

“AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN
DE LA ECONOMÍA PERUANA”.



ESCUELA PROFESIONAL DE CIENCIA DE LA
COMPUTACIÓN
TÓPICOS CIENCIA DE DATOS

Data Wrangling

Estudiante:

Andrea del Rosario Lopez Condori

Docente :

ANA MARIA CUADROS
VALDIVIA



Índice

1. Introducción	2
2. Dataset	2
2.1. Objeto de Estudio	2
2.2. Descripción del Dataset	2
3. Exploración Inicial de Datos	3
3.1. Recolección de Datos	3
3.2. Inspección Inicial	4
3.2.1. Tipos de Variables en el Dataset	4
3.3. Estadísticas Descriptivas para Datos Numericos	5
3.3.1. Ubicación geográfica (<code>latitude</code> , <code>longitude</code>)	5
3.3.2. Profundidad (<code>depth</code>)	6
3.3.3. Magnitud (<code>mag</code>)	7
3.3.4. Número de estaciones (<code>nst</code>)	8
3.3.5. Errores asociados	8
3.3.6. Número de estaciones (<code>nst</code>)	10
3.3.7. Gap (<code>gap</code>)	11
3.3.8. Distancia mínima (<code>dmin</code>)	11
3.3.9. Error cuadrático medio (<code>rms</code>)	12
3.3.10. Número de estaciones para magnitud (<code>magNst</code>)	13
3.4. Estadísticas Descriptivas para Datos Textuales	14
3.4.1. Ubicación geográfica legible del epicentro del sismo(<code>place</code>)	14
3.4.2. Tipo de magnitud del evento sísmico (<code>magType</code>)	14
4. Limpieza de Datos	15
4.1. Manejo de Valores Faltantes	15
4.2. Detección y eliminación de registros duplicados	16
4.3. Corrección de Inconsistencias	16
5. Transformación de Datos	16
5.1. Conversión de Tipos de Datos	16
5.2. Creación de Nuevas Variables	16
5.3. Normalización de la columna <code>place</code>	17
6. Validación de Datos	17
6.1. Verificación de Consistencia	17
6.1.1. Verificación y corrección de la columna <code>mag</code>	17
6.2. Detección de Outliers	18
7. Conclusión	18
7.1. Limitaciones	18
8. Apéndices	19
8.1. Código Utilizado	19

1. Introducción

El presente informe tiene como objetivo documentar el proceso de data wrangling aplicado a un conjunto de datos de sismos registrados por el Servicio Geológico de los Estados Unidos (USGS) desde el año 1990 hasta 2024. El data wrangling es un paso esencial en el análisis de datos, ya que permite transformar datos crudos en información estructurada y útil para el análisis posterior.

2. Dataset

2.1. Objeto de Estudio

El objeto de estudio son los eventos sísmicos registrados por las estaciones sismológicas globales, con el fin de analizarlos y clasificar su origen (natural o artificial) en base a sus características físicas. El dataset fue obtenido desde el sitio web oficial del USGS y contiene datos tabulares estructurados en formato texto (CSV).

2.2. Descripción del Dataset

Un registro del dataset es una entidad que representa un evento sísmico individual dentro del dataset. Cada registro contiene información detallada y específica sobre un único sismo ocurrido, incluyendo datos como:

- La fecha y hora en que ocurrió el evento (`time`).
- La ubicación geográfica, expresada en latitud (`latitude`) y longitud (`longitude`).
- La profundidad a la que se originó el sismo (`depth`).
- La magnitud del evento (`mag`) y su tipo de medición (`magType`).
- Parámetros adicionales como errores asociados (`horizontalError`, `depthError`, `magError`), número de estaciones que detectaron el evento (`nst`), y el estatus o tipo de evento (`status`, `type`).

En resumen, cada registro describe un sismo particular con sus características técnicas y geográficas, permitiendo analizar la actividad sísmica con granularidad a nivel de evento.

A continuación se presenta una descripción detallada de cada columna en el dataset de sismos:

Columna	Descripción Detallada
time	La fecha y hora en que ocurrió el sismo, generalmente en formato ISO 8601.
latitude	La latitud geográfica del epicentro del sismo, medida en grados decimales.
longitude	La longitud geográfica del epicentro del sismo, medida en grados decimales.

Columna	Descripción Detallada
depth	La profundidad a la que ocurrió el sismo, generalmente medida en kilómetros.
mag	La magnitud del sismo, que cuantifica su tamaño o energía liberada.
magType	El tipo de magnitud reportada, como "Mb", "Ms", "Mw", que indica el método utilizado para calcular la magnitud.
nst	El número de estaciones sísmicas que reportaron el evento.
gap	El ángulo de azimut más grande entre estaciones sísmicas adyacentes que registraron el evento.
dmin	La distancia mínima desde la estación sísmica más cercana al epicentro del sismo, medida en kilómetros.
rms	El valor RMS (Root Mean Square) del tiempo residual del sismo, que indica la calidad del ajuste de los datos.
net	La red que reportó el sismo, como us, ci, ak.
id	Un identificador único para el evento sísmico.
updated	La fecha y hora de la última actualización del registro del sismo, generalmente en formato ISO 8601.
place	La ubicación geográfica legible del epicentro del sismo.
type	El tipo de evento sísmico, como earthquake, quarry blast.
horizontalError	El error horizontal en la ubicación del epicentro, generalmente en kilómetros.
depthError	El error en la estimación de la profundidad del sismo, generalmente en kilómetros.
magError	El error en la estimación de la magnitud del sismo.
magNst	El número de estaciones utilizadas para calcular la magnitud del sismo.
status	El estado del registro del sismo, como automatic, reviewed.
locationSource	La fuente de la información de ubicación del sismo.
magSource	La fuente de la información de magnitud del sismo.

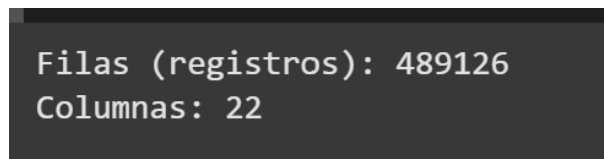
3. Exploración Inicial de Datos

3.1. Recolección de Datos

Los datos fueron obtenidos directamente del sitio web del USGS, donde se encuentran disponibles para su descarga en formato CSV. El archivo contiene registros de sismos a nivel global durante el período mencionado.

3.2. Inspección Inicial

El dataset original tras la carga inicial, se contabilizaron 489 126 filas y 22 columna, sin ningún tipo de preprocesamiento aplicado (es decir, en su estado crudo o sin limpiar). Las columnas incluyen información relevante para el análisis sísmico, tales como `time`, `latitude`, `longitude`, `depth`, `mag`, `place`, entre otras.



```
Filas (registros): 489126
Columnas: 22
```

Figura 1: Cantidad de Registros del DataSet

3.2.1. Tipos de Variables en el Dataset

El conjunto de datos está compuesto por un total de **22 columnas** y **489,056 registros**, abarcando un rango temporal desde el **1 de enero de 1990** hasta el **31 de diciembre de 2024**. La Figura 2 resume la distribución de las variables por tipo. En total, se identifican **12 columnas numéricas**, como `latitude`, `longitude`, `depth`, `mag`, entre otras; estas presentan valores continuos, por ejemplo, `latitude` tiene un rango entre -84.42 y 87.39, y una media de 7.08. Además, existen **9 columnas de texto**, tales como `magType`, `net` y `place`, y finalmente, una única columna con formato `datetime`: `time`, que almacena la fecha y hora exacta de cada evento sísmico. Esta estructura permite aplicar distintos tipos de análisis estadístico y temporal con precisión y sin necesidad inicial de tratamiento de valores faltantes.

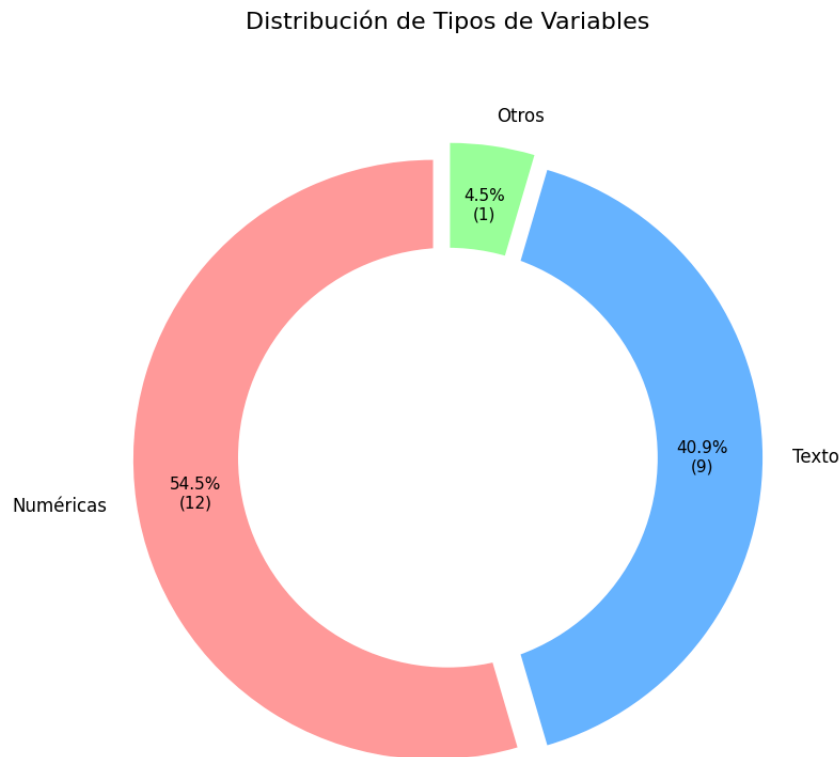


Figura 2: Distribución de columnas según tipo de dato en el dataset.

3.3. Estadísticas Descriptivas para Datos Numericos

Se calcularon estadísticas básicas para las variables numéricas clave:

3.3.1. Ubicación geográfica (latitude, longitude)

- **latitude** Como se muestra en la Figura 3, la columna 'latitude' contiene valores que van desde -84.4220 (mínimo) hasta 87.3860 (máximo), con una media de 7.0806 y una mediana de 5.7680. La similitud entre la media y la mediana indica una distribución equilibrada, con el 50 % central de los datos ubicado entre -16.8664 (primer cuartil) y 34.5340 (tercer cuartil).

El conjunto de datos contiene 226042 valores únicos en esta columna. No se encontraron valores nulos en esta columna.

La mayor concentración de valores se encuentra en el rango [30.1167-44.4340] con 88,248.0 registros (18.0 % del total).

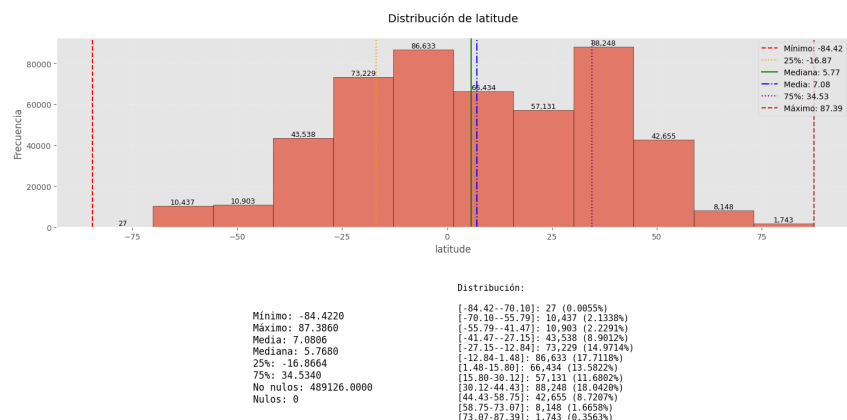


Figura 3: Columnna: Latitude

- **longitude** En la Figura 4 ,la columna 'longitude' contiene valores que van desde -179.9997 (mínimo) hasta 180.0000 (máximo), con una media de 29.7551 y una mediana de 75.9048. La diferencia entre la media y la mediana sugiere una distribución asimétrica, con el 50 % central de los datos ubicado entre -74.3107 (primer cuartil) y 140.7040 (tercer cuartil).

El conjunto de datos contiene 260256 valores únicos en esta columna. No se encontraron valores nulos en esta columna.

La mayor concentración de valores se encuentra en el rango [120.0000-150.0000] con 129,009.0 registros (26.4 % del total).

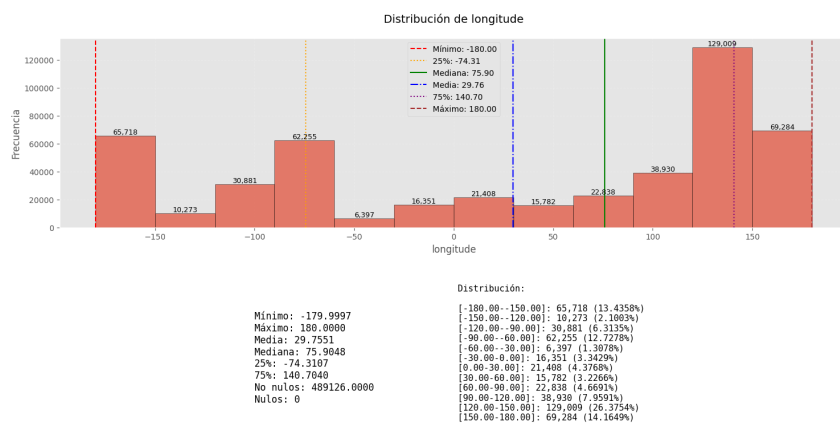


Figura 4: Columnna: Longitude

3.3.2. Profundidad (depth)

En la Figura 5 ,la columna 'depth' contiene valores que van desde -3.7400 (mínimo) hasta 735.8000 (máximo), con una media de 73.6598 y una mediana de 33.0000. La diferencia entre la media y la mediana sugiere una distribución asimétrica, con el 50 % central de los datos ubicado entre 10.0000 (primer cuartil) y 70.6000 (tercer cuartil).

El conjunto de datos contiene 53609 valores únicos en esta columna. No se encontraron valores nulos en esta columna.

La mayor concentración de valores se encuentra en el rango [-3.7400-57.8883] con 348,281.0 registros (71.2 % del total). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

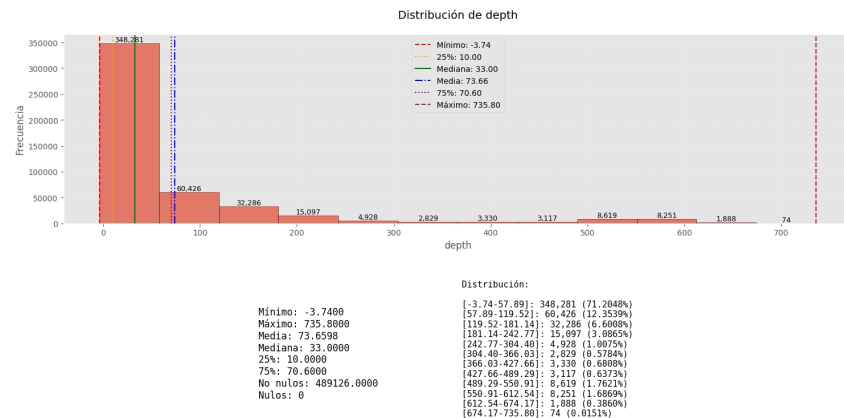


Figura 5: Columnna: Depth

3.3.3. Magnitud (mag)

En la Figura 6 ,la columna 'mag' contiene valores que van desde 3.3800 (mínimo) hasta 9.1000 (máximo), con una media de 4.4409 y una mediana de 4.4000. La similitud entre la media y la mediana indica una distribución equilibrada, con el 50 % central de los datos ubicado entre 4.1000 (primer cuartil) y 4.7000 (tercer cuartil).

El conjunto de datos contiene 250 valores únicos en esta columna. No se encontraron valores nulos en esta columna.

La mayor concentración de valores se encuentra en el rango [4.3333-4.8100] con 193,143.0 registros (39.5 % del total).

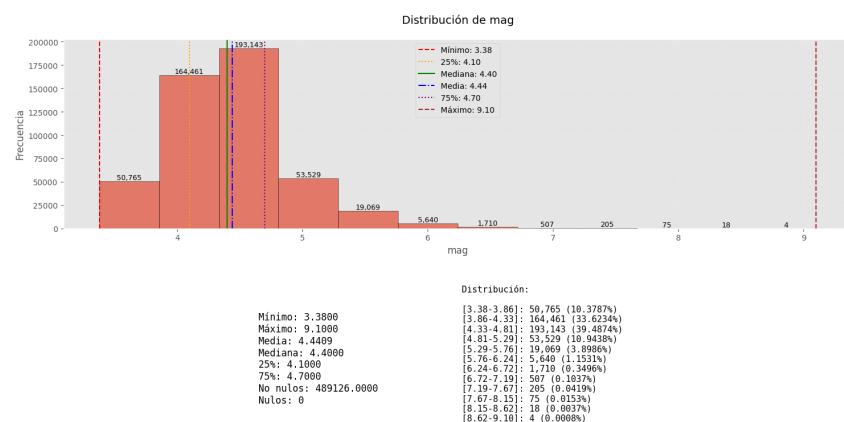


Figura 6: Columnna: Mag

3.3.4. Número de estaciones (nst)

En la Figura 11, la columna 'nst' contiene valores que van desde 0.0000 (mínimo) hasta 934.0000 (máximo), con una media de 53.4436 y una mediana de 28.0000. La diferencia entre la media y la mediana sugiere una distribución asimétrica, con el 50 % central de los datos ubicado entre 16.0000 (primer cuartil) y 59.0000 (tercer cuartil). El conjunto de datos contiene 728 valores únicos en esta columna. Existen 215,375 valores nulos (44.0 % del total), lo cual es una proporción significativa que podría requerir tratamiento especial. La mayor concentración de valores se encuentra en el rango [0.0000-77.8333] con 223,982.0 registros (81.8 % del total). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

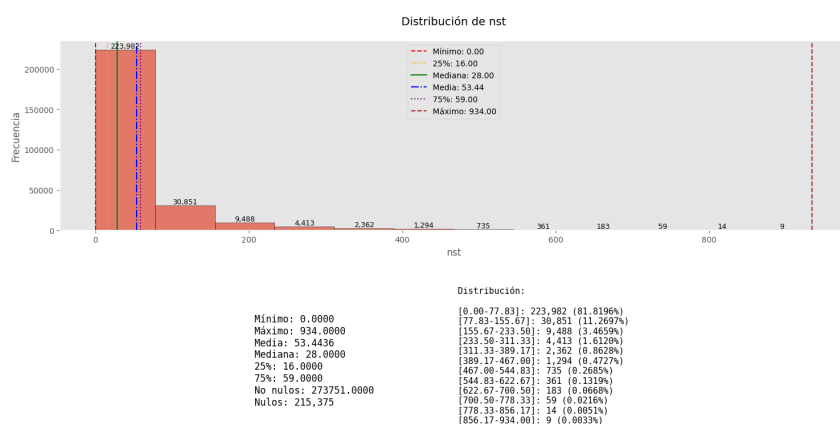


Figura 7: Columna: Nst

3.3.5. Errores asociados

Incluye errores estimados en las mediciones:

- **horizontalError**: error en la ubicación horizontal.

En la Figura 8, la columna 'horizontalError' contiene valores que van desde 0.0000 (mínimo) hasta 99.0000 (máximo), con una media de 8.1787 y una mediana de 7.9000. La similitud entre la media y la mediana indica una distribución equilibrada, con el 50 % central de los datos ubicado entre 5.7000 (primer cuartil) y 10.5000 (tercer cuartil). El conjunto de datos contiene 2405 valores únicos en esta columna. Existen 299,007 valores nulos (61.1%) lo cual es una proporción significativa que podría requerir tratamiento especial.

La mayor concentración de valores se encuentra en el rango [0.0000-8.2500] con 102,794.0 registros (54.1 % del total). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

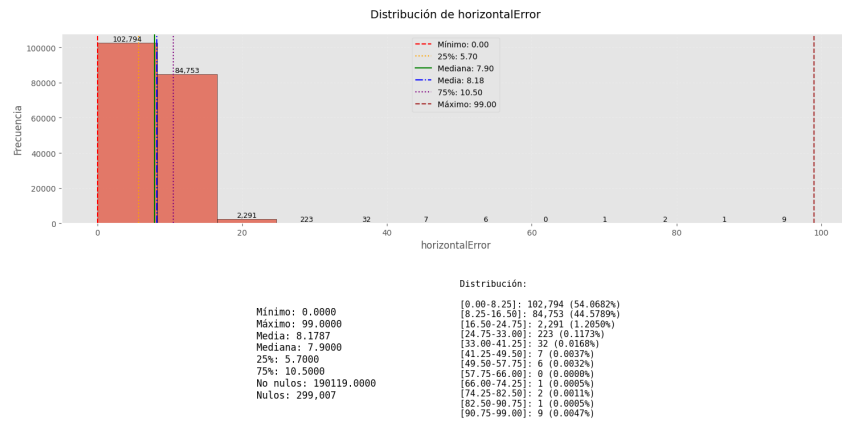


Figura 8: Columnna: horizontalError

- **depthError**: error en la estimación de profundidad.

En la Figura 9, La columna 'depthError' contiene valores que van desde -1.0000 (mínimo) hasta 367558.1000 (máximo), con una media de 11.2795 y una mediana de 5.8735. La similitud entre la media y la mediana indica una distribución equilibrada, con el 50 % central de los datos ubicado entre 1.9050 (primer cuartil) y 9.5000 (tercer cuartil).

El conjunto de datos contiene 11665 valores únicos en esta columna. Existen 171,692 valores nulos (35.1 % del total), lo cual es una proporción significativa que podría requerir tratamiento especial.

La mayor concentración de valores se encuentra en el rango [-1.0000-30628.9250] con 317,428.0 registros (100.0%). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

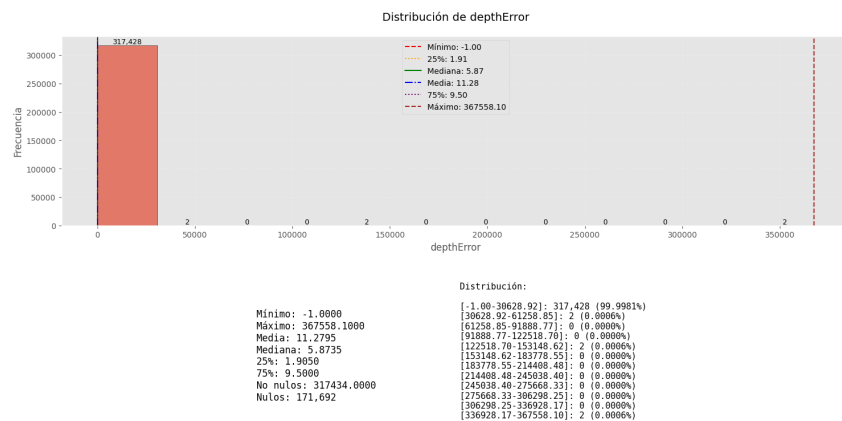


Figura 9: Columnna: depthError

- **magError**: error en la estimación de magnitud. En la Figura 10, la columna 'magError' contiene valores que van desde 0.0000 (mínimo) hasta 5.0000 (máximo), con una media de 0.1289 y una mediana de 0.1180. La diferencia entre la media y la mediana sugiere una distribución asimétrica, con el 50 % central de los datos ubicado entre 0.0780 (primer cuartil) y 0.1620 (tercer cuartil).

El conjunto de datos contiene 756 valores únicos en esta columna. Existen 281,753 valores nulos (57.6 % del total), lo cual es una proporción significativa que podría requerir tratamiento especial. La mayor concentración de valores se encuentra en el rango [0.0000-0.4167] con 206,145.0 registros (99.4 % del total). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

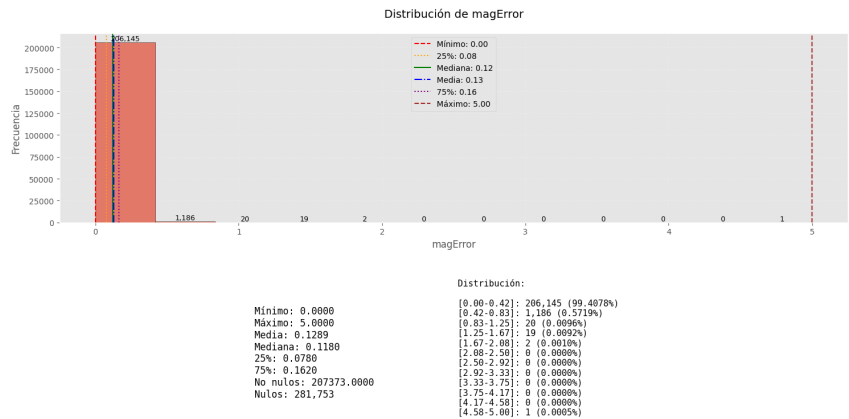


Figura 10: Columna: magError

3.3.6. Número de estaciones (nst)

Cantidad de estaciones sísmicas que registraron el evento. Este valor influye directamente en la precisión de la localización del sismo.

En la Figura 11 ,la columna 'nst' contiene valores que van desde 0.0000 (mínimo) hasta 934.0000 (máximo), con una media de 53.4445 y una mediana de 28.0000. La diferencia entre la media y la mediana sugiere una distribución asimétrica, con el 50 % central de los datos ubicado entre 16.0000 (primer cuartil) y 59.0000 (tercer cuartil). El conjunto de datos contiene 728 valores únicos en esta columna. Existen 215,334 valores nulos (44.0 % del total), lo cual es una proporción significativa que podría requerir tratamiento especial. La mayor concentración de valores se encuentra en el rango [0.0000-77.8333] con 223,959.0 registros (81.8 % del total). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

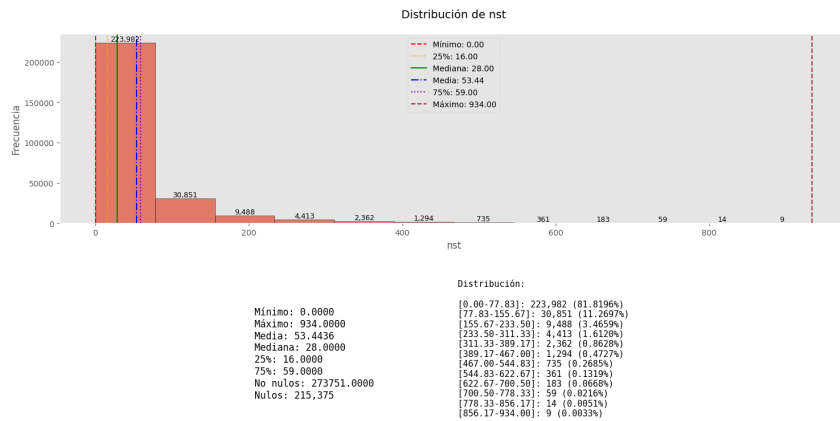


Figura 11: Columnna:nst

3.3.7. Gap (gap)

Ángulo en grados entre las estaciones más alejadas que detectaron el sismo. Un menor valor indica una mejor cobertura.

En la Figura 12, la columna 'gap' contiene valores que van desde 0.0000 (mínimo) hasta 358.3000 (máximo), con una media de 117.1729 y una mediana de 111.0000. La diferencia entre la media y la mediana sugiere una distribución asimétrica, con el 50 % central de los datos ubicado entre 74.0000 (primer cuartil) y 148.0000 (tercer cuartil). El conjunto de datos contiene 3906 valores únicos en esta columna. Existen 68,952 valores nulos (14.1 % del total), lo cual es una proporción significativa que podría requerir tratamiento especial.

La mayor concentración de valores se encuentra en el rango [89.5750-119.4333] con 87,801.0 registros (20.9 % del total).

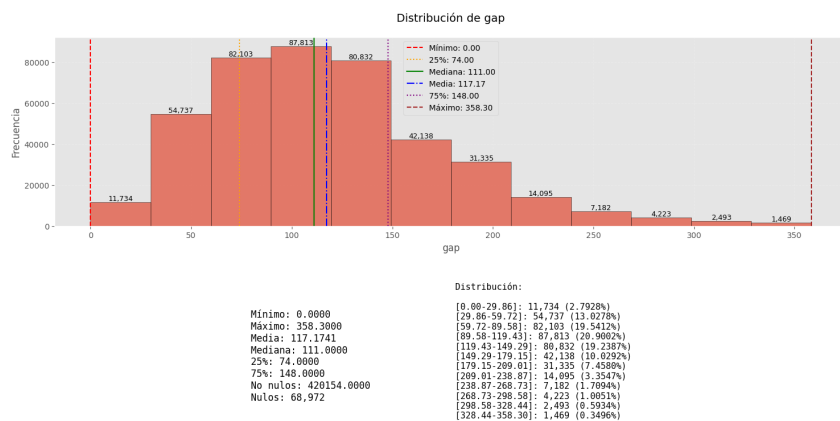


Figura 12: Columnna:gap

3.3.8. Distancia mínima (dmin)

Distancia mínima en grados desde el epicentro hasta una estación sísmica.

En la Figura 13, la columna 'dmin' contiene valores que van desde 0.0000 (mínimo) hasta 64.4980 (máximo), con una media de 3.6485 y una mediana de 2.2300. La diferencia entre

la media y la mediana sugiere una distribución asimétrica, con el 50 % central de los datos ubicado entre 1.0230 (primer cuartil) y 4.2140 (tercer cuartil). El conjunto de datos contiene 22931 valores únicos en esta columna. Existen 280,581 valores nulos (57.4 % del total), lo cual es una proporción significativa que podría requerir tratamiento especial.

La mayor concentración de valores se encuentra en el rango [0.0000-5.3748] con 172,015.0 registros (82.5 % del total). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

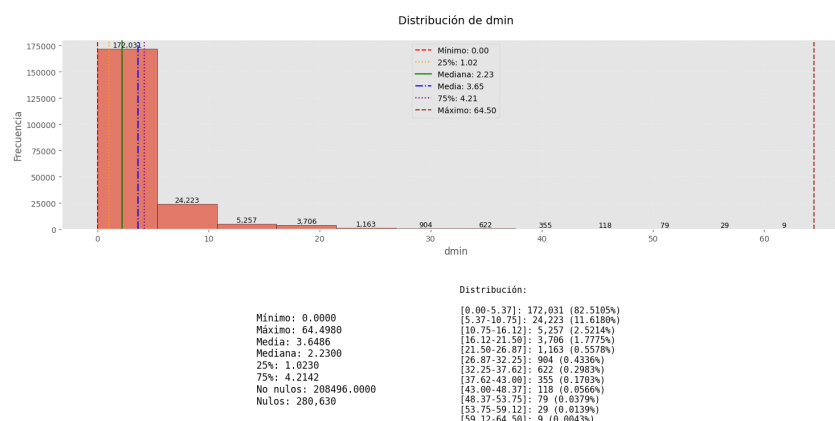


Figura 13: Columnna:dmin

3.3.9. Error cuadrático medio (rms)

Error cuadrático medio del ajuste entre las estaciones y el modelo de propagación de ondas sísmicas.

En la Figura 14 ,La columna 'rms' contiene valores que van desde -1.0000 (mínimo) hasta 69.3200 (máximo), con una media de 0.8555 y una mediana de 0.8500. La similitud entre la media y la mediana indica una distribución equilibrada, con el 50 % central de los datos ubicado entre 0.6600 (primer cuartil) y 1.0500 (tercer cuartil).

El conjunto de datos contiene 1000 valores únicos en esta columna. Existen 42,815 valores nulos (8.8 % del total), lo cual es una proporción significativa que podría requerir tratamiento especial.

La mayor concentración de valores se encuentra en el rango [-1.0000-4.8600] con 446,230.0 registros (100.0 % del total). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

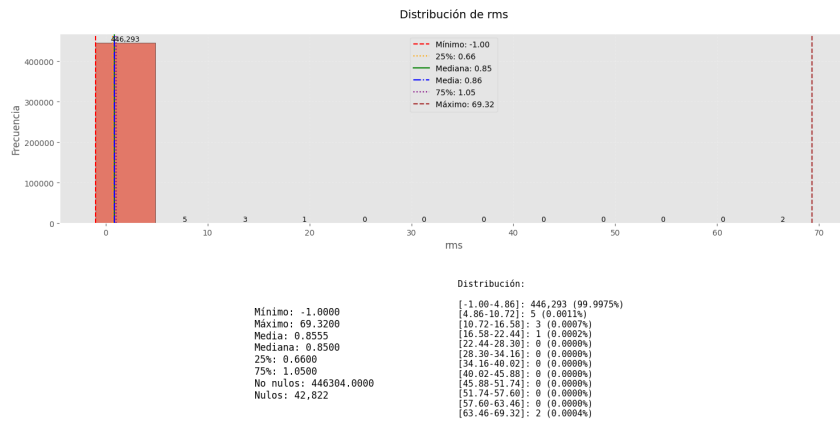


Figura 14: Columnna:rms

3.3.10. Número de estaciones para magnitud (magNst)

Cantidad de estaciones utilizadas específicamente para calcular la magnitud del evento.

En la Figura 15 ,la columna 'magNst' contiene valores que van desde 0.0000 (mínimo) hasta 954.0000 (máximo), con una media de 27.3925 y una mediana de 13.0000. La diferencia entre la media y la mediana sugiere una distribución asimétrica, con el 50 % central de los datos ubicado entre 5.0000 (primer cuartil) y 28.0000 (tercer cuartil).

El conjunto de datos contiene 650 valores únicos en esta columna. Existen 89,933 valores nulos (18.4 % del total), lo cual es una proporción significativa que podría requerir tratamiento especial.

La mayor concentración de valores se encuentra en el rango [0.0000-79.5000] con 369,891.0 registros (92.7 % del total). Se observa una amplia dispersión en los datos, con valores que se extienden mucho más allá del rango intercuartílico.

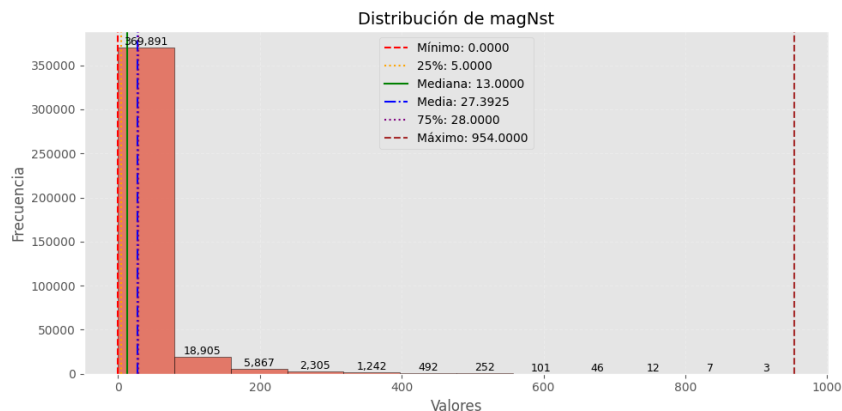


Figura 15: Columnna:magns

3.4. Estadísticas Descriptivas para Datos Textuales

3.4.1. Ubicación geográfica legible del epicentro del sismo(place)

La columna `place`, que describe la ubicación textual del epicentro del sismo, es de tipo `object` y contiene un total de **192,050 valores únicos** sin valores nulos. En cuanto a las características textuales, la longitud promedio de las entradas es de **29.5 caracteres**, con un mínimo de **4** y un máximo de **66 caracteres**.

Los valores más frecuentes en esta columna incluyen:

- `South Sandwich Islands region`: 8,235 ocurrencias (1.7 %)
- `south of the Fiji Islands`: 7,943 ocurrencias (1.6 %)
- `Kermadec Islands region`: 6,230 ocurrencias (1.3 %)
- `Fiji region`: 5,412 ocurrencias (1.1 %)
- `Bonin Islands, Japan region`: 4,065 ocurrencias (0.8 %)

Esta distribución puede observarse en la Figura 16, la cual muestra las ubicaciones textuales más comunes dentro del conjunto de datos.

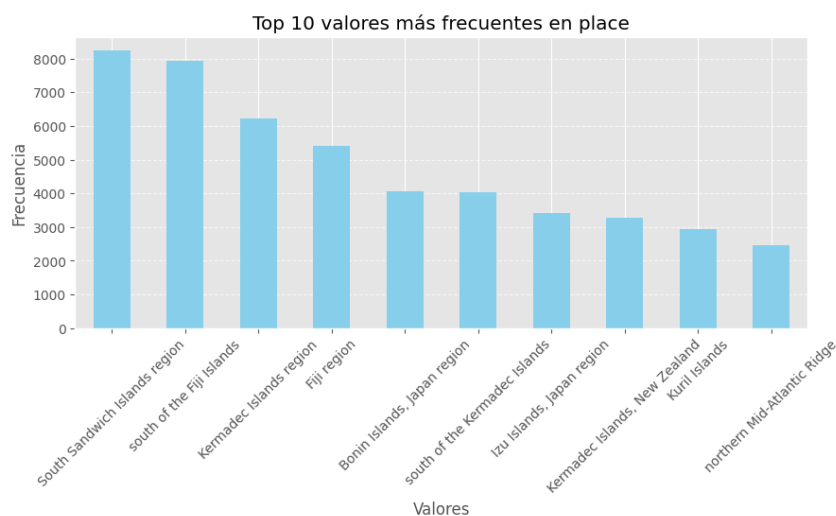


Figura 16: Distribución de las ubicaciones más frecuentes en la columna `place`.

3.4.2. Tipo de magnitud del evento sísmico (magType)

La columna `magType`, correspondiente al tipo de magnitud utilizada para estimar la energía del evento sísmico, es de tipo `object` y contiene **25 valores únicos**, sin registros nulos. Desde el punto de vista textual, las entradas presentan una longitud promedio de **2.1 caracteres**, con un mínimo de **1 carácter** y un máximo de **10 caracteres**. Los valores más frecuentes en esta columna reflejan la predominancia de ciertos métodos de medición: `mb` (magnitud de ondas de cuerpo) aparece en **373,693 registros** (76.4 %), seguido por `ml` (magnitud local) con **28,521 ocurrencias** (5.8 %), `md` (magnitud de duración) con **25,176** (5.1 %), `mwc` con **19,388** (4.0 %) y `mww` con **14,098** (2.9 %). Esta distribución es representada visualmente en la Figura 17, donde se destacan los tipos de magnitud más comúnmente utilizados por las estaciones sísmicas.

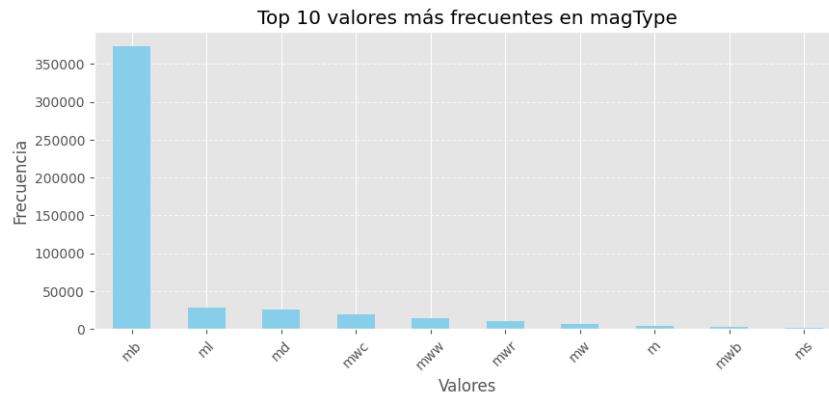


Figura 17: Distribución de los tipos de magnitud más frecuentes en la columna `magType`.

4. Limpieza de Datos

4.1. Manejo de Valores Faltantes

En el análisis preliminar del dataset se identificaron varias columnas con un alto porcentaje de valores faltantes (nulos). Como se observa en la Figura 18, las columnas `horizontalError`, `magError`, `dmin`, `nst`, `depthError`, `magNst`, `gap` y `rms` presentan un número considerable de registros con datos ausentes, que van desde aproximadamente 40,000 hasta casi 300,000 valores faltantes. Los registros con valores nulos o incompletos se eliminaron para garantizar la calidad y consistencia del análisis. La presencia de datos faltantes en columnas clave podría sesgar los resultados o dificultar la aplicación de técnicas estadísticas y modelos de aprendizaje automático. Por ello, se optó por depurar el conjunto de datos, manteniendo únicamente aquellos registros que contuvieran información completa y confiable, lo cual mejora la robustez de los análisis posteriores.

📌 Columnas con valores nulos:

<code>horizontalError</code>	299007
<code>magError</code>	281753
<code>dmin</code>	280630
<code>nst</code>	215375
<code>depthError</code>	171692
<code>magNst</code>	89951
<code>gap</code>	68972
<code>rms</code>	42822
<code>dtype:</code>	<code>int64</code>

Figura 18: Número de valores faltantes por columna en el dataset.

4.2. Detección y eliminación de registros duplicados

En el conjunto de datos original se encontraron 489 126 registros, de los cuales 114 836 corresponden a registros duplicados, mientras que 374 290 son únicos. La presencia de duplicados puede afectar negativamente el análisis, ya que introduce sesgos y distorsiona las estadísticas descriptivas.

Por ello, se procedió a eliminar estos registros duplicados mediante técnicas de limpieza de datos, dejando como resultado un dataset con únicamente registros únicos. Esta operación mejora la calidad y confiabilidad del análisis posterior.

La Figura 19 muestra visualmente la distribución entre los registros originales, duplicados y registros únicos en el conjunto de datos.

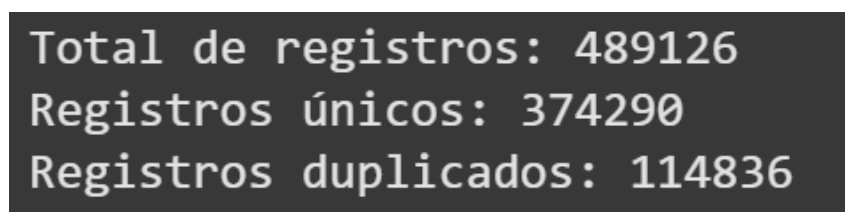


Figura 19: Cantidad de registros totales, duplicados y únicos en el dataset.

4.3. Corrección de Inconsistencias

Se corrigieron formatos inconsistentes en la columna `time` para asegurar que todas las fechas y horas estuvieran en el formato ISO 8601.

5. Transformación de Datos

5.1. Conversión de Tipos de Datos

La columna `time` se convirtió de string a un objeto de fecha y hora para facilitar el análisis temporal.

5.2. Creación de Nuevas Variables

Se creó una nueva variable llamada `year` extrayendo el año de la columna `time`, que fue previamente convertida a un objeto de fecha y hora. Esta nueva variable facilita el análisis temporal anual de la frecuencia de sismos. A continuación, se muestran los primeros cinco registros con la columna `time` y la nueva columna `year`:

		time	year
0	1990-01-31	22:58:18.630000+00:00	1990
1	1990-01-31	21:00:46.120000+00:00	1990
2	1990-01-31	20:22:13.690000+00:00	1990
3	1990-01-31	19:58:31.010000+00:00	1990
4	1990-01-31	16:23:14.010000+00:00	1990

Figura 20: Creacion de la columna year

5.3. Normalización de la columna place

La columna `place` se normalizó para asegurar consistencia en los nombres de los lugares, eliminando espacios adicionales al inicio y final, y convirtiendo todos los textos a minúsculas. Esta limpieza ayuda a corregir errores tipográficos comunes y facilita el análisis posterior.

En la Figura 21 se muestran los primeros cinco registros de la columna `place` después de la normalización.

	place
0	9 km ene of isole tremiti, italy
1	2 km sw of ignacio zaragoza, mexico
2	14 km ene of isole tremiti, italy
3	7 km s of eresós, greece
4	36 km w of filiatrá, greece

Figura 21: Primeros cinco registros de la columna `place` normalizados.

6. Validación de Datos

6.1. Verificación de Consistencia

6.1.1. Verificación y corrección de la columna `mag`

Se verificó que todos los valores en la columna `mag` estuvieran dentro del rango esperado de 3.5 a 10. Se identificaron algunos valores fuera de este rango y se corrigieron reemplazándolos por los límites correspondientes para mantener la calidad y coherencia de los datos. La Figura 22 muestra la verificación.

```
Valores fuera del rango esperado: 0
count    489056.000000
mean      4.440876
std       0.494054
min       3.500000
25%       4.100000
50%       4.400000
75%       4.700000
max       9.100000
```

Figura 22: Verificación y correlación de la columna mag

6.2. Detección de Outliers

Se utilizaron diagramas de caja para identificar outliers en las columnas `depth` y `mag`. Los valores atípicos se mantuvieron en el dataset, ya que representan eventos sísmicos reales y significativos.

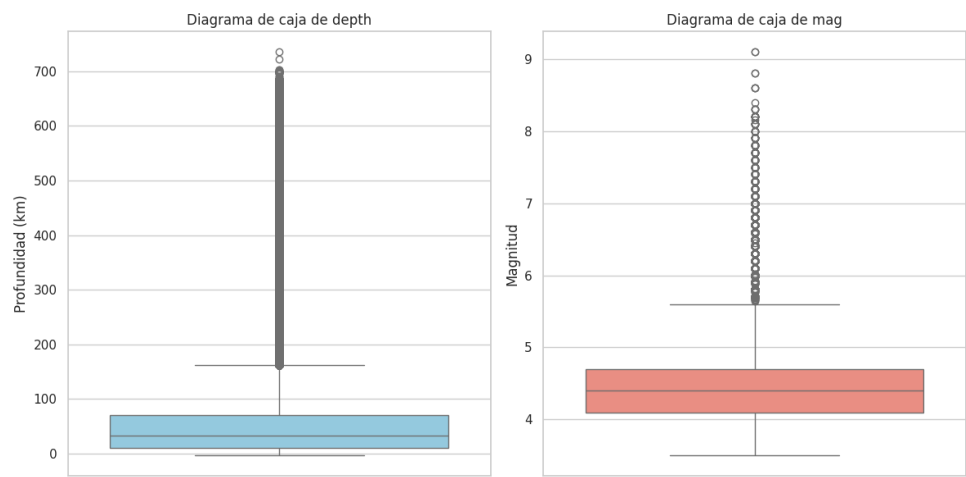


Figura 23: Diagramas de caja para las columnas `depth` y `mag`.

7. Conclusión

El proceso de data wrangling aplicado al dataset de sismos registrados por el USGS desde 1990 hasta 2024 ha permitido transformar los datos crudos en un conjunto de datos limpio y estructurado, listo para análisis posteriores. Se identificaron y corrigieron valores faltantes, duplicados e inconsistencias, y se realizaron transformaciones necesarias para facilitar el análisis.

7.1. Limitaciones

Una limitación importante es la presencia de algunos valores atípicos que, aunque se mantuvieron por su relevancia, pueden afectar ciertos tipos de análisis estadísticos.

8. Apéndices

8.1. Código Utilizado

El código utilizado para el análisis se encuentra disponible en el siguiente enlace:
Repositorio en GitHub.