estatus con `unique()`, se obtiene ‘review’, ‘automatic’ y ‘manual’. Estos tres tienen duplicados con su mismo nombre pero en mayúsculas, datos que deberían estar integrados en sus contrapartes originales. Además existen datos llamados ‘bbb’ que son inválidos en la columna, y NaNs que no cuentan con información en lo absoluto.

Basado en lo anterior, se utilizó un *diccionario* para corregir los nombres de los valores en la columna. Agrupando los duplicados en los valores con minúsculas por preferencia, y asignando ‘manual’ a los ‘bbb’ por ser asignados de esta manera.

```
#Corrección de valores duplicados en Estatus (cambio de nombre)
estatus_correcto = {
    'REVIEWED': 'reviewed',
    'AUTOMATIC': 'automatic',
    'MANUAL': 'manual',
    'bbb': 'manual'
}
✓ 0.0s

#Aplicación de correcciones de Estatus en df
df2['Estatus'] = df2['Estatus'].replace(estatus_correcto)
df2
✓ 0.3s
```

Posteriormente, se reemplazaron los NaN por ‘manual’ por ser asignados de esta manera. Luego se verifican los cambios:

```
#Revisión de cambios
df2['Estatus'].unique()
✓ 0.1s

array(['reviewed', 'manual', 'automatic'], dtype=object)
```

Igualmente, se tradujeron los nombres de los datos al español para ser mejor comprendidos.

```
estatus_trad = {
    'reviewed': 'Revisado',
    'automatic': 'Automático',
    'manual': 'Manual'
}
✓ 0.0s

df2['Estatus'] = df2['Estatus'].replace(estatus_trad)
✓ 0.2s
```

Finalmente se cambia el *data-type* de ‘Estatus’ en ‘string’ para guardarla como una serie de caracteres exclusivamente.

```
#Cambio del datatype a string
df2['Estatus']=df2['Estatus'].astype('string')
```

Comprobación final:

#Revisión de dt	#Revisión de NaN
print(df2.dtypes)	df2.isnull().sum()
✓ 0.0s	✓ 0.5s
Tiempo object	Tiempo 106843
Lugar object	Lugar 106842
Estatus string[python]	Estatus 0
Tsunami float64	Tsunami 106855
Relevancia float64	Relevancia 175821
Magnitud float64	Magnitud 106829
Estado object	Estado 106847
Longitud float64	Longitud 106849
Latitud float64	Latitud 175936
Profundidad object	Profundidad 106861
Fecha object	Fecha 106864
dtype: object	dtype: int64

		Tiempo	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	bbb	12 km NW of Meadow Lakes, Alaska	Revisado	0.0	96.0	2.50	Alaska	-149.669200	61.730200	30.1		1990-01-01 00:22:39.90000+00:00
1	631153491210	14 km S of Volcano, Hawaii	Revisado	0.0	31.0	1.41	Hawaii	-155.212333	19.317667	6.585		1990-01-01 00:24:51.210000+00:00
2	631154083450	7 km W of Cobb, California	Revisado	0.0	19.0	1.11	California	-122.806167	38.821000	3.22		1990-01-01 00:34:43.450000+00:00
3	631155512130	11 km E of Mammoth Lakes, California	Revisado	0.0	15.0	0.98	California	-118.846333	37.664333	-0.584		1990-01-01 00:58:32.130000+00:00
4	631155824490	16km N of Fillmore, CA	Revisado	0.0	134.0	2.95	California	-118.934000	34.546000	16.122		1990-01-01 01:03:44.490000+00:00
...
3563483	939668724560	13 km SW of Aspen Springs, California	Revisado	0.0	28.0	1.36	California	-118.842500	37.483167	-1.271		1999-10-11 19:05:24.560000+00:00
3563484	1373781400820	16km ESE of Anza, CA	Revisado	0.0	NaN	0.17	California	-116.503833	33.515500	14.44		2013-07-14 05:56:40.820000+00:00
3563485	1030993456100	NaN	Revisado	0.0	284.0	4.30	New Zealand	175.050000	-40.810000	36.0		2002-09-02 19:04:16.100000+00:00
3563489	1132870985109	16 km N of Sutcliffe, Nevada	Revisado	0.0	44.0	1.70	Nevada	NaN	40.094600	12.8		2005-11-24 22:23:05.109000+00:00
3563491	697353582860	8 km SW of Coleville, California	Revisado	0.0	NaN	2.10	California	-119.577000	38.517500	4.391		1992-02-06 05:19:42.860000+00:00

3506146 rows × 11 columns



4. Eliminación de columna ‘Tiempo’

La columna ‘Tiempo’ aparenta ser una de gran importancia, y técnicamente lo es, sin embargo también existe la columna de ‘Fecha’. Esta muestra la misma información que ‘Tiempo’, pero de manera más clara, ya que no utiliza milisegundos para expresarse, más bien, utiliza la notación natural donde está claro el día, mes, año, horas, minutos y segundos, incluso decimales.

Por esta razón se optó por eliminar en su totalidad la columna ‘Tiempo’; quedando en su lugar ‘Fecha’, con su información más clara.

```
#Eliminación de columna 'Tiempo'
df3 = df3.drop(columns=['Tiempo'])
df3
```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NNW of Meadow Lakes, Alaska	reviewed	0.0	96.0	2.50	Alaska	-149.669200	61.730200	30.1	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	reviewed	0.0	31.0	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	reviewed	0.0	19.0	1.11	California	-122.806167	38.821000	3.22	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	reviewed	0.0	15.0	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	reviewed	0.0	134.0	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00
...
3563483	13 km SW of Aspen Springs, California	reviewed	0.0	28.0	1.36	California	-118.842500	37.483167	-1.271	1999-10-11 19:05:24.560000+00:00
3563484	16km ESE of Anza, CA	reviewed	0.0	NaN	0.17	California	-116.503833	33.515500	14.44	2013-07-14 05:56:40.820000+00:00
3563485	NaN	reviewed	0.0	284.0	4.30	New Zealand	175.050000	-40.810000	36.0	2002-09-02 19:04:16.100000+00:00
3563489	16 km N of Sutcliffe, Nevada	reviewed	0.0	44.0	1.70	Nevada	NaN	40.094600	12.8	2005-11-24 22:23:05.109000+00:00
3563491	8 km SW of Coleville, California	reviewed	0.0	NaN	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00

5. Limpieza de columna ‘Relevancia’

Con esta columna sólo era necesario reemplazar los NaN. Cualquier repetición de datos no es relevante ya que no es una variable que requiera tener exclusividad de estos.

Para llenar los NaN se optó por utilizar el promedio de todos los datos en ‘Relevancia’. Esta es la forma más adecuada para llenar datos numéricos, el promedio proporciona el valor ideal para no perder tanta precisión.

```
#Rellenar NaN en 'Relevancia' con el promedio de sus datos
df3['Relevancia'].fillna(df3['Relevancia'].mean(), inplace=True)
```

✓ 0.0s

Se comprueba el reemplazo y los NaN:

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NNW of Meadow Lakes, Alaska	Revisado	0.0	96.000000	2.50	Alaska	-149.669200	61.730200	30.1	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	Revisado	0.0	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	Revisado	0.0	19.000000	1.11	California	-122.806167	38.821000	3.22	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	Revisado	0.0	15.000000	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	Revisado	0.0	134.000000	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00
...
3563483	13 km SW of Aspen Springs, California	Revisado	0.0	28.000000	1.36	California	-118.842500	37.483167	-1.271	1999-10-11 19:05:24.560000+00:00
3563484	16km ESE of Anza, CA	Revisado	0.0	74.009708	0.17	California	-116.503833	33.515500	14.44	2013-07-14 05:56:40.820000+00:00
3563485	Nan	Revisado	0.0	284.000000	4.30	New Zealand	175.050000	-40.810000	36.0	2002-09-02 19:04:16.100000+00:00
3563489	16 km N of Sutcliffe, Nevada	Revisado	0.0	44.000000	1.70	Nevada	Nan	40.094600	12.8	2005-11-24 22:23:05.109000+00:00
3563491	8 km SW of Coleville, California	Revisado	0.0	74.009708	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00

3506146 rows × 10 columns

```
#Revisión Cambios
df3.isnull().sum()
```

✓ 0.4s

```
Lugar          106842
Estatus         0
Tsunami        106855
Relevancia      0
Magnitud       106829
Estado          106847
Longitud        106849
Latitud         175936
Profundidad     106861
Fecha           106864
dtype: int64
```

6. Limpieza de columna ‘Magnitud’

Para la magnitud se utiliza el mismo principio y proceso que ‘Relevancia’ y se reemplazan los NaN con el promedio de ‘Magnitud’

```
#Rellenar NaN en 'Magnitud' con el promedio de sus datos
df3['Magnitud'].fillna(df3['Magnitud'].mean(), inplace=True)
df3
```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NNW of Meadow Lakes, Alaska	Revisado	0.0	96.000000	2.50	Alaska	-149.669200	61.730200	30.1	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	Revisado	0.0	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	Revisado	0.0	19.000000	1.11	California	-122.806167	38.821000	3.22	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	Revisado	0.0	15.000000	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	Revisado	0.0	134.000000	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00
...
3563483	13 km SW of Aspen Springs, California	Revisado	0.0	28.000000	1.36	California	-118.842500	37.483167	-1.271	1999-10-11 19:05:24.560000+00:00
3563484	16km ESE of Anza, CA	Revisado	0.0	74.009708	0.17	California	-116.503833	33.515500	14.44	2013-07-14 05:56:40.820000+00:00
3563485	NaN	Revisado	0.0	284.000000	4.30	New Zealand	175.050000	-40.810000	36.0	2002-09-02 19:04:16.100000+00:00
3563489	16 km N of Sutcliffe, Nevada	Revisado	0.0	44.000000	1.70	Nevada	NaN	40.094600	12.8	2005-11-24 22:23:05.109000+00:00
3563491	8 km SW of Coleville, California	Revisado	0.0	74.009708	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00

3506146 rows × 10 columns

```
#Revisión de cambios
df3.isnull().sum()
✓ 0.4s
```

Lugar	106842
Estatus	0
Tsunami	106855
Relevancia	0
Magnitud	0
Estado	106847
Longitud	106849
Latitud	175936
Profundidad	106861
Fecha	106864
dtype: int64	

7. Limpieza de columna ‘Profundidad’

‘Profundidad’ es un poco diferente a los anteriores. El principio es el mismo, pero la columna es de tipo ‘object’, es decir, que cuenta con datos no numéricos, y no se puede calcular su promedio.

Por lo tanto, primero se observan los valores únicos de la columna, en donde se encuentra a ‘bbb’ como uno de ellos. Este valor está causando el problema. Para corregirlo, se convierten todos los ‘bbb’ a NaN, se convierte el *data-type* a ‘float’ y posteriormente se sigue el mismo proceso de las columnas anteriores.

```
#Reemplazo de 'bbb' en Profundidad por NaN
df3['Profundidad'] = df3['Profundidad'].replace('bbb', np.nan)
✓ 0.1s

#Cambio del datatype de 'Profundidad' a float
df3['Profundidad']=df3['Profundidad'].astype(float)
✓ 0.3s

#Rellenar NaN en 'Profundidad' con el promedio de sus datos
df3['Profundidad'].fillna(df3['Profundidad'].mean(), inplace=True)
df3
✓ 0.0s
```

Es en este paso de convertir a NaN que se utilizó la librería de numpy para ‘np.nan’

<pre>#Revisión de cambios df3.isnull().sum() ✓ 0.3s</pre> <p>Lugar 106842 Estatus 0 Tsunami 106855 Relevancia 0 Magnitud 0 Estado 106847 Longitud 106849 Latitud 175936 Profundidad 0 Fecha 106864 dtype: int64</p>	<pre>#Revisión de cambios df3.info() ✓ 0.0s</pre> <p><class 'pandas.core.frame.DataFrame'> Index: 3506146 entries, 0 to 3563491 Data columns (total 10 columns): # Column Dtype 0 Lugar object 1 Estatus string 2 Tsunami float64 3 Relevancia float64 4 Magnitud float64 5 Estado object 6 Longitud float64 7 Latitud float64 8 Profundidad float64 9 Fecha object dtypes: float64(6), object(3), string(1) memory usage: 294.2+ MB</p>
---	--

8. Convertir ‘Lugar’ y ‘Estado’ a String

A un paso más cerca de tener todas las columnas con sus respectivos data-types; se convirtieron a la columna ‘Lugar’ y ‘Estado’ en ‘string’

```
#Cambio de datatype de 'Lugar' a string
df4['Lugar']=df4['Lugar'].astype('string')
✓ 0.1s

#Cambio de datatype de 'Estado' a string
df4['Estado']=df4['Estado'].astype('string')
✓ 0.0s

#Revisión de cambios
df4.info()
✓ 0.0s

<class 'pandas.core.frame.DataFrame'>
Index: 3506146 entries, 0 to 3563491
Data columns (total 10 columns):
 #   Column      Dtype  
 0   Lugar        string 
 1   Estatus       string 
 2   Tsunami      float64
 3   Relevancia   float64
 4   Magnitud     float64
 5   Estado        string 
 6   Longitud     float64
 7   Latitud      float64
 8   Profundidad  float64
 9   Fecha         object  
dtypes: float64(6), object(1), string(3)
memory usage: 294.2+ MB
```

9. Limpieza de columna ‘Tsunami’

Para limpiar esta columna, se optó por utilizar el 0 y reemplazarlo en los NaN, para luego convertir el data-type a booleano.

Se utilizó el 0 ya que ‘Tsunami’ está pensado para ser una variable booleana, es decir, de Verdadero o Falso (0 y 1). El 0 representa Falso, dato que es el más abundante ya que son pocos los terremotos causantes de tsunamis. Reemplazar los NaN con 1 implicaría que la mayoría de los terremotos causó un tsunami, lo cual es completamente falso.

The screenshot shows two code cells and their execution results. The first cell contains code to find unique values in the 'Tsunami' column and to fill NaN values with 0. The second cell shows the resulting DataFrame statistics and a preview of the data.

```
#Valores Únicos en Tsunami
df4['Tsunami'].unique()
✓ 0.0s
array([ 0., nan,  1.])

#relleno de NaN de 'Tsunami' con 0
df4['Tsunami'].fillna(0, inplace=True)
df4
✓ 0.0s
```

```
#Comprobar cambios
df4.isnull().sum()
✓ 0.3s
Lugar          106842
Estatus         0
Tsunami         0
Relevancia      0
Magnitud        0
Estado          106847
Longitud        106849
Latitud         175936
Profundidad     0
Fecha           106864
dtype: int64
```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NNW of Meadow Lakes, Alaska	Revisado	0.0	96.000000	2.50	Alaska	-149.669200	61.730200	30.100	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	Revisado	0.0	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	Revisado	0.0	19.000000	1.11	California	-122.806167	38.821000	3.220	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	Revisado	0.0	15.000000	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	Revisado	0.0	134.000000	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00
...
3563483	13 km SW of Aspen Springs, California	Revisado	0.0	28.000000	1.36	California	-118.842500	37.483167	-1.271	1999-10-11 19:05:24.560000+00:00
3563484	16km ESE of Anza, CA	Revisado	0.0	74.009708	0.17	California	-116.503833	33.515500	14.440	2013-07-14 05:56:40.820000+00:00
3563485	<NA>	Revisado	0.0	284.000000	4.30	New Zealand	175.050000	-40.810000	36.000	2002-09-02 19:04:16.100000+00:00
3563489	16 km N of Sutcliffe, Nevada	Revisado	0.0	44.000000	1.70	Nevada	NaN	40.094600	12.800	2005-11-24 22:23:05.109000+00:00
3563491	8 km SW of Coleville, California	Revisado	0.0	74.009708	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00

Finalmente se cambia el *data-type* de la columna a ‘booleano’:

```

#Cambio del datatype de 'Tsunami' a booleano
df4['Tsunami']=df4['Tsunami'].astype(bool)

✓ 0.0s

#Revisión de cambios
df4.info()

✓ 0.0s

<class 'pandas.core.frame.DataFrame'>
Index: 3506146 entries, 0 to 3563491
Data columns (total 10 columns):
 #   Column      Dtype  
 0   Lugar        string 
 1   Estatus       string 
 2   Tsunami      bool    
 3   Relevancia   float64
 4   Magnitud     float64
 5   Estado        string 
 6   Longitud     float64
 7   Latitud      float64
 8   Profundidad  float64
 9   Fecha         object 
dtypes: bool(1), float64(5), object(1), string(3)
memory usage: 270.8+ MB

```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NW of Meadow Lakes, Alaska	Revisado	False	96.000000	2.50	Alaska	-149.669200	61.730200	30.100	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	Revisado	False	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	Revisado	False	19.000000	1.11	California	-122.806167	38.821000	3.220	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	Revisado	False	15.000000	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	Revisado	False	134.000000	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00
...
3563483	13 km SW of Aspen Springs, California	Revisado	False	28.000000	1.36	California	-118.842500	37.483167	-1.271	1999-10-11 19:05:24.560000+00:00
3563484	16km ESE of Anza, CA	Revisado	False	74.009708	0.17	California	-116.503833	33.515500	14.440	2013-07-14 05:56:40.820000+00:00
3563485	<NA>	Revisado	False	284.000000	4.30	New Zealand	175.050000	-40.810000	36.000	2002-09-02 19:04:16.100000+00:00
3563489	16 km N of Sutcliffe, Nevada	Revisado	False	44.000000	1.70	Nevada	NaN	40.094600	12.800	2005-11-24 22:23:05.109000+00:00
3563491	8 km SW of Coleville, California	Revisado	False	74.009708	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00

3506146 rows × 10 columns

```

#Revisión de columnas por limpiar
df5.isnull().sum()

✓ 0.3s

Lugar          106842
Estatus         0
Tsunami         0
Relevancia     0
Magnitud        0
Estado          106847
Longitud        106849
Latitud         175936
Profundidad     0
Fecha           106864
dtype: int64

```

10. Limpieza de columna ‘Fecha’

Para reemplazar los NaN en ‘Fecha’ se optó por el método de *forward fill* qué consiste el llenar el NaN con el valor anterior más cercano que no sea NaN.

Se utilizó este método ya que es el que menos afecta la exactitud de las fechas. Hace que dos terremotos aparenten haber ocurrido simultáneamente, pero al ser una minoría de los datos, el cambio no es tan significativo y no hay otro método que haga el mismo trabajo sin afectar considerablemente los resultados.

```
#Relleno de NaN en 'Fecha' con metodo forward fill
df5['Fecha'] = df5['Fecha'].fillna(method='ffill')
✓ 0.2s
```

```
#Revisión de cambios
df5.isnull().sum()
✓ 0.2s

Lugar          106842
Estatus         0
Tsunami         0
Relevancia      0
Magnitud        0
Estado          106847
Longitud        106849
Latitud         175936
Profundidad     0
Fecha           0
dtype: int64
```



Posteriormente se cambia el *data-type* a *datetime* en formato mixto, ya que la columna ya se encuentra en formato, pero sin el *data-type* correcto.

```

#Cambiar datatype de 'Fecha' a datetime
df5['Fecha'] = pd.to_datetime(df5['Fecha'], format='mixed')
✓ 4.2s

#Comprobación de cambios
df5.info()

✓ 0.0s

<class 'pandas.core.frame.DataFrame'>
Index: 3506146 entries, 0 to 3563491
Data columns (total 10 columns):
 #   Column            Dtype    
0   Lugar              string   
1   Estatus             string   
2   Tsunami            bool      
3   Relevancia          float64  
4   Magnitud            float64  
5   Estado              string   
6   Longitud            float64  
7   Latitud              float64  
8   Profundidad         float64  
9   Fecha               datetime64[ns, UTC] 
dtypes: bool(1), datetime64[ns, UTC](1), float64(5), string(3)
memory usage: 270.8 MB

```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NNW of Meadow Lakes, Alaska	Revisado	False	96.000000	2.50	Alaska	-149.669200	61.730200	30.100	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	Revisado	False	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	Revisado	False	19.000000	1.11	California	-122.806167	38.821000	3.220	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	Revisado	False	15.000000	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	Revisado	False	134.000000	2.95	California	-118.934000	34.546000	16.122	1990-01-01 01:03:44.490000+00:00
...
3563483	13 km SW of Aspen Springs, California	Revisado	False	28.000000	1.36	California	-118.842500	37.483167	-1.271	1999-10-11 19:05:24.560000+00:00
3563484	16km ESE of Anza, CA	Revisado	False	74.009708	0.17	California	-116.503833	33.515500	14.440	2013-07-14 05:56:40.820000+00:00
3563485	<NA>	Revisado	False	284.000000	4.30	New Zealand	175.050000	-40.810000	36.000	2002-09-02 19:04:16.100000+00:00
3563489	16 km N of Sutcliffe, Nevada	Revisado	False	44.000000	1.70	Nevada	NaN	40.094600	12.800	2005-11-24 22:23:05.109000+00:00
3563491	8 km SW of Coleville, California	Revisado	False	74.009708	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00

3506146 rows × 10 columns

11. Limpieza de columnas restantes ('Lugar', 'Estado', 'Latitud', 'Longitud')

Con este paso se tuvieron muchos problemas a la hora de encontrar la manera más adecuada de eliminar NaN sin alterar la ubicación de los terremotos, ya que el modelo a realizar con estos datos depende principalmente de estas cuatro columnas.

Al final se llegó a la conclusión de que lo más ideal consiste en reemplazar los NaN con texto que indicará la ausencia del dato. De esta manera se evitaba reemplazar con datos incorrectos que afectarán directamente la ubicación de los eventos.

Se procedió a cambiar los NaN en 'Lugar' y 'Estado' con el texto 'No definido'.

```
#Rellenar NaN de 'Lugar' con 'No definido'  
df6['Lugar'] = df6['Lugar'].fillna('No definido')  
  
#Rellenar NaN de 'Estado' con 'No definido'  
df6['Estado'] = df6['Estado'].fillna('No definido')
```

También se reemplazó en 'Latitud' y 'Longitud', aunque con un espacio vacío:

```
#Rellenar NaN de 'Latitud' con ''  
df6['Latitud'] = df6['Latitud'].fillna('')  
✓ 0.1s  
  
#Rellenar NaN de 'Longitud' con ''  
df6['Longitud'] = df6['Longitud'].fillna('')  
✓ 0.1s
```

El único problema con este proceso es que, para ‘Latitud’ y ‘Longitud’, el reemplazo de NaN con cualquier dato diferente a datos numéricos, significa el cambio de *data-type* a ‘object’. Lo cuál no es ideal, pero no se pudo encontrar una solución a esto, ya que reemplazar por cualquier dato numérico implica colocar coordenadas potencialmente incorrectas que afectarían directamente la precisión del modelo a realizar.

Dejar NaN suena como la única opción viable para mantener el *data-type*. Se decidió eliminarlos, aunque esto quedaría a criterio del futuro.

```
#Comprobar limpieza total
df6.isnull().sum()
✓ 0.3s
Lugar      0
Estatus     0
Tsunami    0
Relevancia  0
Magnitud    0
Estado      0
Longitud    0
Latitud     0
Profundidad 0
Fecha       0
dtype: int64
```



```
#Comprobar datatype final
df6.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
Index: 3506146 entries, 0 to 3563491
Data columns (total 10 columns):
 #   Column        Dtype  
 0   Lugar         string 
 1   Estatus        string 
 2   Tsunami       bool   
 3   Relevancia    float64
 4   Magnitud      float64
 5   Estado         string 
 6   Longitud       object  
 7   Latitud        object  
 8   Profundidad   float64
 9   Fecha          datetime64[ns, UTC]
dtypes: bool(1), datetime64[ns, UTC](1), float64(3), object(2), string(3)
memory usage: 270.8+ MB
```

#BASE DE DATOS LIMPIA											
df6											
✓ 0.0s											
	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha	
0	12 km NNW of Meadow Lakes, Alaska	Revisado	False	96.000000	2.50	Alaska	-149.6692	61.7302	30.100	1990-01-01 00:22:33.990000+00:00	
1	14 km S of Volcano, Hawaii	Revisado	False	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585	1990-01-01 00:24:51.210000+00:00	
2	7 km W of Cobb, California	Revisado	False	19.000000	1.11	California	-122.806167	38.821	3.220	1990-01-01 00:34:43.450000+00:00	
3	11 km E of Mammoth Lakes, California	Revisado	False	15.000000	0.98	California	-118.846333	37.664333	-0.584	1990-01-01 00:58:32.130000+00:00	
4	16km N of Fillmore, CA	Revisado	False	134.000000	2.95	California	-118.934	34.546	16.122	1990-01-01 01:03:44.490000+00:00	
...	
3563483	13 km SW of Aspen Springs, California	Revisado	False	28.000000	1.36	California	-118.8425	37.483167	-1.271	1999-10-11 19:05:24.560000+00:00	
3563484	16km ESE of Anza, CA	Revisado	False	74.009708	0.17	California	-116.503833	33.5155	14.440	2013-07-14 05:56:40.820000+00:00	
3563485	No definido	Revisado	False	284.000000	4.30	New Zealand	175.05	-40.81	36.000	2002-09-02 19:04:16.100000+00:00	
3563489	16 km N of Sutcliffe, Nevada	Revisado	False	44.000000	1.70	Nevada		40.0946	12.800	2005-11-24 22:23:05.109000+00:00	
3563491	8 km SW of Coleville, California	Revisado	False	74.009708	2.10	California	-119.577	38.5175	4.391	1992-02-06 05:19:42.860000+00:00	

3506146 rows × 10 columns



12. Eliminación de datos negativos en ‘Magnitud’ y ‘Profundidad’

Se puede observar que en estas columnas, existen valores negativos que van en contra de la lógica de sus propósitos, ya que es imposible contar con magnitudes o profundidades negativas.

Para solucionar esto, se obtuvo el promedio de los valores positivos de ambas columnas, y posteriormente se llenaron los valores negativos con este promedio. De esta manera se eliminan los valores imposibles, sin que sus reemplazos alteren la precisión de la base de datos de manera significativa.

Aquí también se usa numpy

```
#Promedio de los datos positivos de 'Magnitud' y 'Profundidad'
promedio_mag = df7['Magnitud'][df7['Magnitud'] > 0].mean()
promedio_prof = df7['Profundidad'][df7['Profundidad'] > 0].mean()

#Eliminación de valores negativos en 'Magnitud' y 'Profundidad'
df7['Magnitud'] = df7['Magnitud'].apply(lambda x: promedio_mag if x < 0 else x)
df7['Profundidad'] = df7['Profundidad'].apply(lambda x: promedio_prof if x < 0 else x)
```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NNW of Meadow Lakes, Alaska	Revisado	False	96.000000	2.50	Alaska	-149.6692	61.7302	30.100000	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	Revisado	False	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585000	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	Revisado	False	19.000000	1.11	California	-122.806167	38.821	3.220000	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	Revisado	False	15.000000	0.98	California	-118.846333	37.664333	24.425023	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	Revisado	False	134.000000	2.95	California	-118.934	34.546	16.122000	1990-01-01 01:03:44.490000+00:00
...
3563483	13 km SW of Aspen Springs, California	Revisado	False	28.000000	1.36	California	-118.8425	37.483167	24.425023	1999-10-11 19:05:24.560000+00:00
3563484	16km ESE of Anza, CA	Revisado	False	74.009708	0.17	California	-116.503833	33.5155	14.440000	2013-07-14 05:56:40.820000+00:00
3563485	No definido	Revisado	False	284.000000	4.30	New Zealand	175.05	-40.81	36.000000	2002-09-02 19:04:16.100000+00:00
3563489	16 km N of Sutcliffe, Nevada	Revisado	False	44.000000	1.70	Nevada		40.0946	12.800000	2005-11-24 22:23:05.109000+00:00
3563491	8 km SW of Coleville, California	Revisado	False	74.009708	2.10	California	-119.577	38.5175	4.391000	1992-02-06 05:19:42.860000+00:00

3506146 rows × 10 columns

13. Correcciones Finales

Finalmente se resetea el index para que encaje con el nuevo número de filas; y se eliminan las filas duplicadas que fueron generadas durante el proceso, por la cantidad de reemplazos NaN realizados.

```
df7.reset_index(drop=True, inplace=True)
df7
✓ 0.0s
```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NNW of Meadow Lakes, Alaska	Revisado	False	96.000000	2.50	Alaska	-149.6692	61.7302	30.100000	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	Revisado	False	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585000	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	Revisado	False	19.000000	1.11	California	-122.806167	38.821	3.220000	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	Revisado	False	15.000000	0.98	California	-118.846333	37.664333	24.425023	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	Revisado	False	134.000000	2.95	California	-118.934	34.546	16.122000	1990-01-01 01:03:44.490000+00:00
...
3506141	13 km SW of Aspen Springs, California	Revisado	False	28.000000	1.36	California	-118.8425	37.483167	24.425023	1999-10-11 19:05:24.560000+00:00
3506142	16km ESE of Anza, CA	Revisado	False	74.009708	0.17	California	-116.503833	33.5155	14.440000	2013-07-14 05:56:40.820000+00:00
3506143	No definido	Revisado	False	284.000000	4.30	New Zealand	175.05	-40.81	36.000000	2002-09-02 19:04:16.100000+00:00
3506144	16 km N of Sutcliffe, Nevada	Revisado	False	44.000000	1.70	Nevada		40.0946	12.800000	2005-11-24 22:23:05.109000+00:00
3506145	8 km SW of Coleville, California	Revisado	False	74.009708	2.10	California	-119.577	38.5175	4.391000	1992-02-06 05:19:42.860000+00:00
3506146	rows × 10 columns									

```
df7.duplicated().sum()
✓ 6.2s
np.int64(10234)

df8 = df7.drop_duplicates()
✓ 5.9s

df8.reset_index(drop=True, inplace=True)
df8
✓ 0.0s
```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha
0	12 km NNW of Meadow Lakes, Alaska	Revisado	False	96.000000	2.50	Alaska	-149.6692	61.7302	30.100000	1990-01-01 00:22:33.990000+00:00
1	14 km S of Volcano, Hawaii	Revisado	False	31.000000	1.41	Hawaii	-155.212333	19.317667	6.585000	1990-01-01 00:24:51.210000+00:00
2	7 km W of Cobb, California	Revisado	False	19.000000	1.11	California	-122.806167	38.821	3.220000	1990-01-01 00:34:43.450000+00:00
3	11 km E of Mammoth Lakes, California	Revisado	False	15.000000	0.98	California	-118.846333	37.664333	24.425023	1990-01-01 00:58:32.130000+00:00
4	16km N of Fillmore, CA	Revisado	False	134.000000	2.95	California	-118.934	34.546	16.122000	1990-01-01 01:03:44.490000+00:00
...
3495907	13 km SW of Aspen Springs, California	Revisado	False	28.000000	1.36	California	-118.8425	37.483167	24.425023	1999-10-11 19:05:24.560000+00:00
3495908	16km ESE of Anza, CA	Revisado	False	74.009708	0.17	California	-116.503833	33.5155	14.440000	2013-07-14 05:56:40.820000+00:00
3495909	No definido	Revisado	False	284.000000	4.30	New Zealand	175.05	-40.81	36.000000	2002-09-02 19:04:16.100000+00:00
3495910	16 km N of Sutcliffe, Nevada	Revisado	False	44.000000	1.70	Nevada		40.0946	12.800000	2005-11-24 22:23:05.109000+00:00
3495911	8 km SW of Coleville, California	Revisado	False	74.009708	2.10	California	-119.577	38.5175	4.391000	1992-02-06 05:19:42.860000+00:00
3495912	rows × 10 columns									

Documentación y Reporte

La base de datos original fue limpiada por medio de Visual Studio Code, Python, Jupyter, Numpy y Pandas. Se eliminaron 2 columnas ('Tiempo' y 'Tipo_dato'). Se eliminaron filas duplicadas y NaN encontrados a lo largo de las 10 columnas restantes.

Se tradujeron columnas al español, se corrigieron detalles en el index, se cambiaron los tipos de datos según era necesario y se eliminaron valores no válidos.

A continuación un resumen de los resultados finales:

TAMAÑO

```
df8.shape  
✓ 0.0s  
(3495912, 10)
```



SIN DUPLICADOS

```
#Sin duplicados  
df8.duplicated().sum()  
✓ 4.8s  
np.int64(0)
```

RESUMEN ESTADÍSTICO

	Relevancia	Magnitud	Profundidad
count	3.495912e+06	3.495912e+06	3.495912e+06
mean	7.401109e+01	1.830468e+00	2.399839e+01
std	9.903614e+01	1.214645e+00	5.321840e+01
min	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.400000e+01	1.000000e+00	4.250000e+00
50%	3.500000e+01	1.560000e+00	9.678000e+00
75%	7.400971e+01	2.210000e+00	2.285256e+01
max	2.910000e+03	9.100000e+00	7.358000e+02

	Relevancia	Magnitud	Profundidad
Cantidad	3,495,912	3,495,912	3,495,912
Promedio	74.01109	1.830468	23.99839
Desviación std	99.03614	1.214645	53.2184
Mínimo	0	0	0
25%	14	1	4.25
50%	35	1.56	9.678
75%	74.00971	2.21	22.85256
Máximo	2910	9.1	735.8

SIN NAN O INVALIDOS

```
Lugar          0  
Estatus        0  
Tsunami        0  
Relevancia     0  
Magnitud       0  
Estado          0  
Longitud        0  
Latitud          0  
Profundidad     0  
Fecha           0  
dtype: int64
```

Columna	Val. Total	Val. NaN	% Faltante
Lugar	3,495,912	0	0%
Estatus	3,495,912	0	0%
Tsunami	3,495,912	0	0%
Relevancia	3,495,912	0	0%
Magnitud	3,495,912	0	0%
Estado	3,495,912	0	0%
Longitud	3,495,912	0	0%
Latitud	3,495,912	0	0%
Profundidad	3,495,912	0	0%
Fecha	3,495,912	0	0%

DATA-TYPES DE COLUMNAS*

Columna	Tipo de Dato	Tipo de Dato Final
Lugar	object	string
Estatus	object	string
Tsunami	float64	boolean
Relevancia	float64	float64
Magnitud	float64	float64
Estado	object	string
Longitud	float64	object
Latitud	float64	object
Profundidad	object	float64
Fecha	object	datetime

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3495912 entries, 0 to 3495911
Data columns (total 10 columns):
 #   Column      Dtype  
 ____ 
 0   Lugar        string 
 1   Estatus       string 
 2   Tsunami      bool   
 3   Relevancia   float64
 4   Magnitud     float64
 5   Estado        string 
 6   Longitud     object  
 7   Latitud       object  
 8   Profundidad  float64
 9   Fecha         datetime64[ns, UTC]
dtypes: bool(1), datetime64[ns, UTC](1), float64(3), object(2), string(3)
memory usage: 243.4+ MB
```