



BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA



INTRODUCCIÓN A LA CIENCIA DE DATOS

Estimación Probabilística de Terremotos Basada en Datos Históricos

Proyecto Final

Introducción a la Ciencia de Datos

Jaime Alejandro Romero Sierra

Samuel Naranjo Madrigal

202470537

25.11.24

Índice

Introducción	3
Metodología	5
Proceso de Limpieza de Datos	5
● Análisis Inicial de la Base de Datos	5
● Limpieza de Datos	8
● Eliminación de columna 'Tipo_dato'	9
● Limpieza de duplicados	10
● Limpieza de columna 'Estatus'	11
● Eliminación de columna 'Tiempo'	14
● Limpieza de columna 'Relevancia'	15
● Limpieza de columna 'Magnitud'	16
● Limpieza de columna 'Profundidad'	17
● Convertir 'Lugar' y 'Estado' a String	18
● Limpieza de columna 'Tsunami'	19
● Limpieza de columna 'Fecha'	21
● Limpieza de columnas ('Lugar', 'Estado')	23
● Eliminación de datos negativos en 'Magnitud' y 'Profundidad'	24
● Eliminación de NaNs en columnas ('Longitud', 'Latitud')	25
● Extracción de información en la columna ('Fecha')	26
● Correcciones Finales	28
● Documentación y Reporte	29
Análisis Exploratorio de Datos (EDA)	32
1. Descripción General de los Datos	32
2. Visualización y Distribución de Variables Individuales	34
3. Correlación entre Variables	47
4. Análisis de Valores Atípicos	53
5. Análisis de Valores Faltantes	54
6. Relación entre Variables Categóricas y Numéricas	58
7. Observaciones y Hallazgos Importantes	61
Modelo de Machine Learning	63
Dashboard	65
Conclusiones y Futuras Líneas de Trabajo	66
Referencias	67
Anexos	67

Introducción

“Para poder afrontar cualquier circunstancia, es necesario un previo conocimiento de esta”

Los terremotos han sido protagonistas de numerosos eventos devastadores a lo largo de la historia. Su gran fuerza, poder y lo impredecible de su naturaleza, han hecho de ellos un fenómeno natural muy estudiado.

Por muchos años se han llevado a cabo incontables investigaciones acerca de este increíble fenómeno; con el fin de llegar a entender qué lo causa, cómo se comporta, dónde ocurre y más.

Con el tiempo se ha podido dar respuesta a muchas preguntas acerca de los terremotos, pero lo que aún la ciencia no ha podido conseguir es la predicción de los mismos.

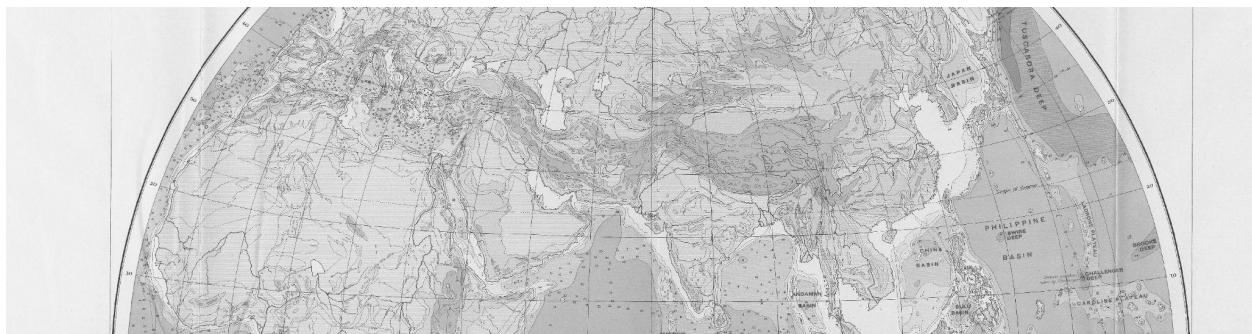
Al ser eventos tan erráticos, llegar a predecir alguno con gran antelación es actualmente imposible con los avances tecnológicos que poseemos. Aunque podemos detectar actividad sísmica, en cierto modo, con antelación; no es ni lo suficientemente grande como para ser llamada una predicción, más bien una alarma para una inminente catástrofe.



*06 de febrero de 2023: En Iskenderun, uno de los lugares más afectados por el terremoto
de magnitud 7,7 con epicentro en Kahramanmaraş*

“Estimación Probabilística de Terremotos Basada en Datos Históricos” (*EPTBDH*) no tiene como objetivo la predicción de terremotos, mas bien, el análisis de datos históricos de estos fenómenos para diseñar un modelo que muestre una aproximación estadística de posibles eventos futuros, lugares propensos a ellos, y magnitudes tentativas. Mostrando así una idea de lo que podría suceder en un futuro, mas no lo que, por hecho, sucederá.

Todo con la finalidad de tomar medidas preventivas a lo largo del mundo entero; como la mejora de estructuras resistentes a sismos en zonas propensas a estos, así como el apoyo gubernamental a la ciudadanía en caso de algún evento. La implementación de educación y medidas de prevención para algún desastre y más. En general, se busca la conciencia ciudadana y la acción de los gobiernos para estar preparados ante cualquier terremoto futuro. Con la debida preparación, afrontar cualquier circunstancia resulta más sencillo.



Para lograr el objetivo y el modelo, se utilizó una base de datos pública de la “*United States Geological Survey*” (*USGS*) extraída desde Kaggle. Esta cuenta con datos históricos desde 1990 a 2023, lo que resulta en +3,000,000 de terremotos registrados.

La base cuenta con fechas, lugares, coordenadas, magnitud y profundidad de cada evento en los últimos 33 años; suficiente para ser viable en la realización de un modelo como el esperado.

Metodología

Proceso de Limpieza de Datos

• 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

np.int64(57346)

57,346

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.

```
#librerias
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:

- **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	

3563492 rows × 12 columns

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.

#Eliminación de columna 'Tipo_dato' df2 = df.drop(columns=['Tipo_dato']) df2													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	2.10	California	-119.577000	38.517500	4.391	1992-02-06 05:19:42.860000+00:00		

3563492 rows × 11 columns

- **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)
```

- **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:

<pre>#Revisión de dt print(df2.dtypes)</pre>	<pre>#Revisión de NaN df2.isnull().sum()</pre>																																																				
✓ 0.0s	✓ 0.5s																																																				
<table border="1"> <thead> <tr> <th></th> <th></th> </tr> </thead> <tbody> <tr><td>Tiempo</td><td>object</td></tr> <tr><td>Lugar</td><td>object</td></tr> <tr><td>Estatus</td><td>string[python]</td></tr> <tr><td>Tsunami</td><td>float64</td></tr> <tr><td>Relevancia</td><td>float64</td></tr> <tr><td>Magnitud</td><td>float64</td></tr> <tr><td>Estado</td><td>object</td></tr> <tr><td>Longitud</td><td>float64</td></tr> <tr><td>Latitud</td><td>float64</td></tr> <tr><td>Profundidad</td><td>object</td></tr> <tr><td>Fecha</td><td>object</td></tr> <tr><td>dtype:</td><td>object</td></tr> </tbody> </table>			Tiempo	object	Lugar	object	Estatus	string[python]	Tsunami	float64	Relevancia	float64	Magnitud	float64	Estado	object	Longitud	float64	Latitud	float64	Profundidad	object	Fecha	object	dtype:	object	<table border="1"> <thead> <tr> <th></th> <th></th> </tr> </thead> <tbody> <tr><td>Tiempo</td><td>106843</td></tr> <tr><td>Lugar</td><td>106842</td></tr> <tr><td>Estatus</td><td>0</td></tr> <tr><td>Tsunami</td><td>106855</td></tr> <tr><td>Relevancia</td><td>175821</td></tr> <tr><td>Magnitud</td><td>106829</td></tr> <tr><td>Estado</td><td>106847</td></tr> <tr><td>Longitud</td><td>106849</td></tr> <tr><td>Latitud</td><td>175936</td></tr> <tr><td>Profundidad</td><td>106861</td></tr> <tr><td>Fecha</td><td>106864</td></tr> <tr><td>dtype:</td><td>int64</td></tr> </tbody> </table>			Tiempo	106843	Lugar	106842	Estatus	0	Tsunami	106855	Relevancia	175821	Magnitud	106829	Estado	106847	Longitud	106849	Latitud	175936	Profundidad	106861	Fecha	106864	dtype:	int64
Tiempo	object																																																				
Lugar	object																																																				
Estatus	string[python]																																																				
Tsunami	float64																																																				
Relevancia	float64																																																				
Magnitud	float64																																																				
Estado	object																																																				
Longitud	float64																																																				
Latitud	float64																																																				
Profundidad	object																																																				
Fecha	object																																																				
dtype:	object																																																				
Tiempo	106843																																																				
Lugar	106842																																																				
Estatus	0																																																				
Tsunami	106855																																																				
Relevancia	175821																																																				
Magnitud	106829																																																				
Estado	106847																																																				
Longitud	106849																																																				
Latitud	175936																																																				
Profundidad	106861																																																				
Fecha	106864																																																				
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	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



- **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

3506146 rows × 10 columns

- **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)
df3
```

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

- **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



- **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> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50px; padding: 2px;">Lugar</td> <td style="padding: 2px;">106842</td> </tr> <tr> <td>Estatus</td> <td>0</td> </tr> <tr> <td>Tsunami</td> <td>106855</td> </tr> <tr> <td>Relevancia</td> <td>0</td> </tr> <tr> <td>Magnitud</td> <td>0</td> </tr> <tr> <td>Estado</td> <td>106847</td> </tr> <tr> <td>Longitud</td> <td>106849</td> </tr> <tr> <td>Latitud</td> <td>175936</td> </tr> <tr> <td>Profundidad</td> <td>0</td> </tr> <tr> <td>Fecha</td> <td>106864</td> </tr> <tr> <td>dtype:</td> <td>int64</td> </tr> </table>	Lugar	106842	Estatus	0	Tsunami	106855	Relevancia	0	Magnitud	0	Estado	106847	Longitud	106849	Latitud	175936	Profundidad	0	Fecha	106864	dtype:	int64	<pre>#Revisión de cambios df3.info() ✓ 0.0s</pre> <pre><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</pre>
Lugar	106842																						
Estatus	0																						
Tsunami	106855																						
Relevancia	0																						
Magnitud	0																						
Estado	106847																						
Longitud	106849																						
Latitud	175936																						
Profundidad	0																						
Fecha	106864																						
dtype:	int64																						

- 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
```

- **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.

```
#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 |


3506146 rows × 10 columns

```

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

```

- **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

- **Limpieza de columnas ('Lugar', 'Estado')**

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')
```

	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	No definido	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



- **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()  
✓ 0.0s  
  
#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)  
✓ 1.0s
```

- **Eliminación de NaNs en columnas ('Longitud', 'Latitud')**

Al ser coordenadas, estas columnas son indispensables para el funcionamiento del modelo; por lo que no pueden tener datos inexactos o vacíos. En este caso, se optó por la forma más rápida para solucionar el problema, la cual consiste simplemente en eliminar todas las filas que cuenten con NaNs.



```
df9 = df8.dropna()

df9.isnull().sum()

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

	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	30.100000	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	38.821000	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	-118.934000	34.546000	16.122000	1990-01-01 01:03:44.490000+00:00
..
3563478	No definido	Revisado	False	74.009708	0.88	California	-121.699000	37.333500	7.207000	2010-03-26 10:31:44.450000+00:00
3563483	13 km SW of Aspen Springs, California	Revisado	False	28.000000	1.36	California	-118.842500	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	33.515500	14.440000	2013-07-14 05:56:40.820000+00:00
3563485	No definido	Revisado	False	284.000000	4.30	New Zealand	175.050000	-40.810000	36.000000	2002-09-02 19:04:16.100000+00:00
3563491	8 km SW of Coleville, California	Revisado	False	74.009708	2.10	California	-119.577000	38.517500	4.391000	1992-02-06 05:19:42.860000+00:00

3228568 rows × 10 columns

- **Extracción de información en la columna ('Fecha')**

Para un uso más práctico de la columna fecha, es bueno tenerla con sus datos por separado; es decir, implementar una columna de 'Año', 'Mes', 'Dia' y 'Hora'.

Estas nuevas columnas son guardadas en automático como enteros.

```
df9['Año'] = df9['Fecha'].dt.year
df9['Mes'] = df9['Fecha'].dt.month
df9['Dia'] = df9['Fecha'].dt.day
df9['Hora'] = df9['Fecha'].dt.hour
df9['DiaNombre'] = df9['Fecha'].dt.day_name()

df9
```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha	Año	Mes	Dia	DiaNombre	Hora
0	12 km NNW of Meadow Lakes, Alaska	Revisado	False	96.000000	2.50	Alaska	-149.669200	61.730200	30.100000	1990-01-01 00:22:33.990000+00:00	1990	1	1	Monday	0
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	1990	1	1	Monday	0
2	7 km W of Cobb, California	Revisado	False	19.000000	1.11	California	-122.806167	38.821000	3.220000	1990-01-01 00:34:43.450000+00:00	1990	1	1	Monday	0
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	1990	1	1	Monday	0
4	16km N of Fillmore, CA	Revisado	False	134.000000	2.95	California	-118.934000	34.546000	16.122000	1990-01-01 01:03:44.490000+00:00	1990	1	1	Monday	1
...
3563478	No definido	Revisado	False	74.009708	0.88	California	-121.699000	37.333500	7.207000	2010-03-26 10:31:44.450000+00:00	2010	3	26	Friday	10
3563483	13 km SW of Aspen Springs, Colorado	Revisado	False	28.000000	1.36	California	-118.842500	37.483167	24.425023	1999-10-11 19:05:24.560000+00:00	1999	10	11	Monday	19
3563484	16km ESE of Anza, CA	Revisado	False	74.009708	0.17	California	-116.503833	33.515500	14.440000	2013-07-14 05:56:40.820000+00:00	2013	7	14	Sunday	5
3563485	No definido	Revisado	False	284.000000	4.30	New Zealand	175.050000	-40.810000	36.000000	2002-09-02 19:04:16.100000+00:00	2002	9	2	Monday	19
3563491	8 km SW of Coleville, California	Revisado	False	74.009708	2.10	California	-119.577000	38.517500	4.391000	1992-02-06 05:19:42.860000+00:00	1992	2	6	Thursday	5

3228568 rows × 15 columns

Se convierte la columna 'DiaNombre' a str.

```
df10['DiaNombre'] = df10['DiaNombre'].astype('string')
```

Como un extra se agregó una última columna con la fecha escrita de forma natural. Esta tomó un tipo de variable objeto ya que cuenta con caracteres y números.

```
df9['FechaFormatoada'] = df9['Fecha'].dt.strftime('%d/%m/%Y %H:%M')
df9
```

	Lugar	Estatus	Tsunami	Relevancia	Magnitud	Estado	Longitud	Latitud	Profundidad	Fecha	Año	Mes	Dia	DiaNombre	Hora	FechaForma
0	12 km NNW of Meadow Lakes, Alaska	Revisado	False	96.000000	2.50	Alaska	-149.669200	61.730200	30.100000	1990-01-01 00:22:33.990000+00:00	1990	1	1	Monday	0	01/01/1990
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	1990	1	1	Monday	0	01/01/1990
2	7 km W of Cobb, California	Revisado	False	19.000000	1.11	California	-122.806167	38.821000	3.220000	1990-01-01 00:34:43.450000+00:00	1990	1	1	Monday	0	01/01/1990
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	1990	1	1	Monday	0	01/01/1990
4	16km N of Fillmore, CA	Revisado	False	134.000000	2.95	California	-118.934000	34.546000	16.122000	1990-01-01 01:03:44.490000+00:00	1990	1	1	Monday	1	01/01/1990
...
3563478	No definido	Revisado	False	74.009708	0.88	California	-121.699000	37.333500	7.207000	2010-03-26 10:31:44.450000+00:00	2010	3	26	Friday	10	26/03/2010
3563483	13 km SW of Aspen Springs, California	Revisado	False	28.000000	1.36	California	-118.842500	37.483167	24.425023	1999-10-11 19:05:24.560000+00:00	1999	10	11	Monday	19	11/10/1999
3563484	16km ESE of Anza, CA	Revisado	False	74.009708	0.17	California	-116.503833	33.515500	14.440000	2013-07-14 05:56:40.820000+00:00	2013	7	14	Sunday	5	14/07/2013



- **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.

```
df9.duplicated().sum()
np.int64(10199)

df10 = df9.drop_duplicates()

df10.duplicated().sum()
np.int64(0)

df10.reset_index(drop=True, inplace=True)
df10

df10.to_csv('BASE LIMPIA PROYECTO.csv', index=False)
```

- Documentación y Reporte

A continuación un resumen de los resultados finales:

TAMAÑO

`df10.shape`

`(3218369, 16)`

SIN DUPLICADOS

`df10.duplicated().sum()`

`np.int64(0)`

RESUMEN ESTADÍSTICO

	Relevancia	Magnitud	Longitud	Latitud	Profundidad	Año	Mes	Día	Hora
count	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06
mean	7.400951e+01	1.830472e+00	-1.012591e+02	3.746653e+01	2.399005e+01	2.009020e+03	6.502443e+00	1.536603e+01	1.143662e+01
std	9.901938e+01	1.214490e+00	7.700059e+01	2.040559e+01	5.320559e+01	9.438443e+00	3.429456e+00	8.661335e+00	6.903292e+00
min	0.000000e+00	0.000000e+00	-1.799997e+02	-8.442200e+01	0.000000e+00	1.990000e+03	1.000000e+00	1.000000e+00	0.000000e+00
25%	1.400000e+01	1.000000e+00	-1.464120e+02	3.406500e+01	4.251000e+00	2.002000e+03	4.000000e+00	8.000000e+00	5.000000e+00
50%	3.500000e+01	1.560000e+00	-1.189523e+02	3.793217e+01	9.670000e+00	2.010000e+03	7.000000e+00	1.500000e+01	1.100000e+01
75%	7.400971e+01	2.210000e+00	-1.159237e+02	4.783383e+01	2.285256e+01	2.017000e+03	9.000000e+00	2.300000e+01	1.700000e+01
max	2.910000e+03	9.100000e+00	1.800000e+02	8.738600e+01	7.358000e+02	2.023000e+03	1.200000e+01	3.000000e+01	2.300000e+01

	Relevancia	Magnitud	Longitud	Latitud	Profundidad	Año	Mes	Día	Hora
Cantidad	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369
Promedio	74.01	1.83	-101.26	37.47	23.99	2009.02	6.50	15.37	11.44
Desviación estándar	99.02	1.21	77.00	20.41	53.53	9.44	3.43	8.66	6.90
Mínimo	0.00	0.00	-179.99	-84.42	0.00	1990.00	1.00	1.00	0.00
25%	14.00	1.00	-146.41	34.07	42.51	2002.00	4.00	8.00	5.00
50%	35.00	1.56	-118.95	37.93	96.70	2010.00	7.00	15.00	11.00
75%	74.01	2.21	-115.92	47.83	228.53	2017.00	9.00	23.00	17.00
Máximo	2,910.00	9.10	180.00	87.38	735.80	2023.00	12.00	30.00	23.00

S/N NAN O INVALIDOS

df10.isnull().sum()

```
Lugar      0
Estatus    0
Tsunami   0
Relevancia 0
Magnitud   0
Estado     0
Longitud   0
Latitud    0
Profundidad 0
Fecha      0
Año        0
Mes        0
Dia        0
DiaNombre  0
Hora        0
FechaFormatoada 0
dtype: int64
```

DATA-TYPES DE COLUMNAS

Columna	Tipo de Dato	Tipo de Dato Final	Columna	Tipo de Dato	Tipo de Dato Final
Lugar	object	string	Profundidad	object	float
Estatus	object	string	Fecha	object	datetime
Tsunami	float	boolean	Año	int	int
Relevancia	float	float	Mes	int	int
Magnitud	float	float	Dia	int	int
Estado	object	string	DiaNombre	object	string
Longitud	float	float	Hora	int	int
Latitud	float	float	Fecha Formateada	object	object

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3218369 entries, 0 to 3218368
Data columns (total 16 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]
 10  Año               int32  
 11  Mes                int32  
 12  Dia                int32  
 13  DiaNombre          string 
 14  Hora               int32  
 15  FechaFormatteada  object 
dtypes: bool(1), datetime64[ns, UTC](1), float64(5), int32(4), object(1), string(4)
memory usage: 322.3+ MB
```

Análisis Exploratorio de Datos (EDA)

En esta fase se busca entender la estructura, patrones y relaciones dentro de los datos antes de construir el modelo de estimación probabilística.

1. Descripción General de los Datos

La base de datos limpia cuenta con 3,218,369 filas o terremotos registrados. Además tiene 16 columnas o variables de cada terremoto; estas son:

Lugar, Estatus, Tsunami, Relevancia, Magnitud, Estado, Longitud, Latitud, Profundidad, Fecha, Año, Mes, Dia, DiaNombre, Hora y FechaFormateada.

Cada una de estas variables cuenta con su respectivo tipo de dato (int, float, str, bool, object, datetime)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3218369 entries, 0 to 3218368
Data columns (total 16 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] 
 10  Año                int64   
 11  Mes                int64   
 12  Dia                int64   
 13  DiaNombre          string  
 14  Hora               int64   
 15  FechaFormatteada   object  
dtypes: bool(1), datetime64[ns, UTC](1), float64(5), int64(4), object(1), string(4)
memory usage: 371.4+ MB
```

bool (1)	datetime (1)	float (5)	int (4)	object (1)	string (4)
- Tsunami	- Fecha	- Relevancia - Magnitud	- Año - Mes	- Fecha Formatoada	- Lugar - Estatus

		- Longitud - Latitud - Profundidad	- Dia - Hora		- Estado - DiaNombre
--	--	--	-----------------	--	-------------------------

Así mismo, las variables numéricas cuentan con estadísticas descriptivas:

	Relevancia	Magnitud	Longitud	Latitud	Profundidad	Año	Mes	Dia	Hora
count	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06	3.218369e+06
mean	7.400951e+01	1.830472e+00	-1.012591e+02	3.746653e+01	2.399005e+01	2.009020e+03	6.502443e+00	1.536603e+01	1.143662e+01
std	9.901938e+01	1.214490e+00	7.700059e+01	2.040559e+01	5.320559e+01	9.438443e+00	3.429456e+00	8.661335e+00	6.903292e+00
min	0.000000e+00	0.000000e+00	-1.799997e+02	-8.442200e+01	0.000000e+00	1.990000e+03	1.000000e+00	1.000000e+00	0.000000e+00
25%	1.400000e+01	1.000000e+00	-1.464120e+02	3.406500e+01	4.251000e+00	2.002000e+03	4.000000e+00	8.000000e+00	5.000000e+00
50%	3.500000e+01	1.560000e+00	-1.189523e+02	3.793217e+01	9.670000e+00	2.010000e+03	7.000000e+00	1.500000e+01	1.100000e+01
75%	7.400971e+01	2.210000e+00	-1.159237e+02	4.783383e+01	2.285256e+01	2.017000e+03	9.000000e+00	2.300000e+01	1.700000e+01
max	2.910000e+03	9.100000e+00	1.800000e+02	8.738600e+01	7.358000e+02	2.023000e+03	1.200000e+01	3.000000e+01	2.300000e+01

	Relevancia	Magnitud	Longitud	Latitud	Profundidad	Año	Mes	Dia	Hora
Cantidad	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369	3,218,369
Promedio	74.01	1.83	-101.26	37.47	23.99	2009.02	6.50	15.37	11.44
Desviación estándar	99.02	1.21	77.00	20.41	53.53	9.44	3.43	8.66	6.90
Mínimo	0.00	0.00	-179.99	-84.42	0.00	1990.00	1.00	1.00	0.00
25%	14.00	1.00	-146.41	34.07	42.51	2002.00	4.00	8.00	5.00
50%	35.00	1.56	-118.95	37.93	96.70	2010.00	7.00	15.00	11.00
75%	74.01	2.21	-115.92	47.83	228.53	2017.00	9.00	23.00	17.00
Máximo	2,910.00	9.10	180.00	87.38	735.80	2023.00	12.00	30.00	23.00

Finalmente, las frecuencias de cada variable:

```

Lugar = 497530
Estatus = 3
Tsunami = 2
Relevancia = 1131
Magnitud = 759
Estado = 851
Longitud = 706319
Latitud = 502118
Profundidad = 71656
Fecha = 3086142
Año = 34
Mes = 12
Dia = 30
DiaNombre = 7
Hora = 24
FechaFormatoada = 2760850

```

2. Visualización y Distribución de Variables Individuales

En la base de datos contamos con 16 variables, de las cuales 11 son numéricas y 5 son categóricas. Es necesario identificar el tipo de variable y su contenido, esto para que los datos tengan sentido a la hora de visualizarlos, y así, poder obtener información relevante de ellos.

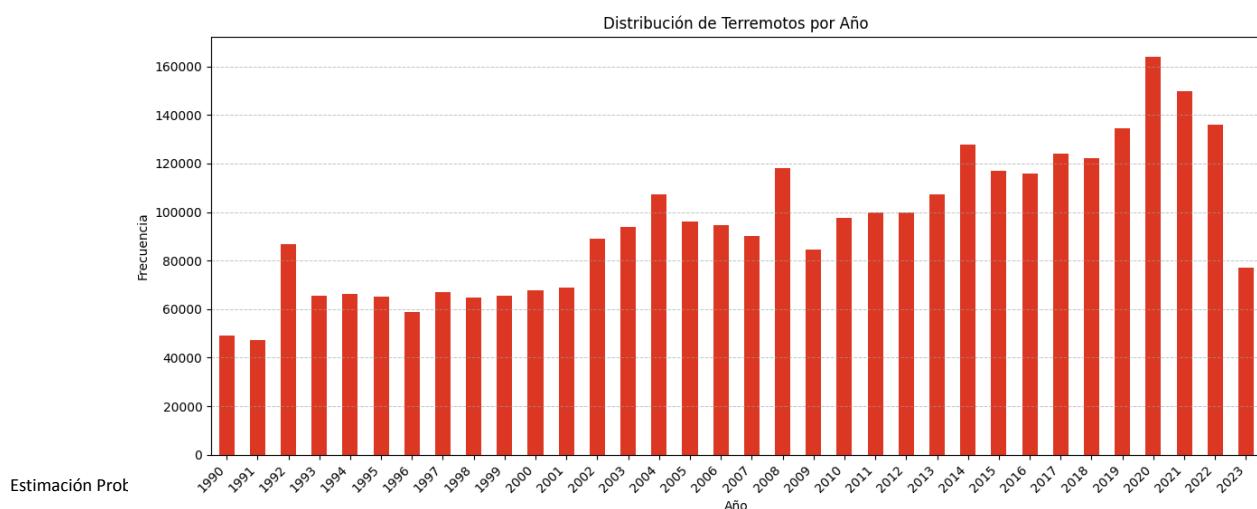
Variables Numéricas	Variables Categóricas
<ul style="list-style-type: none">• Relevancia• Magnitud• Longitud• Latitud• Profundidad• Fecha• Año• Mes• Dia• Hora• FechaFormateada	<ul style="list-style-type: none">• Lugar• Estatus• Tsunami• Estado• DiaNombre

• Variables Numéricas

Podemos dividirlas en dos:

- Valores únicos altos (Relevancia, Magnitud, Longitud, Latitud, Profundidad, Fecha, FechaFormateada)
- Valores únicos bajos (Año, Mes, Dia, Hora)

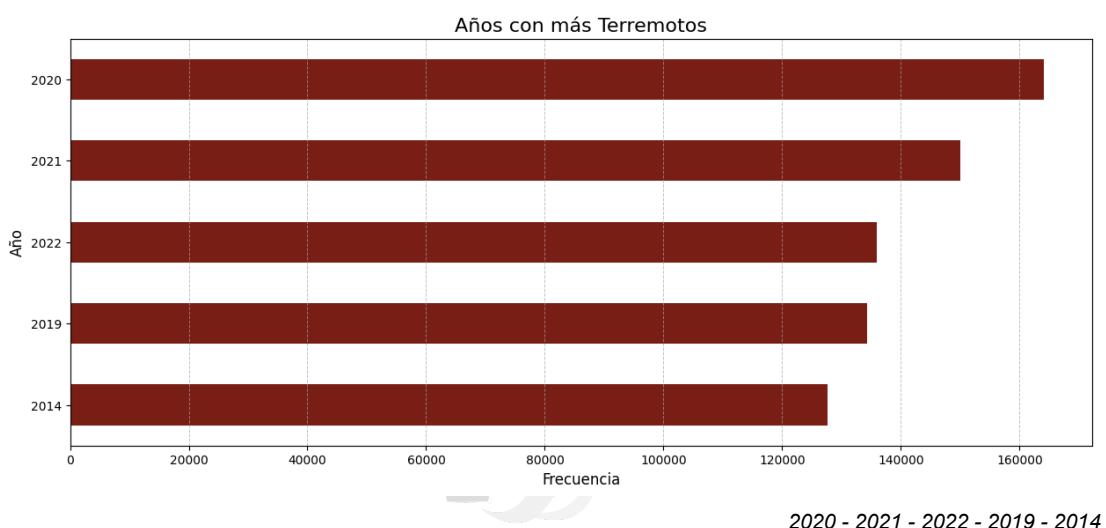
Utilizando gráficos de barras para visualizar las variables numéricas con valores únicos bajos, obtenemos:



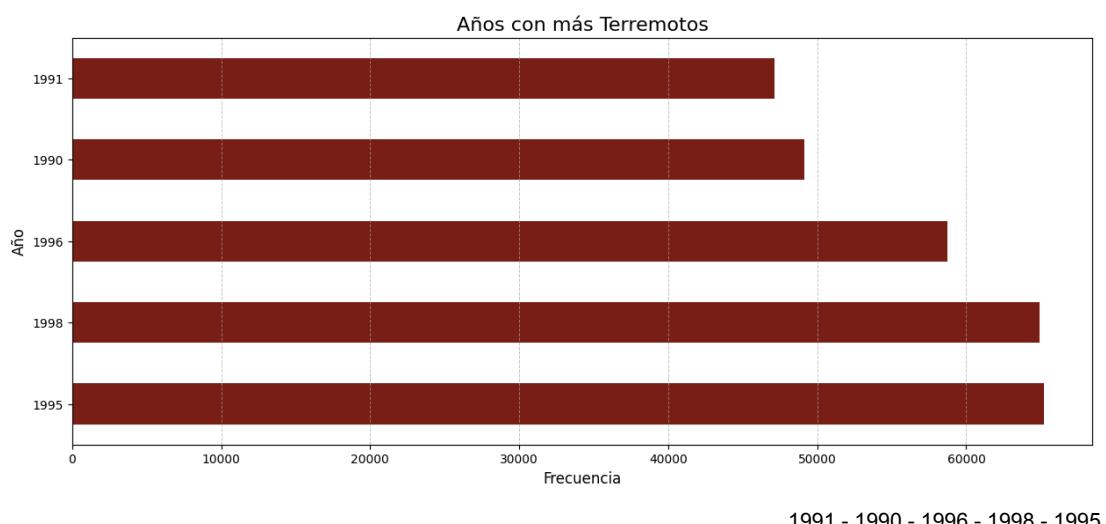
La gráfica muestra la frecuencia de los sismos por cada año; desde 1990 a 2023.

Aquí se observa una tendencia ascendente en el transcurso del tiempo. La cantidad de terremotos por año fue aumentando desde el principio, llegando a un pico en el año 2020, con más de 160,000 terremotos registrados ese año. Posterior al pico, se observa una decadencia gradual de 2020 a 2022, para después descender muy drásticamente en 2023.

Los 5 años con más terremotos registrados:

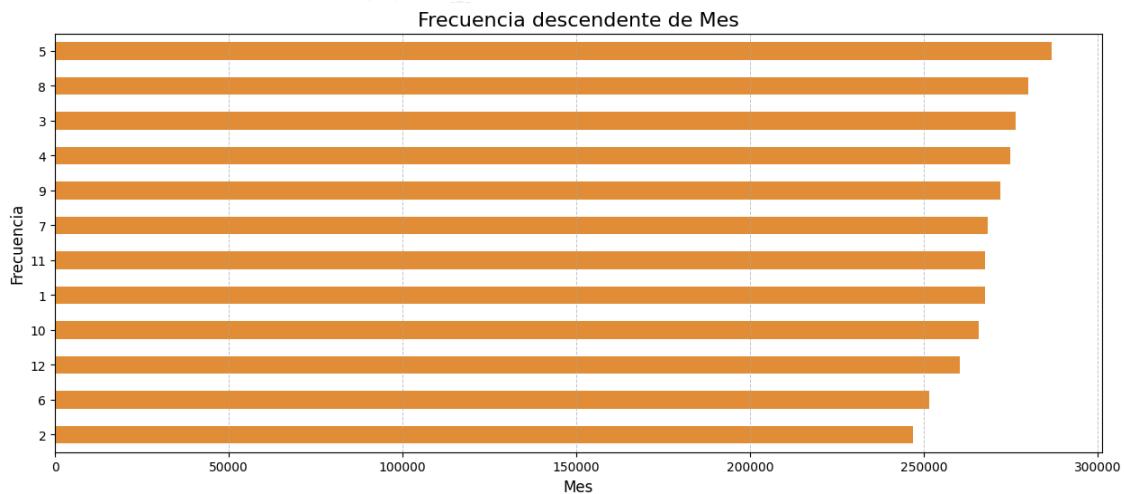
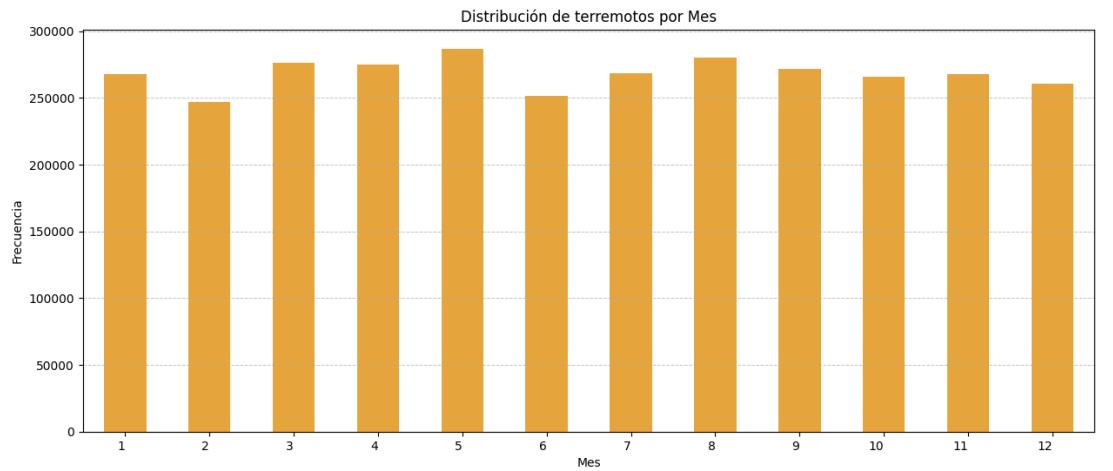


Los 5 años con menores terremotos registrados:



Si hablamos de la distribución por mes, día u hora, se observa que no existe una tendencia en específico. En todos estos casos la frecuencia de sismos es muy similar.

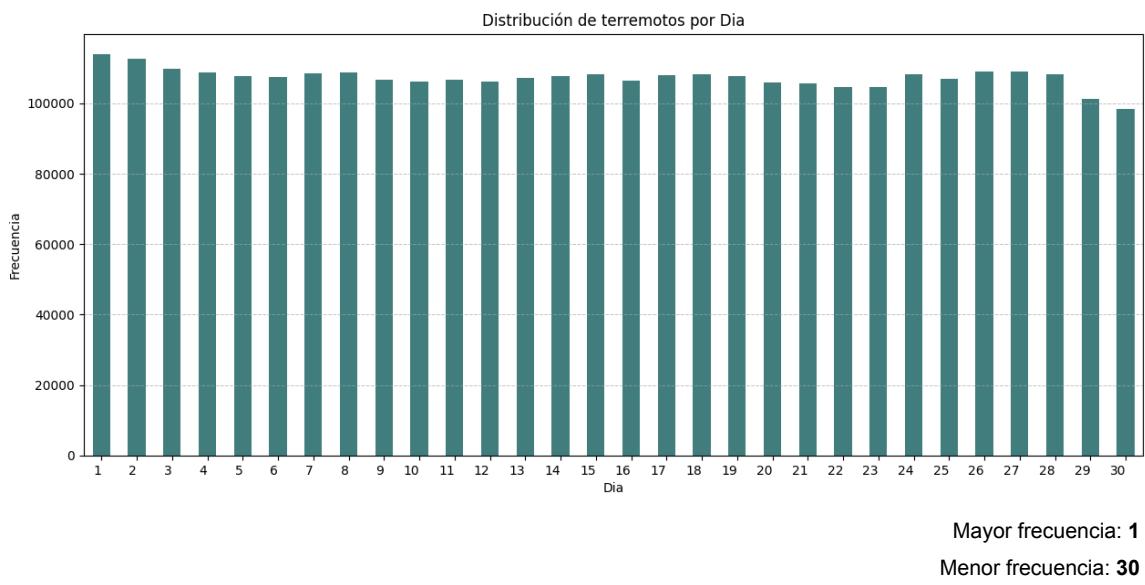
Distribución por Mes



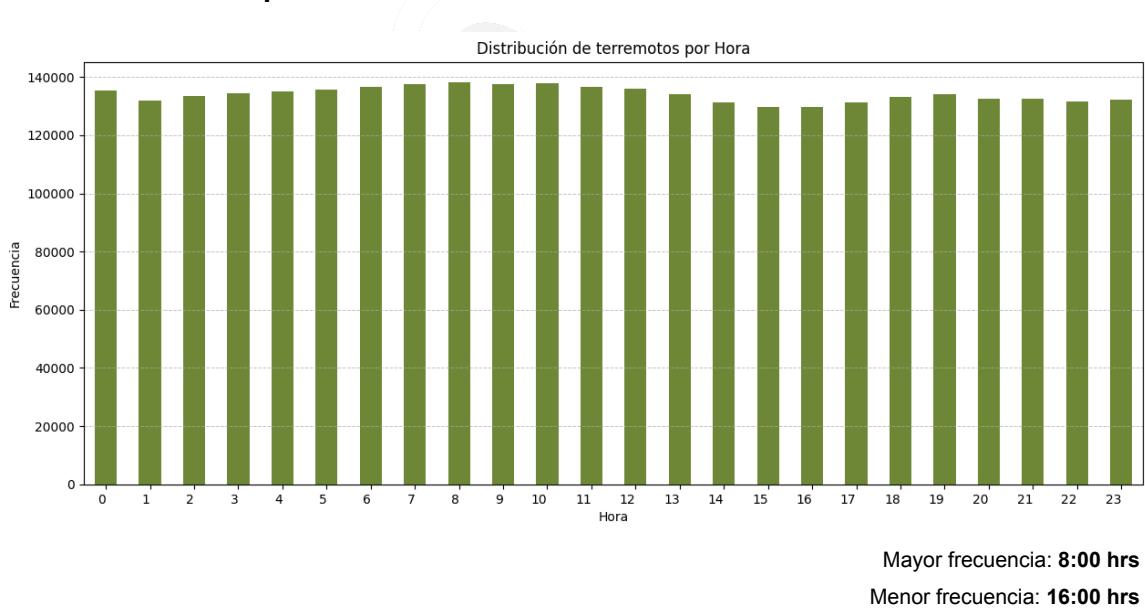
Mayor frecuencia: **Mayo**

Menor frecuencia: **Febrero**

Distribución por Dia

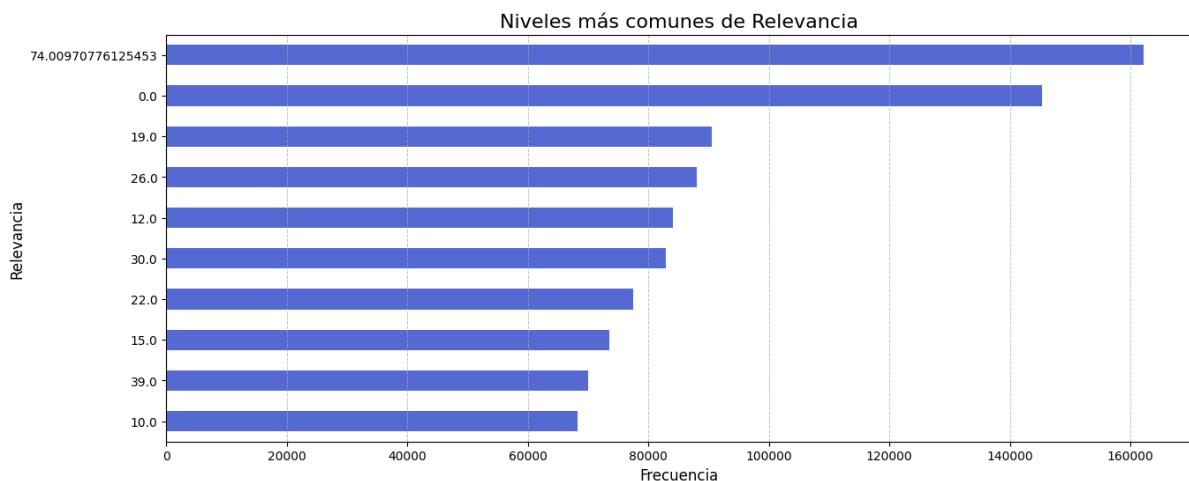


Distribución por Hora



Cambiando a las variables con alto número de valores, se utilizaron gráficos de barras para encontrar las frecuencias más y menos abundantes. Obteniendo lo siguiente:

Relevancia

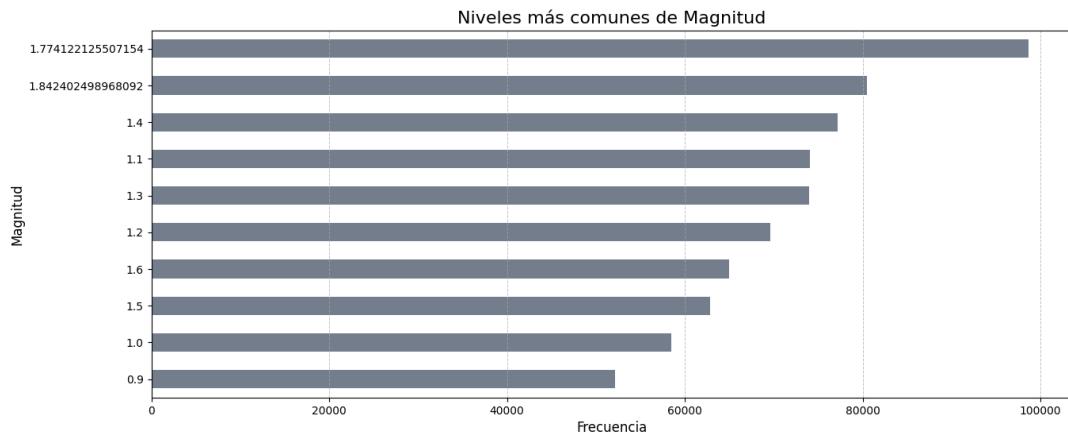


El nivel de relevancia más común es de 74, lo cual es considerable. Esto muestra que muchos terremotos han tenido un impacto, aunque no muy significativo.

Le sigue una relevancia de 0, mostrando que existe una gran cantidad de terremotos que son imperceptibles para nosotros como humanos. Estos eventos no representan ningún riesgo y ocurren frecuentemente.

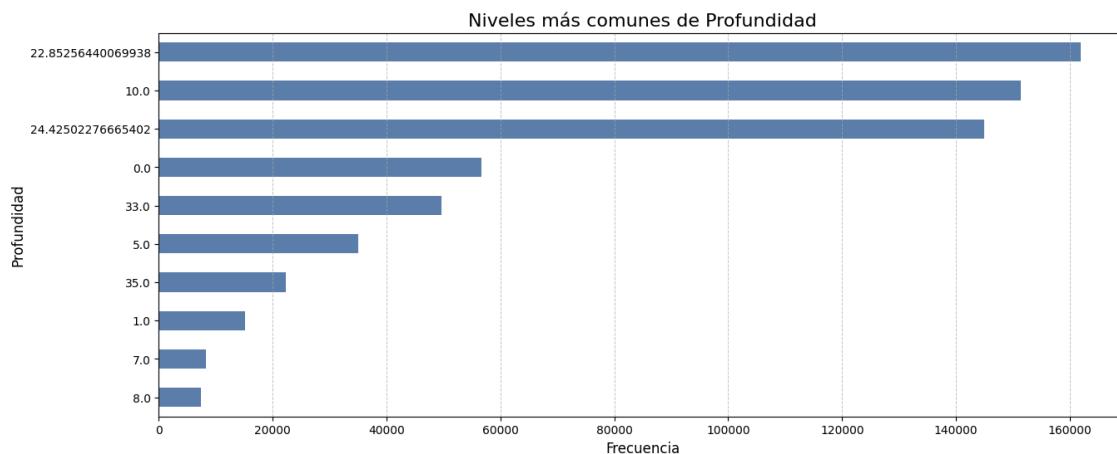
Después le siguen relevancias más pequeñas a 74, concluyendo con que la gran mayoría de terremotos son de poco impacto.

Magnitud



Los niveles de magnitud más frecuentes se encuentran alrededor de 1 en la escala de Richter; con 1.77 siendo el más abundante con casi 100,000 sismos. Esto demuestra que la gran mayoría de los terremotos que ocurren, cuentan con muy poca magnitud, y son en su mayoría, inofensivos.

Profundidad



En cuanto a la profundidad, no existe una tendencia principal. Se observa que más de 160,000 eventos se dan a 22.8 km de la superficie terrestre, seguido de 10 km y 24.4, ambos superando los 140,000 eventos.

Lo interesante es la gran cantidad de sismos a los 0 km, estando en 4ta posición con casi 60,000 eventos.

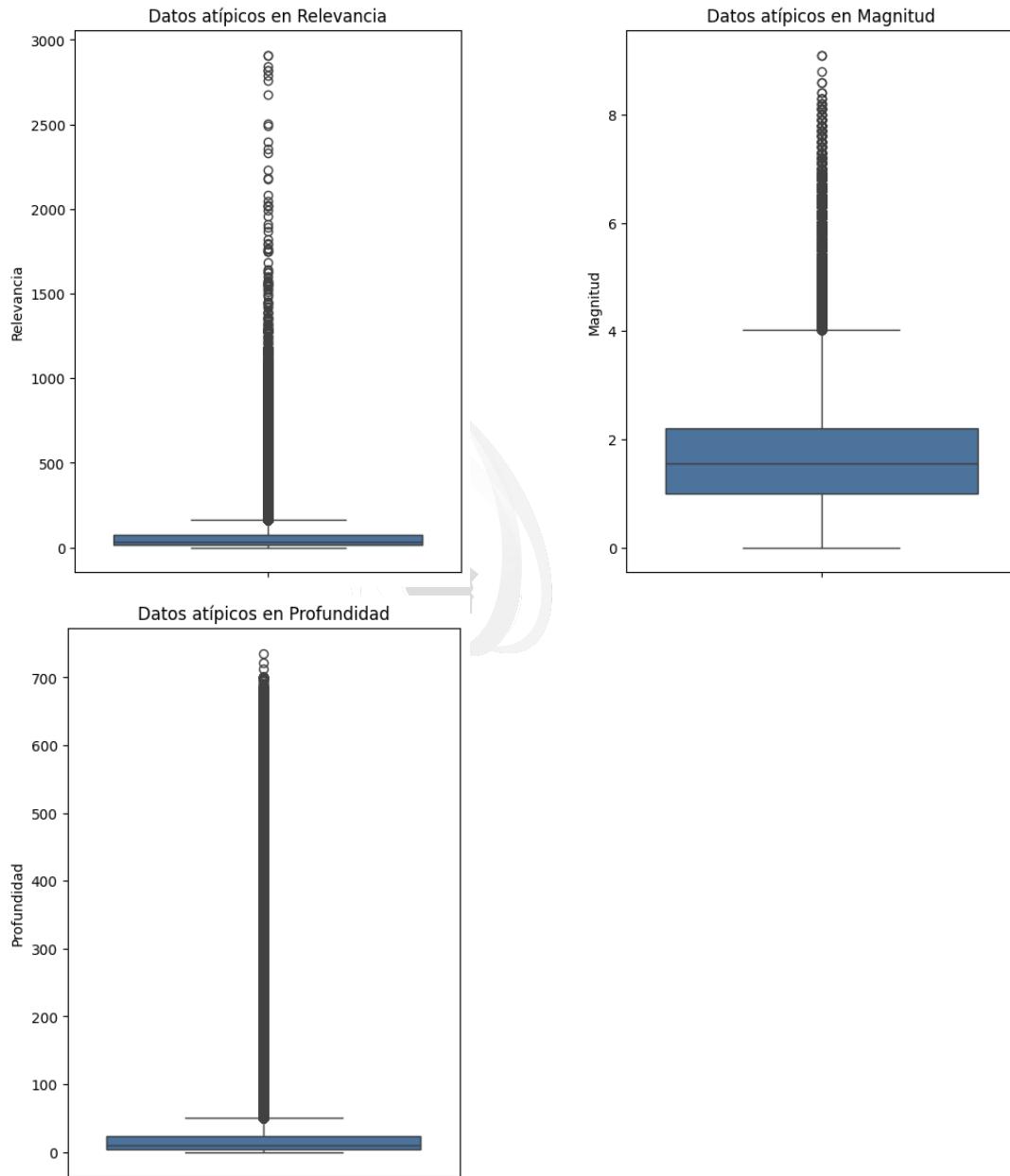
Esto indica 2 cosas: puede tratarse de un sismo muy desastroso debido a su proximidad con la superficie, o bien puede ser el resultado de acciones humanas como explosiones lo que causa un sismo en esa profundidad.

- **Valores Atípicos**

Para identificar los valores atípicos de las variables numéricas se utilizaron boxplots. Solo se analizaron aquellas variables que seguramente tendrían este tipo de valores, por lo tanto el año, mes, día y hora no son tomadas en cuenta.

Además las variables de coordenadas no se analizaron con boxplot ya que es mejor un análisis correlacionado entre ambas, para conseguir información más clara.

Los resultados de las variables numéricas restantes son los siguientes:

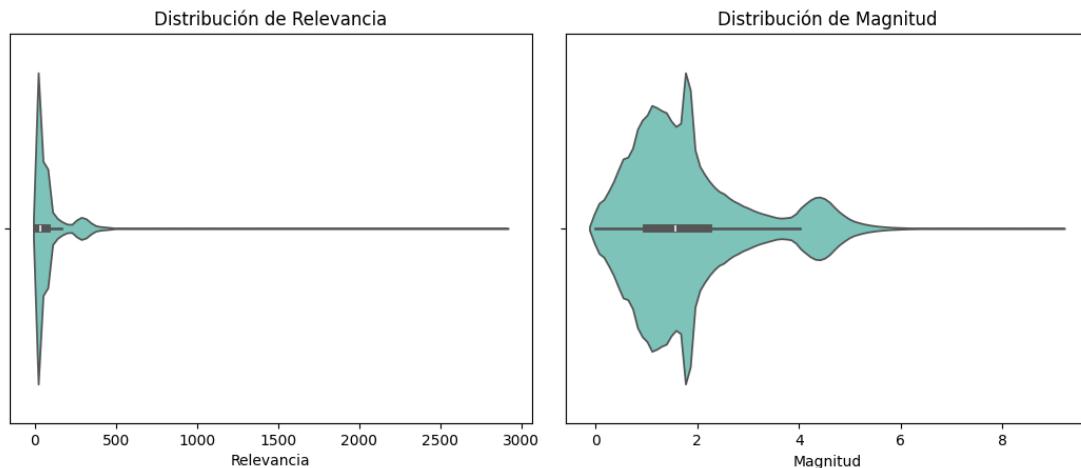


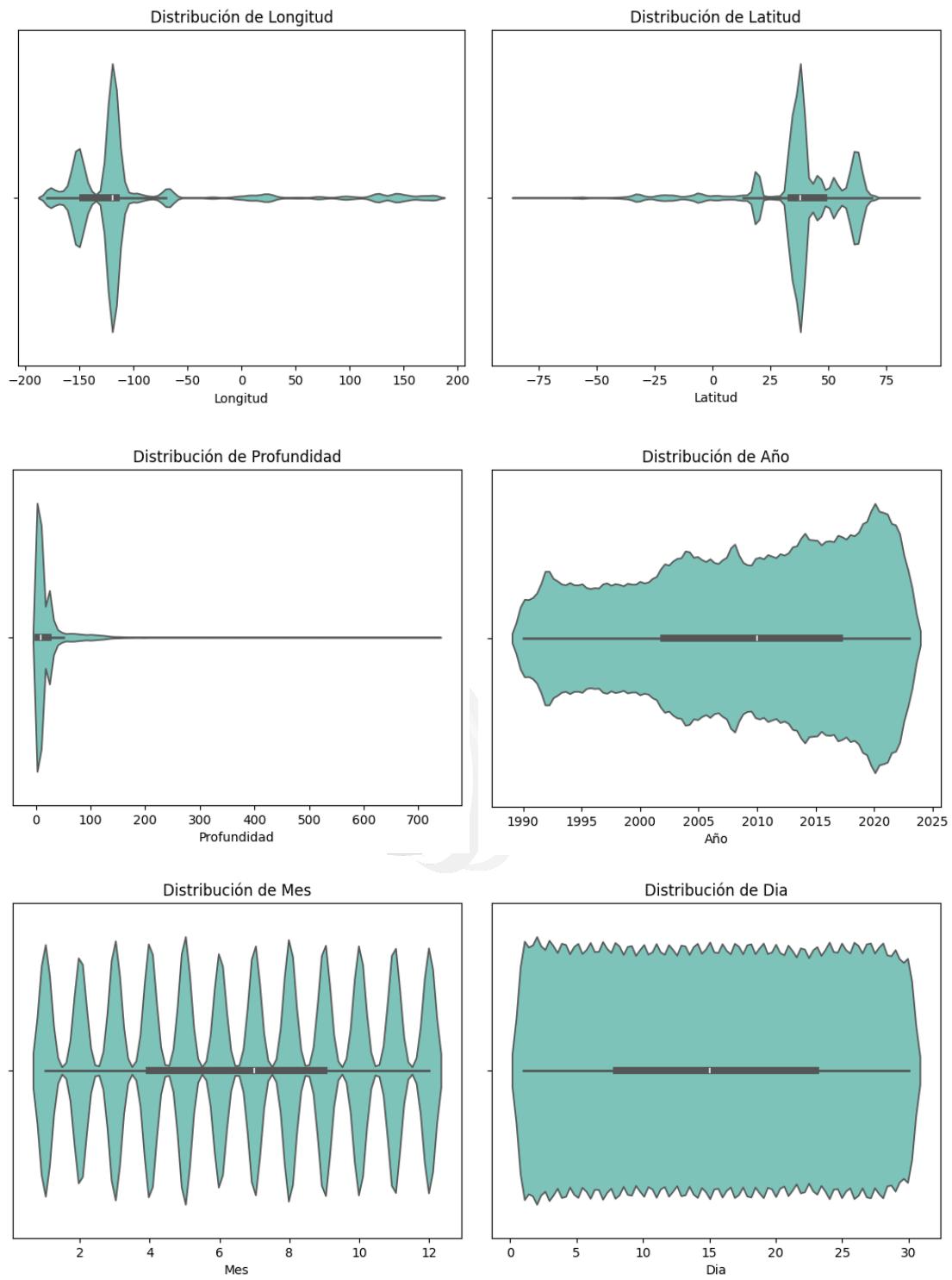
Se puede observar que tanto la relevancia, como la magnitud y la profundidad, cuentan con una gran cantidad de valores fuera de lo esperado. Sin embargo, no hay nada de qué preocuparse en este caso, ya que esto ocurre por la naturaleza de las variables analizadas.

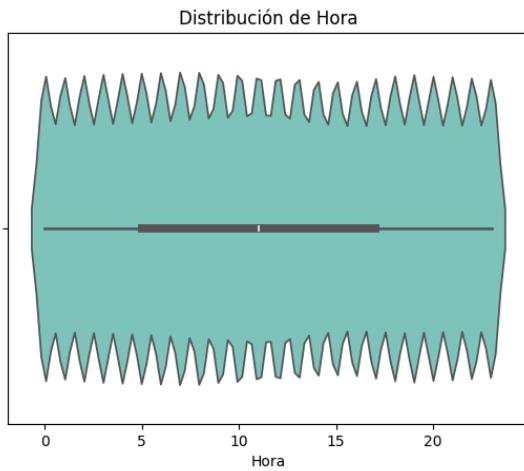
Verdaderamente no hay nada en el análisis de esta base de datos, que impida la existencia de valores atípicos. Lo que se muestra fuera de lo normal simplemente son eventos que han sobresalido gracias a lo diferente de una de sus frecuencias.

Los outliers de relevancia muestran terremotos más relevantes que la norma. Asimismo, en magnitud muestra los sismos más significativos y con mayor fuerza. Finalmente la profundidad muestra gran cantidad de eventos que ocurrieron a grandes profundidades, las cuales no son tan comunes.

Para finalizar la visualización de las variables numéricas, se presentan todas en su diagrama de violín:

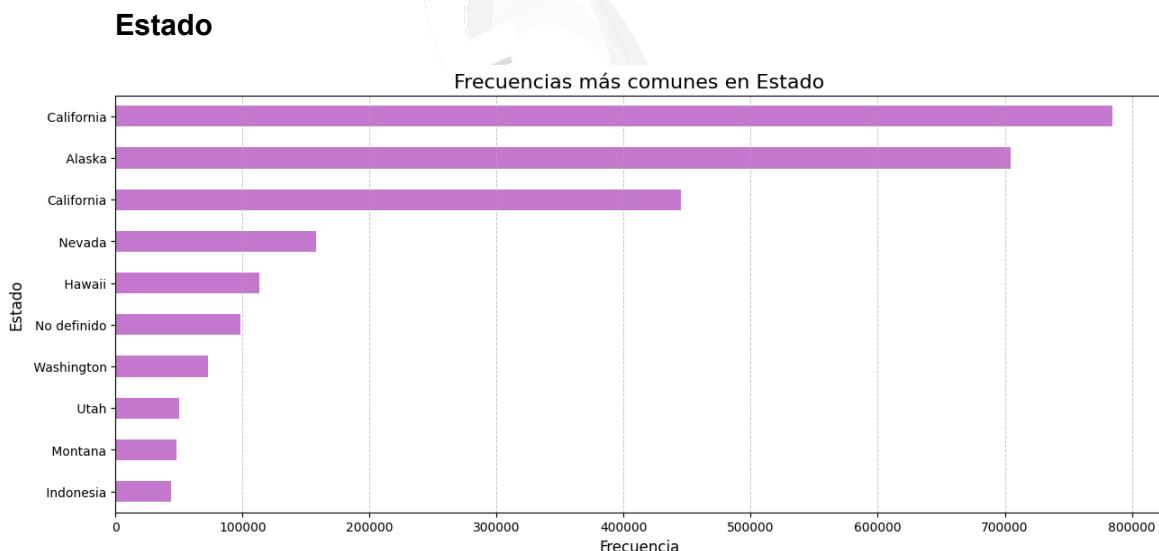






- **Variables Categóricas**

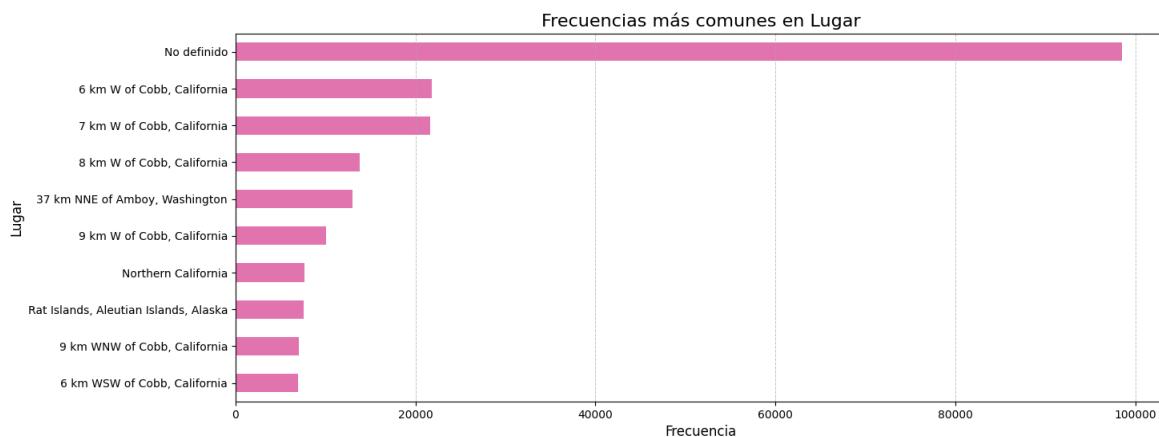
Para las variables categóricas se utilizaron gráficos de barras y los resultados fueron los siguientes:



Podemos observar que California lidera ampliamente, siendo el estado con más terremotos registrados con alrededor de 1,200,000.

Seguido de Alaska, Nevada y demás estados que pertenecen a EUA; concluyendo que es en este país donde más terremotos se presentaron en los últimos años.

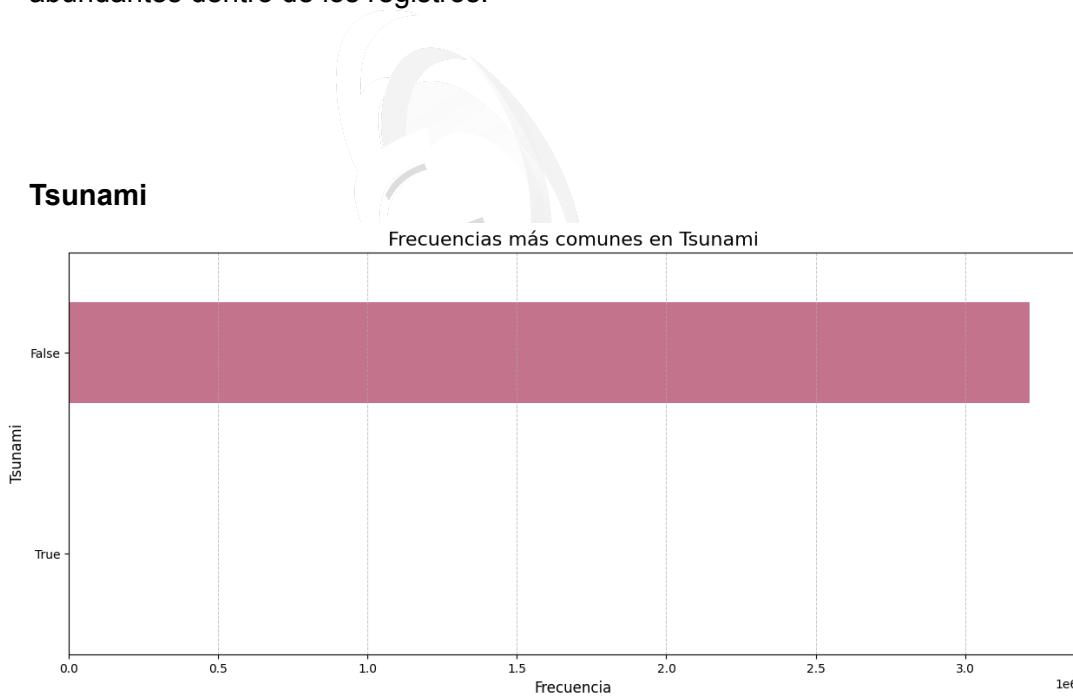
Lugar



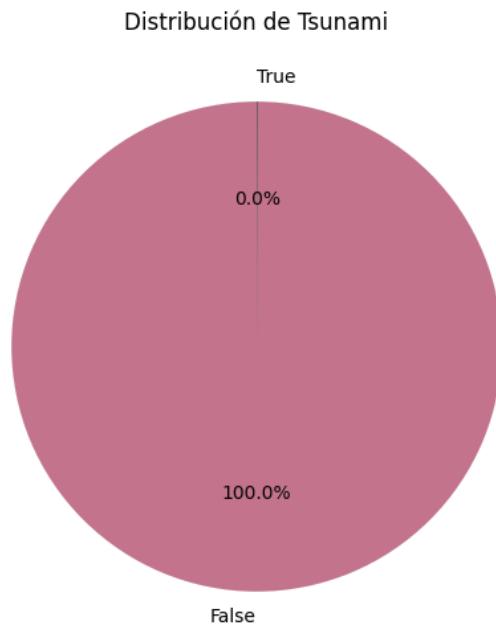
Con el lugar vemos que muchos son los terremotos que no fueron registrados con una localización específica, al menos en forma de texto.

Aunque de nuevo se observa que las localizaciones en California son las más abundantes dentro de los registros.

Tsunami



Visualizado más claramente en un gráfico de pastel, obtenemos:



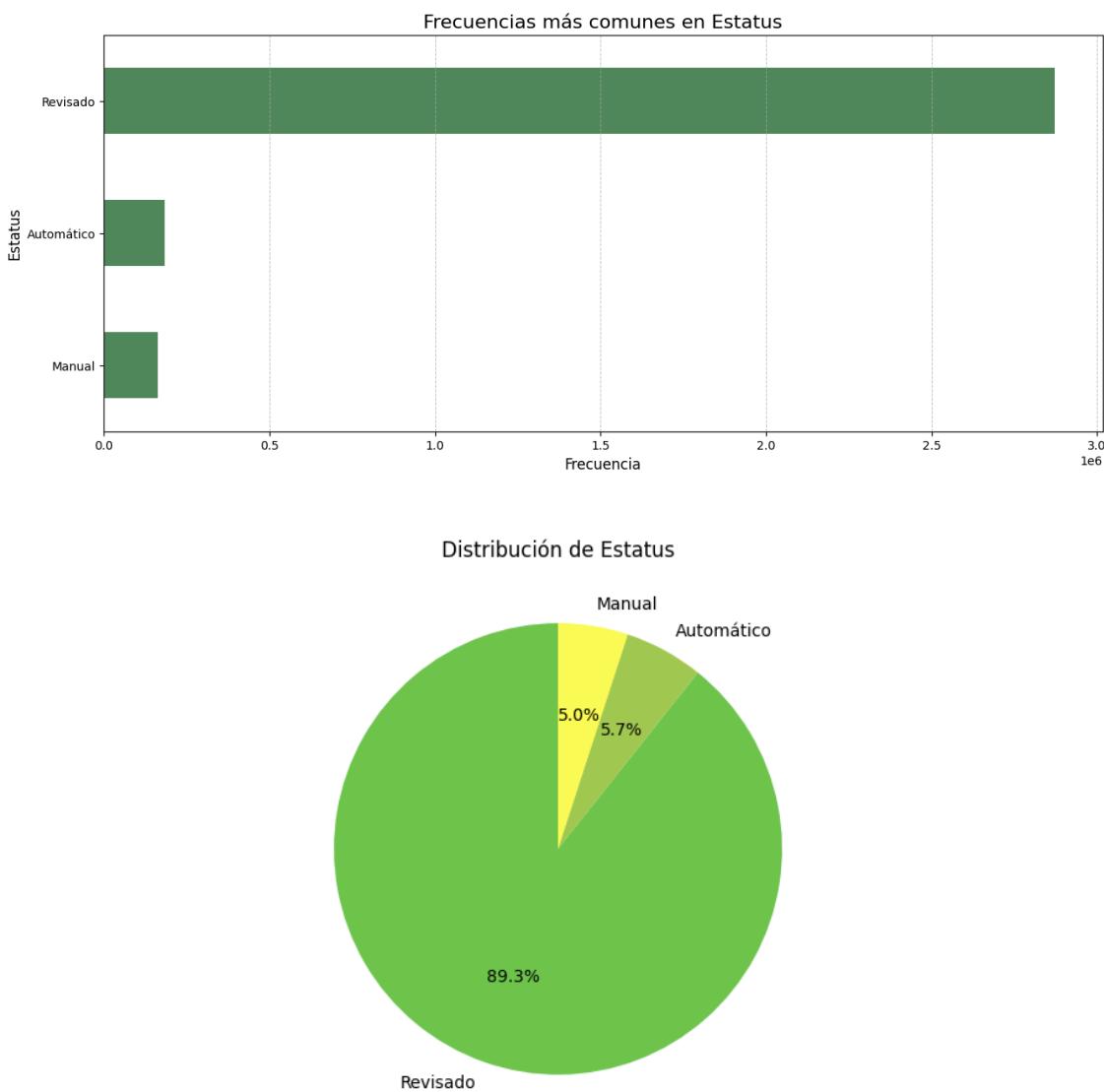
La gráfica muestra que el 100% de los terremotos no presentaron tsunamis posteriores. Aunque puede que la cantidad de estos sea tan mínima que representa menos de 1.0%.

Verificando con python obtenemos que existen 1376 sismos que provocaron un tsunami, lo cual representa el 0.0428% del total, y la razón de porque aparenta ser un 0.0% en las gráficas.

```
# Numero de Tsunamis
df['Tsunami'].value_counts()[True]
✓ 0.0s
np.int64(1376)
```

Se concluye que solo unos pocos sismos han generado un tsunami en los últimos años.

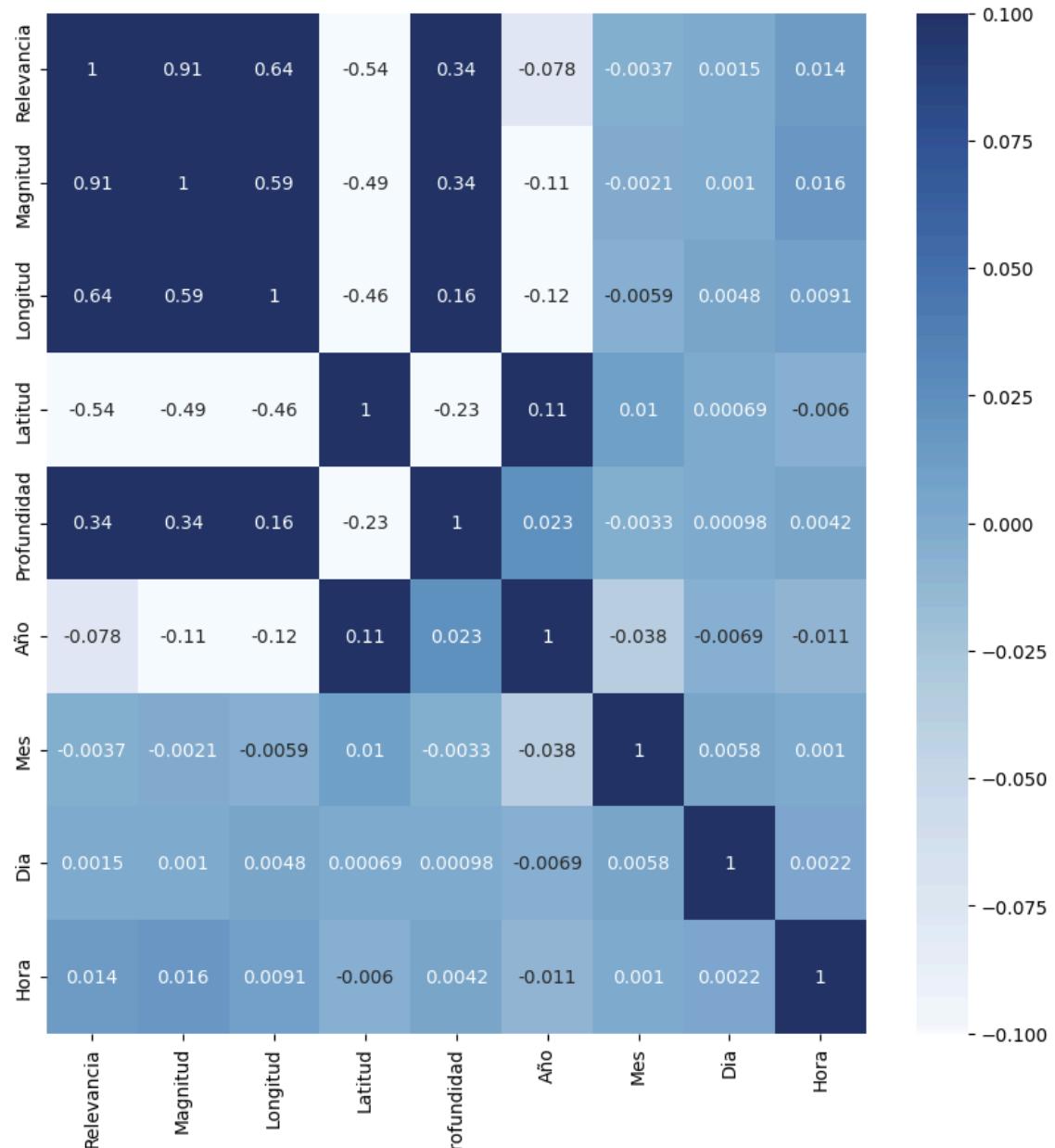
Estatus



El 89.3% de los terremotos fueron previamente revisados, con el 5.7% registrados automáticamente y el 5.0% registrado manualmente.

3. Correlación entre Variables

Se diseñó una matriz de correlación para identificar la relación de las variables numéricas y estos fueron los resultados:

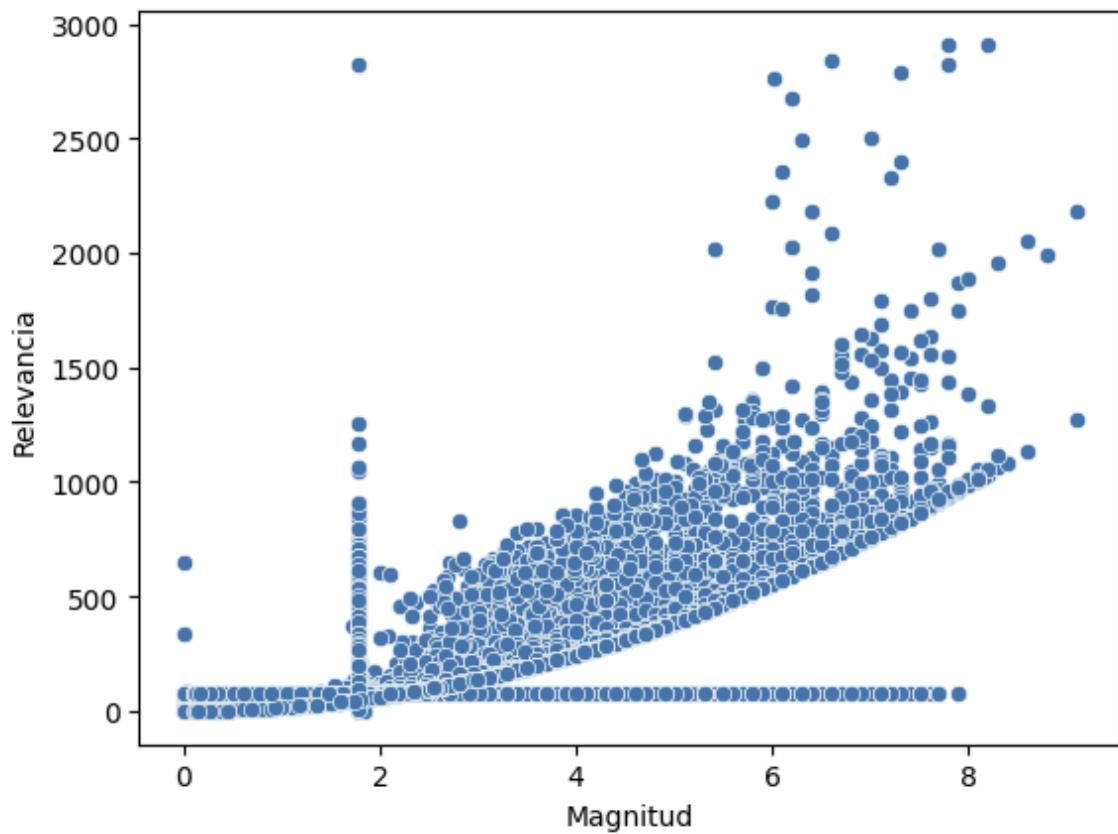


Se observa una relación importante en las siguientes variables:

Magnitud	Relevancia
Longitud	Relevancia
Longitud	Magnitud
Profundidad	Relevancia
Profundidad	Magnitud

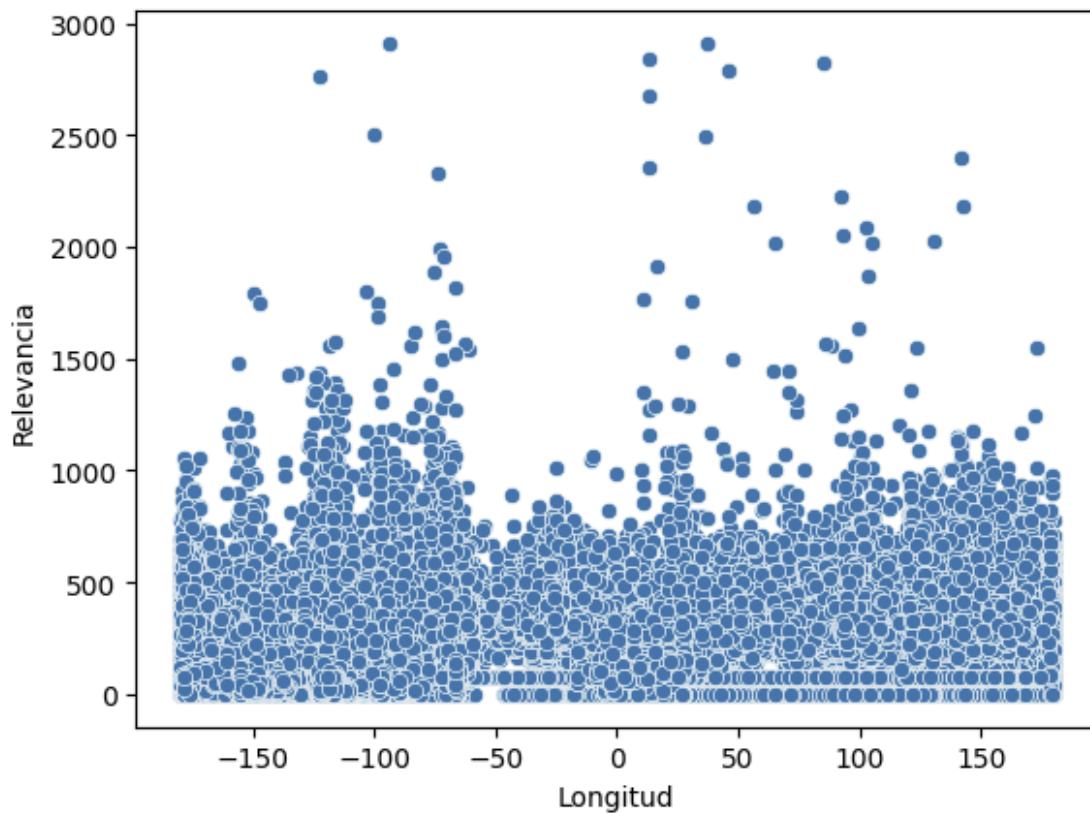
Estas relaciones fueron tomadas en cuenta para la realización de diagramas de dispersión con los cuales se visualizan la relación de ambas variables. Los resultados fueron los siguientes:

Magnitud - Relevancia



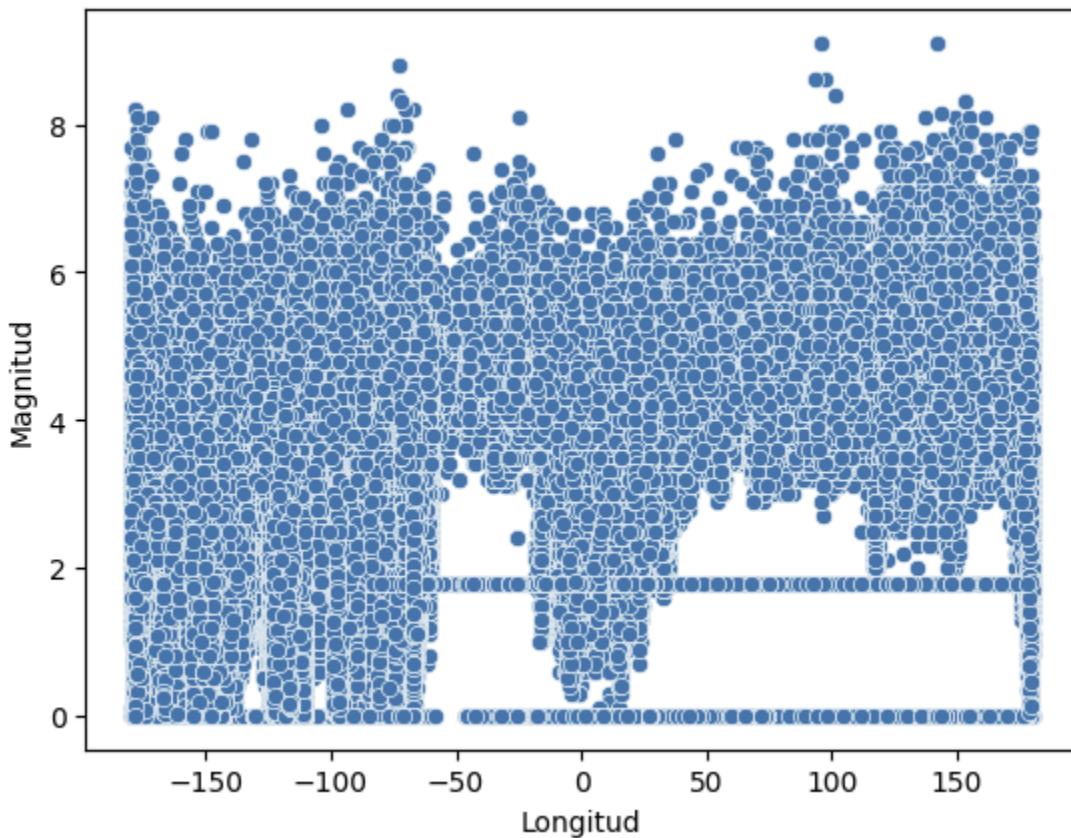
La relación muestra una tendencia donde se concluye que, entre mayor magnitud, mayor es la relevancia del terremoto.

Longitud - Relevancia



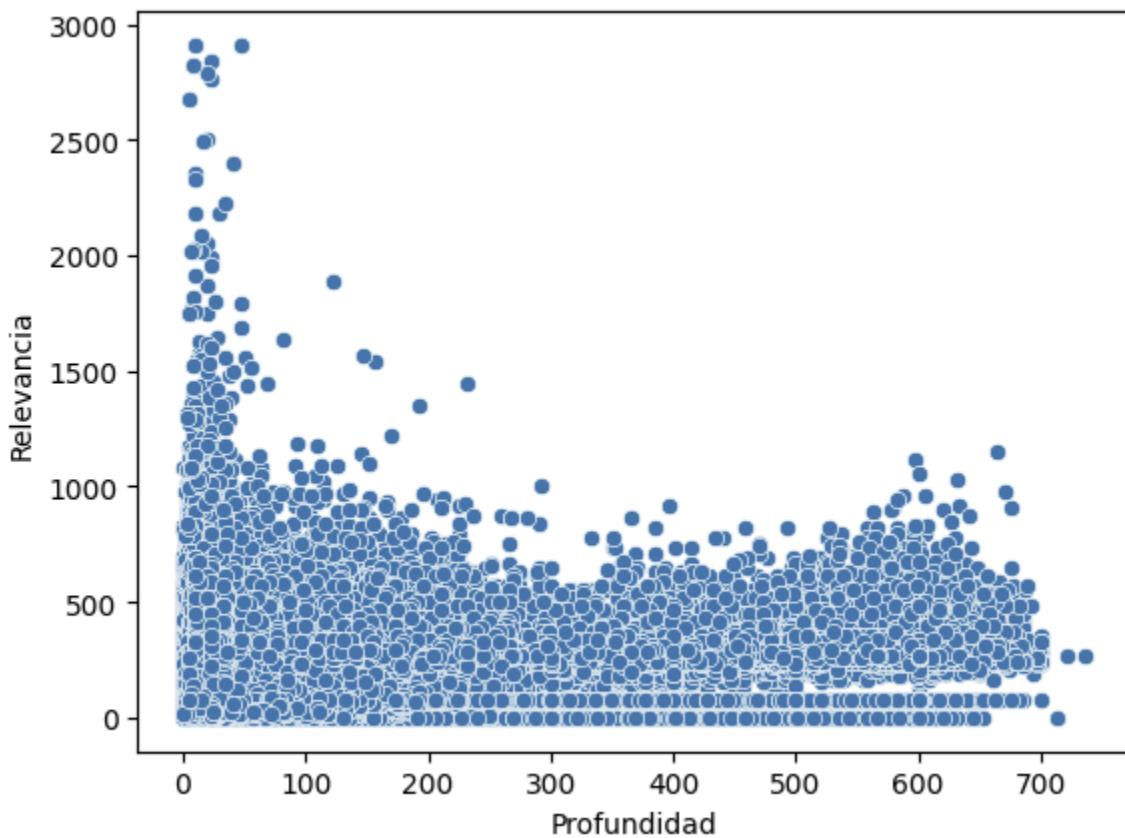
En este diagrama no se logró identificar una tendencia en particular. Los datos están distribuidos de manera similar a lo largo de cualquier coordenada de longitud.

Longitud - Magnitud



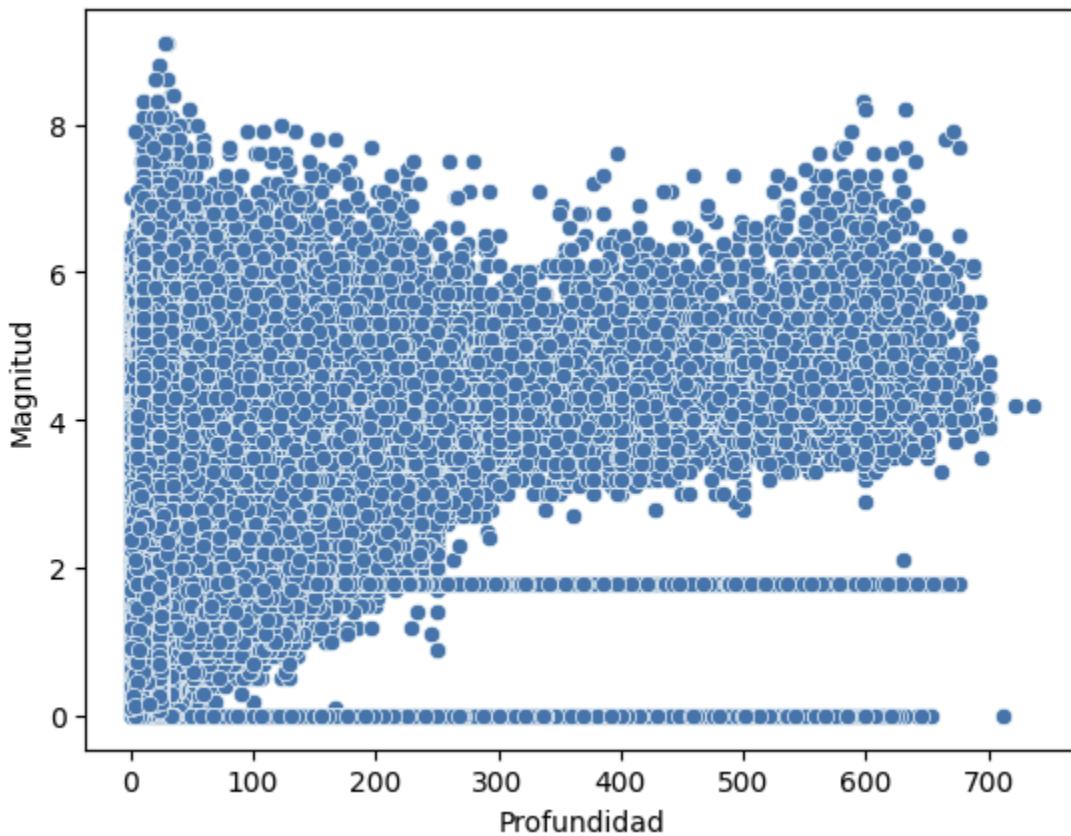
Misma situación que la anterior, no se logra observar una tendencia dominante entre la magnitud y las coordenadas de longitud.

Profundidad - Relevancia



En este caso se puede observar que, entre menor sea la profundidad de un terremoto, este tiende a aumentar su nivel de relevancia. Esto se debe a que, al no ser mucha la profundidad del sismo, la fuerza de este no se disipa lo suficiente antes de llegar a la superficie, lo que causa terremotos más fuertes y por tanto, más relevantes.

Profundidad - Magnitud



Aquí no hay una tendencia dominante, pero podemos llegar a conclusiones.

Se observa que los terremotos de mayor magnitud tienden a ocurrir a poca profundidad, pero también a mucha. También se observa que entre mayor profundidad, son menos los terremotos que tienen magnitudes pequeñas, siendo estos los que ocurren mucho a profundidades más cercanas a la superficie.

4. Análisis de Valores Atípicos

Para identificar valores atípicos en la base de datos, se utilizaron gráficas de boxplot anteriormente visualizadas. Estas gráficas tienen la capacidad de mostrar datos fuera de un rango esperado. Lo usual es que todos los datos de una frecuencia se encuentren dentro de este rango; sin embargo, la base de datos de sismos se comporta de manera diferente.

Los datos atípicos encontrados con anterioridad, no son inservibles ni tampoco son indicadores de una mala limpieza de los datos. Al ser frecuencias como magnitud, profundidad y más; siendo además de datos que abarcan todo el mundo; no existe parámetro que limite la variabilidad de los datos para su relevancia.

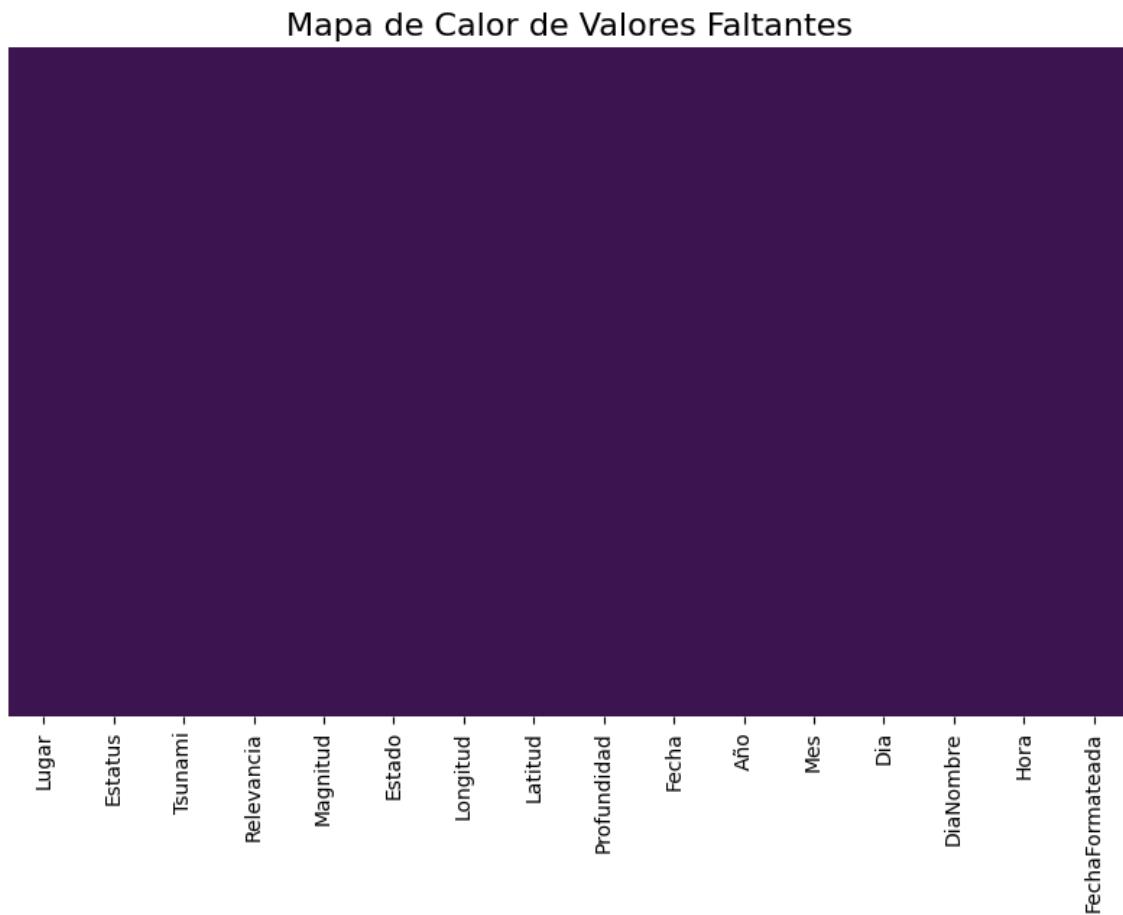
Cualquier dato es importante por más alejado que esté de lo usual, ya que esto solo quiere decir que hubo un evento más significativo que la mayoría.

Si se estudiara el mismo tema pero de manera local o con alguna condición en específico, veríamos como muchos de los datos atípicos si afectaría el resultado de lo que se busca; pero en este análisis no es el caso.

Con todo lo dicho antes, se decidió mantener los valores atípicos, sabiendo que no afectan al modelo y que son necesarios para la construcción de este.

5. Análisis de Valores Faltantes

Diseñando un mapa de calor para identificar los valores faltantes en las funciones, obtenemos lo siguiente:



La gráfica muestra un color sólido en todas las iteraciones, lo que quiere decir que no existe ningún valor nulo. Esto era de esperarse ya que todos los valores faltantes fueron manipulados en el proceso de limpieza de datos. Aquí un resumen de esto:

```
print(df.isnull().sum())
✓ 1.6s
Lugar          0
Estatus        0
Tsunami        0
Relevancia     0
Magnitud       0
Estado         0
Longitud       0
Latitud         0
Profundidad    0
Fecha          0
Año            0
Mes            0
Dia             0
DiaNombre      0
Hora            0
FechaFormatoada 0
dtype: int64
```

Valores nulos en variable ‘Lugar’

- Se le asignó el str ‘No definido’ a todas las filas que no tuvieran datos en lugar. Se tomó esta decisión ya que no es confiable colocar localizaciones falsas sin fundamentos. Además de que el lugar también viene definido en variables como ‘Estado’ o las coordenadas; es decir, no conocer el valor de esta variable no era un gran problema.

Valores nulos en variable ‘Estatus’

- Se optó por colocar el str ‘Manual’. Esto ya que forma parte de los valores únicos de variable, además de que el estatus manual hace referencia a que el sismo fue colocado de manera manual en la base de datos, cosa que se hizo al manualmente cambiar sus atributos.

Valores nulos en variable ‘Tsunami’

- Se eliminaron los NaNs con el valor booleano de False. La decisión radica en la poca probabilidad de que un terremoto causó un tsunami, cosa observada anteriormente con un 0.04% de que ocurriera. Cambiar el valor a falso es algo que no altera en lo absoluto la fiabilidad de la base.

Valores nulos en variable ‘Relevancia’

- Los datos nulos de relevancia fueron sustituidos por la media de la misma, es decir, el promedio. Se tomó esta decisión ya que es el dato que menos cambia la exactitud de la base, sin eliminar filas que pueden ser necesarias para el modelo.

Valores nulos en variable ‘Magnitud’

- Para la magnitud se tomó el mismo principio que la relevancia. Eliminado los datos nulos por el promedio de la variable. Alterando lo menos posible la fiabilidad de los datos.

Valores nulos en variable ‘Estado’

- Para el estado se tomó el mismo principio que el lugar. Reemplazando los datos nulos por un str ‘No definido’, recordando que podemos obtener la localización exacta a partir de las coordenadas.

Valores nulos en variable ‘Latitud’

- Para las coordenadas en latitud se optó por eliminar por completo las filas que contarán con valores nulos. La razón radica en que esta variable es la principal indicadora de una localización, la cual es indispensable para el modelo. Cambiar manualmente las coordenadas afecta directamente la exactitud del trabajo, dando como única opción eliminar toda fila que no cuente con esta variable.

Valores nulos en variable ‘Longitud’

- El mismo principio, al ser una variable de coordenada, se vuelve indispensable en el modelo. Cualquier fila que no contara con esta, fue eliminada para no generar errores.

Valores nulos en variable ‘Profundidad’

-

Para la profundidad se utilizó el principio de magnitud y relevancia. Utilizando la media o el promedio para llenar los valores nulos; generando el menor cambio posible a la exactitud de la base de datos.

Valores nulos en variable ‘Fecha’

-

Para las fechas se utilizó el método de forward-fill; el cual consiste en llenar los NaNs con el valor anterior más cercano que no sea nulo; es decir, si una fecha aparece nula, se sustituye por la fecha anterior que no sea nula.

Se utilizó este método ya que se sabe que los terremotos en la base de datos van en orden cronológico, lo que significa que uno ocurre después de otro de manera consecutiva. Colocando la fecha del terremoto anterior, es la manera más sencilla para generar pocos cambios significativos que dañen al modelo. El orden de los terremotos se mantiene.



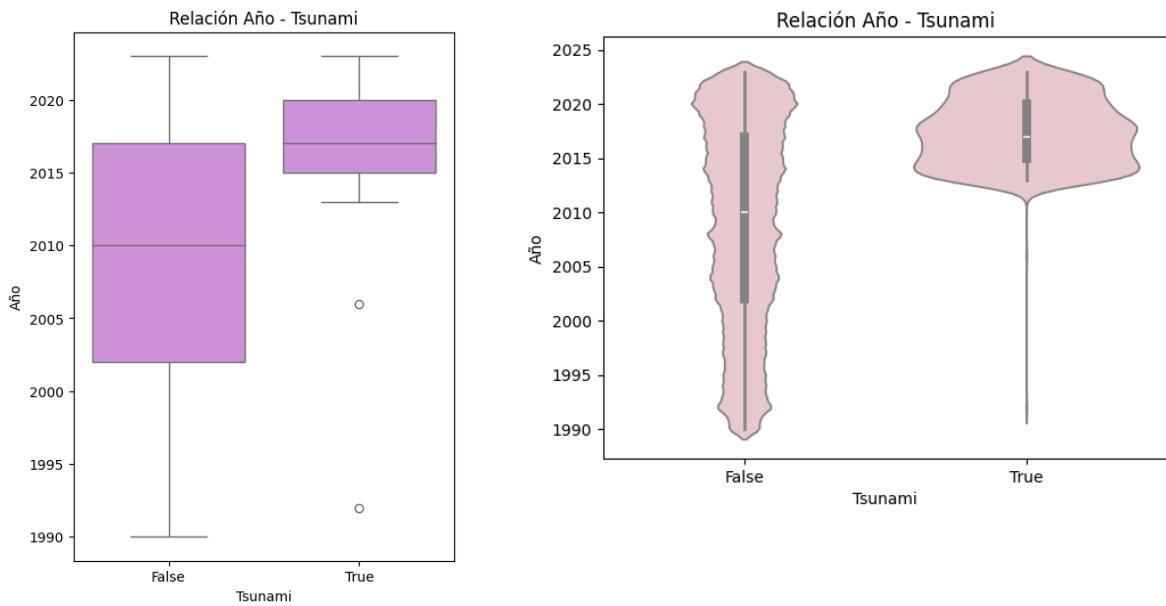
6. Relación entre Variables Categóricas y Numéricas

Recordemos las variables y sus tipos:

Variables Numéricas	Variables Categóricas
<ul style="list-style-type: none">• Relevancia• Magnitud• Longitud• Latitud• Profundidad• Fecha• Año• Mes• Dia• Hora• FechaFormatteada	<ul style="list-style-type: none">• Lugar• Estatus• Tsunami• Estado• DiaNombre

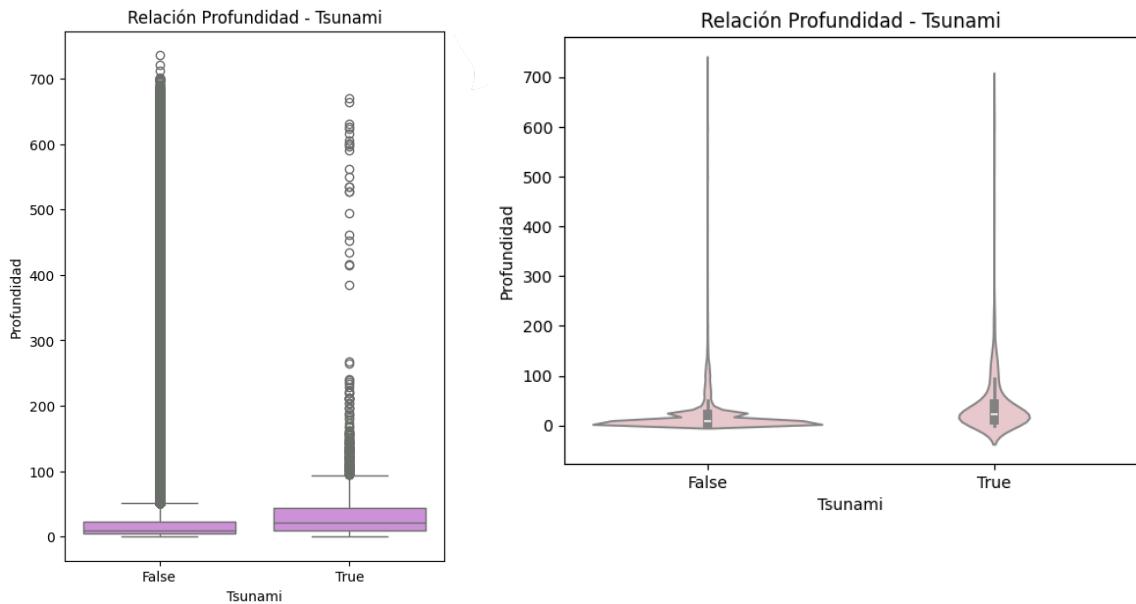
Analizando las variables y sus posibles relaciones entre sí, se generaron los siguientes gráficos:

Año - Tsunami



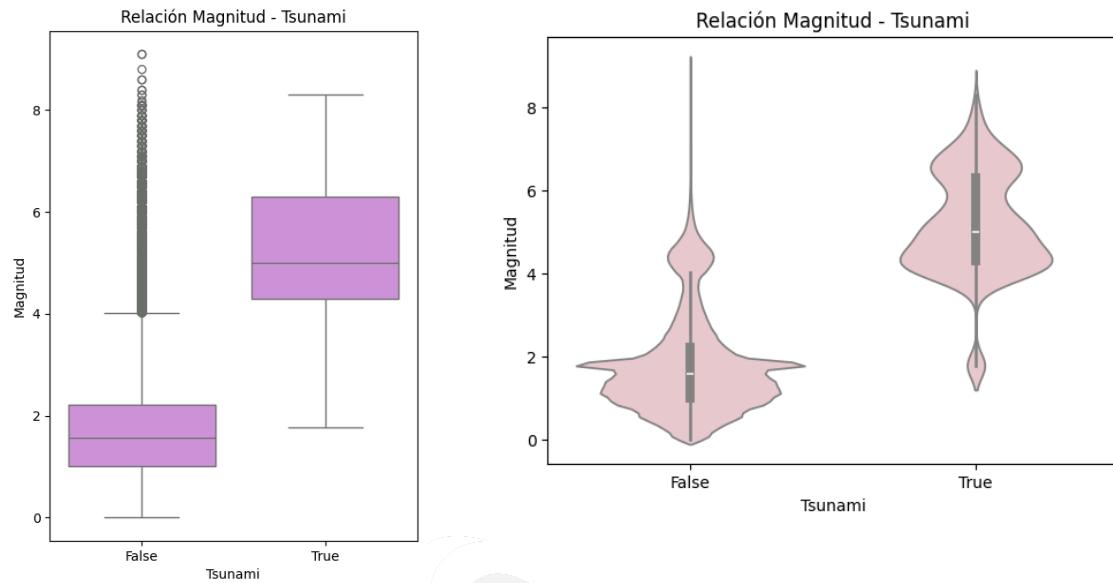
La gran parte de los tsunamis ocurrieron entre 2015 y 2020, con algunas excepciones. Los eventos sin tsunamis se distribuyen de manera natural, creciendo en los últimos años por la mayor cantidad de eventos registrados entre el 2000 y el 2020.

Profundidad - Tsunami



Los tsunamis ocurrieron en eventos donde la profundidad rondaba entre 0 - 100 km. Esto cobra sentido ya que a menos profundidad, mayor es el impacto, aunque también depende la localización.

Magnitud - Tsunami



A mayor magnitud, mayor probabilidad de un tsunami. Las gráficas muestran a los tsunamis aumentar, cuando la magnitud es superior a 4.

7. Observaciones y Hallazgos Importantes

El objetivo del EDA consiste en encontrar las tendencias que los terremotos tuvieron en los últimos años, para así poder dar una respuesta de lo que podrían hacer en el futuro.

Tendencias a través del tiempo

- Se descubrió que la cantidad de terremotos aumentaba con cada año, llegando a un pico en 2020 con más de 160,000 terremotos. Posterior a ello, la abundancia de eventos ha disminuido drásticamente.

Si se habla de la frecuencia por mes, día u hora; se observa que es muy constante y no hay una tendencia en particular.

Tendencias de localización

- La mayoría de los terremotos que se registraron, ocurrieron en el país de Estados Unidos (principalmente California y Alaska), seguido de Indonesia.

Los terremotos tienden a ocurrir en estas zonas debido a las capas tectónicas de Norte América y la India. Estas parecen ser las de mayor actividad recientemente.

Tendencias de magnitud

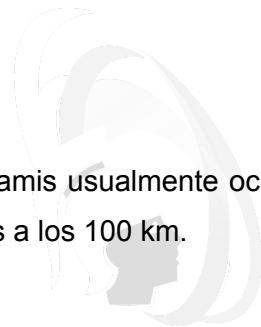
-

Las magnitudes de los sismos más frecuentes rondan entre los 0.0 a los 2.0. Concluyendo que la mayoría de los terremotos son insignificantes y pocos son los que causan un gran impacto.

Tendencias de tsunamis

-

Como era esperado, los tsunamis usualmente ocurren en terremotos de gran magnitud, y a profundidades no mayores a los 100 km.



Tendencias de profundidad

-

Las profundidades más comunes rondan entre los 0km a los 35 km; con 22 km siendo la más común de todas. Esto indica que no tienden a surgir a grandes profundidades, más bien, ocurren relativamente cerca de la superficie.

Tendencias de relevancia

-

La relevancia más común es de 74, seguido de 0 y menores al primero son las posteriores. Concluyendo que la mayoría de los terremotos tiene una relevancia baja y no impactan mucho a la vida humana.

CONCLUSIONES FINALES DEL EDA

- La relevancia de un sismo está conectada directamente con su magnitud y profundidad. A mayor magnitud y menor profundidad, mayor es el impacto. Esto puede venir acompañado de un tsunami, que aumenta aún más la relevancia y son causados usualmente por terremotos de gran magnitud.

El número de sismos está en un momento de receso, con tendencias a disminuir. Sin embargo también con el pasar de los años, han sido más y más los terremotos que han tenido mayor impacto.

Los lugares con mayor número de terremotos son: EUA, Chile, Europa central y oriental, Indonesia, Malasia y demás países vecinos.



Modelo de Machine Learning

DISCLAIMER

El modelo de este proyecto no pudo llevarse a cabo debido a la complejidad del mismo, sumado a las incapacidades de hardware con las que se tenía contadas para trabajar.

Enfoque de modelado

Se selecciona el tipo de modelo más adecuado según los datos y el objetivo del proyecto. Para este proyecto existen tres enfoques principales.

- Series temporales:
 - Modelos ARIMA, SARIMA o Prophet; útiles para predecir magnitudes futuras o frecuencias de terremotos a lo largo del tiempo.
 - Redes neuronales LSTM o GRU, las cuales pueden capturar patrones más complejos en datos temporales.
- Modelos probabilísticos:
 - Naive Bayes o regresión logística son capaces de estimar la probabilidad de que ocurra un terremoto en una región o en un intervalo de tiempo.
- Aprendizaje supervisado:
 - Para clasificar regiones de alto o bajo riesgo, se usan algoritmos como Random Forest, XG Boost o redes neuronales.

Evaluación del modelo

Se entrena el modelo para poder medir su desempeño usando métricas específicas. La evaluación permite comparar los distintos modelos y ajustar sus parámetros.

Los datos se dividen en entrenamiento (70-80%), para ajustar los parámetros del modelo; y prueba (20-30%); para evaluar su desempeño en datos no vistos.

También se seleccionan métricas según el tipo de modelo. Para probabilidades se usan log loss o curvas ROC-AUC. Para predicción de magnitudes o frecuencias, se usan métricas como error cuadrático medio (MSE) o R². Para la clasificación se analiza la precisión, recall, y F1-score.

Validación del modelo

Este paso asegura que las predicciones generadas por el modelo tengan sentido. Es importante comprobar que el modelo no esté sobreajustado a los datos históricos.

Se comparan las predicciones con eventos históricos y se hacen las siguientes preguntas.

- ¿El modelo predice bien las zonas de alta actividad sísmica?
- ¿Los patrones temporales estimados coinciden con lo observado?

Otra forma sería consultar directamente a expertos o estudios geológicos para validar los resultados generados por el modelo. O bien, realizar simulaciones con diferentes configuraciones para evaluar la fiabilidad del modelo.

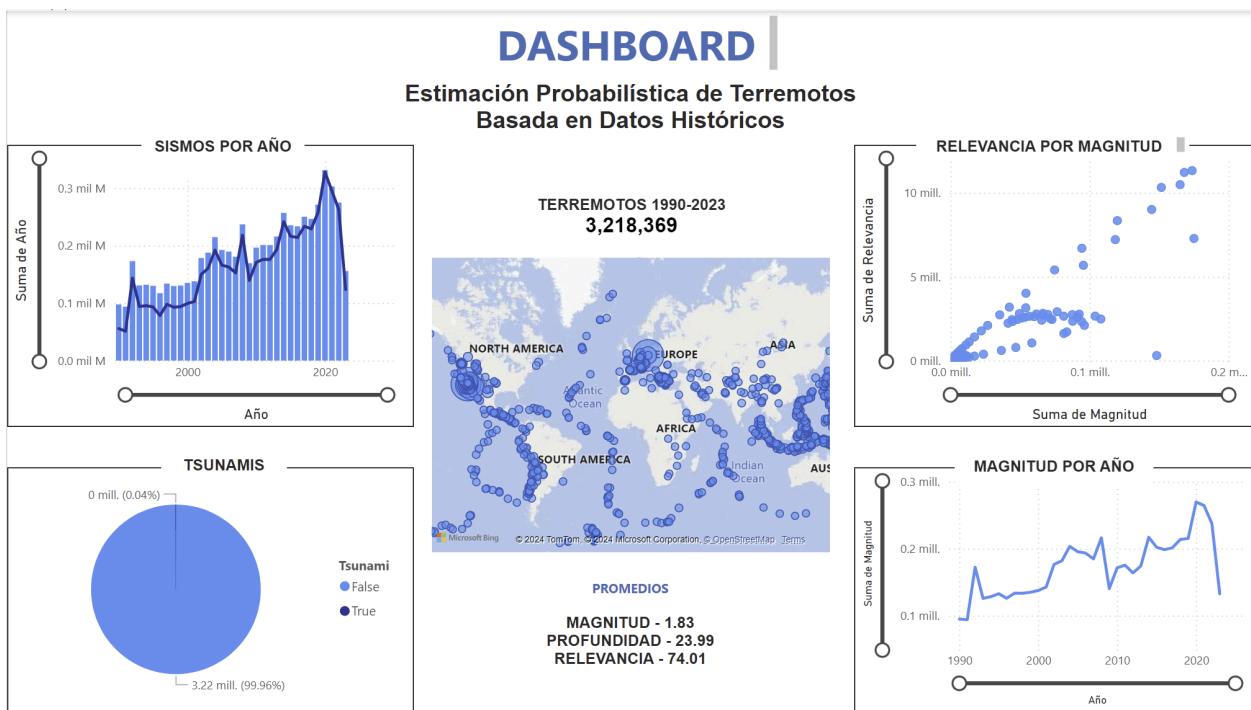
Herramientas

Para todo este proceso, es necesario contar con ciertas herramientas que nos permitan llegar a los resultados esperados por el modelo:

- Pandas y Numpy para procesamiento y análisis.
- Matplotlib y Seaborn para gráficos simples.
- Plotly para visualizaciones interactivas, como mapas 3D o gráficos dinámicos.
- Scikit-learn para modelos básicos y métricas.
- Tensorflow o Pytorch para redes neuronales.
- Statsmodels para series temporales.
- Geopandas para manejar datos geoespaciales.
- Folium para crear mapas interactivos.



Dashboard



Conclusiones y Futuras Líneas de Trabajo

Con el análisis realizado podemos concluir que EUA, Sudamérica, Europa Central, Indonesia y Malasia son los lugares más propensos a sufrir terremotos en el futuro. Son estos los países que necesitan implementar más medidas de prevención para evitar un posible desastre.

Los gobiernos deben educar a su población acerca de estos fenómenos naturales, como actuar en caso de uno y qué hacer al ocurrir un evento. Las infraestructuras tienen que ser acorde a la gran actividad sísmica que se presenta; esto con el fin de evitar desgracias.

También son estos los lugares que los científicos pueden estudiar para comprender aún más estos eventos y así estar cada vez más cerca de poder predecir con mucha antelación.

La realización del modelo se vería en gran ventaja si se tuviera una máquina capaz de analizar muchos datos en poco tiempo. Asimismo la base de datos es en su gran mayoría un acierto, pero no está de más poder conseguir datos que se acerquen a la perfección para poder dar una conclusión más verídica.

En conclusión, los sismos ocurren muy frecuentemente a lo largo de todo el mundo. Aunque sus magnitudes en su mayoría son muy pequeñas, en los últimos años se ha visto un aumento de terremotos y por tanto, el aumento de sismos cada vez más fuertes a una frecuencia mayor. Las tendencias de donde ocurren los sismos no han cambiado en los últimos años, y son las placas tectónicas de Norteamérica y la India las que más actividad presentan.

Por otro lado los tsunamis ocurren muy pocas veces, pero es necesario a los lugares cerca de las costas que estén alertas ante cualquier indicio después de un evento.

Los sismos se generan a profundidades relativamente bajas, lo que aumenta el riesgo de un evento de gran relevancia.

Referencias

Introduc. (s/f)

Lo Bello, A. (2023). All the Earthquakes Dataset : From 1990-2023 [Data set].

Anexos

Reporte Fase 1 - Descubrimiento del proyecto - SNM.pdf at main · Samolongo/Proyecto-EPTBDH. (s/f).

Reporte 2 - Limpieza de una Base de Datos Ensuciada -SNM.pdf at main · Samolongo/Proyecto-EPTBDH. (s/f).

Database Cleaning.ipynb at main · Samolongo/Proyecto-EPTBDH. (s/f).

Análisis Exploratorio de Datos.ipynb at main · Samolongo/Proyecto-EPTBDH. (s/f).

VIDEO