"Año de la recuperación y consolidación de la economía peruana".

Escuela Profesional de Ciencia de la Computación

Tópicos Ciencia de Datos

Informe Pipeline de Cs. de Datos

Estudiante:
Andrea del Rosario Lopez Condori

Docente :
ANA MARIA CUADROS VALDIVIA

Índice

1. Introducción

El presente informe tiene como objetivo aplicar un pipeline completo sobre un conjunto de datos sísmicos obtenidos del Servicio Geológico de los Estados Unidos (USGS), que comprende eventos registrados entre los años 1990 y 2024. A lo largo del pipeline se identifican problemas en los datos, se corrigen mediante técnicas de *data wrangling*, se exploran tendencias significativas y se plantean preguntas de investigación que dan origen a hipótesis verificables.

Este proceso no solo garantiza la calidad y consistencia del análisis, sino que también permite generar nuevos conocimientos sobre la actividad sísmica global mediante el uso riguroso y metódico de los datos.

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.

Columna	Descripción Detallada
latitude	La latitud geográfica del epicentro del sismo, medida en grados decimales.
longitude	La longitud geográfica del epicentro del sismo, medida en grados decimales.
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. Preguntas con respecto al Dataset

¿Qué problema se identificó en los datos?

Durante la revisión inicial del conjunto de datos sísmicos proporcionado por el USGS, se detectaron múltiples problemas que comprometen la calidad del análisis posterior. Estos problemas no solo afectan la integridad del dataset, sino también la validez de cualquier

conclusión que pudiera extraerse de forma directa sin aplicar un proceso riguroso de limpieza. Los principales problemas identificados fueron:

1. **Registros duplicados:** Se encontraron más de 114,000 registros idénticos, lo cual representa aproximadamente el 23 % del total. Si no se eliminan, estos inflan la frecuencia de eventos y distorsionan los análisis estadísticos.

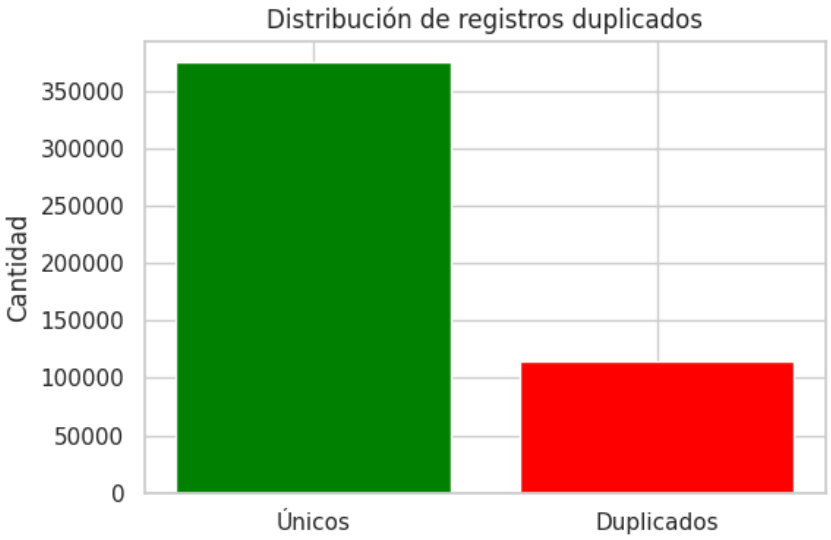


Figura 1: Distribución de registros duplicados en el dataset.

2. **Valores faltantes:** Varias columnas presentan un porcentaje elevado de valores nulos, especialmente:

   - `magError` (error de magnitud): 57.6 % de valores faltantes.
   - `horizontalError` (error en ubicación): 61.1 %.
   - `dmin` (distancia mínima): 57.4 %.
   - `nst` (número de estaciones): 44.0 %.

   Estos valores faltantes pueden impedir análisis robustos si no se tratan adecuadamente.

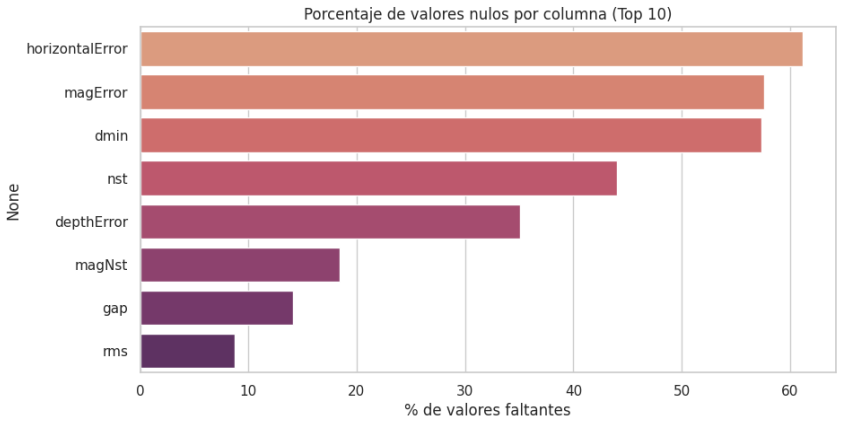


Figura 2: Distribución de registros faltantes por columna.

3. **Inconsistencias de formato:** La columna `time`, que contiene la fecha y hora del evento, estaba almacenada como texto y no como objeto de fecha, lo que dificultaba su uso en análisis temporales.

4. **Problemas de estandarización textual:** La columna `place`, que indica la ubicación legible del sismo, presentaba formatos diversos, inconsistencias de mayúsculas/minúsculas, y espacios adicionales. Esto hacía difícil agrupar correctamente por regiones o países.

5. **Presencia de valores extremos:** Algunas variables como `depthError` y `rms` contenían valores que superaban ampliamente los rangos esperados. Aunque no todos los valores extremos son errores, su presencia exige validación adicional.

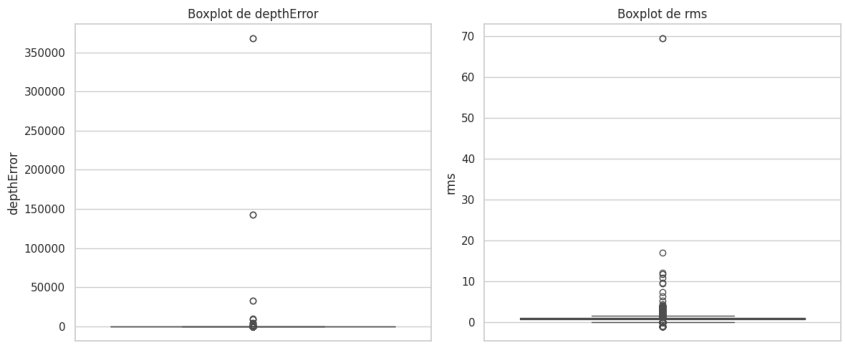


Figura 3: Distribución de registros faltantes pro columna.

Estos problemas demuestran la necesidad de aplicar un proceso riguroso de *data wrangling*, que permita limpiar, transformar y preparar los datos antes de realizar cualquier análisis exploratorio o formular hipótesis confiables.

3.1. ¿Qué descubren al analizar los datos?

Se observan patrones claros en los datos:

- **Magnitud promedio por año:** Se observó una variación en la magnitud promedio de los sismos a lo largo del tiempo. En ciertos años se registraron promedios significativamente más altos, lo que puede estar relacionado con eventos sísmicos de gran magnitud que ocurrieron en esos periodos.

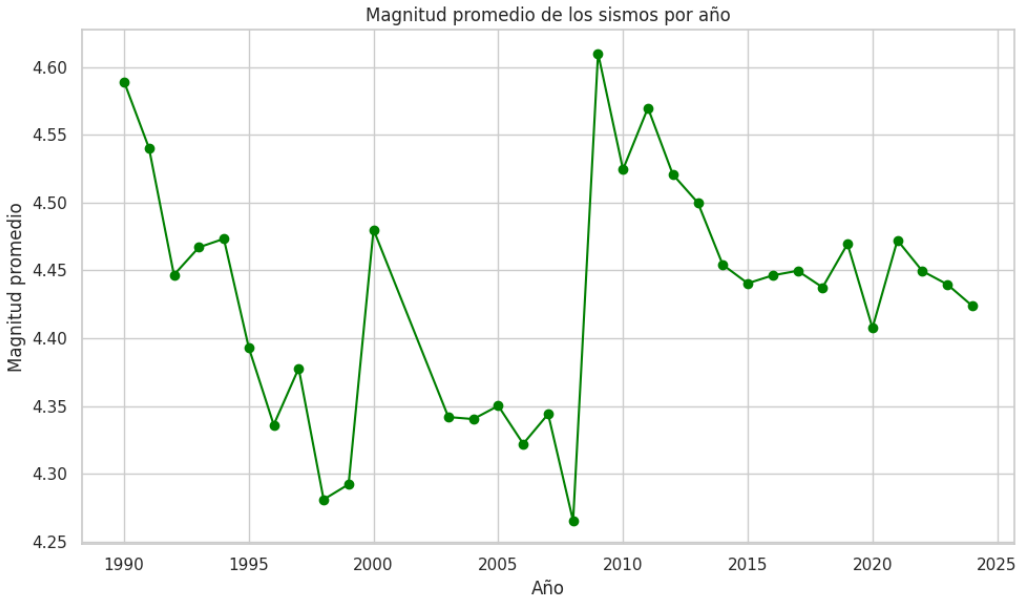


Figura 4: Magnitud promedio de los sismos por año.

- **Relación entre magnitud y profundidad:** El gráfico de dispersión muestra que los sismos con magnitudes altas pueden ocurrir tanto en profundidades bajas como altas. Sin embargo, existe una mayor concentración de eventos con magnitudes moderadas (3 a 5) a profundidades menores a 100 km, lo cual concuerda con la distribución típica de sismos tectónicos.

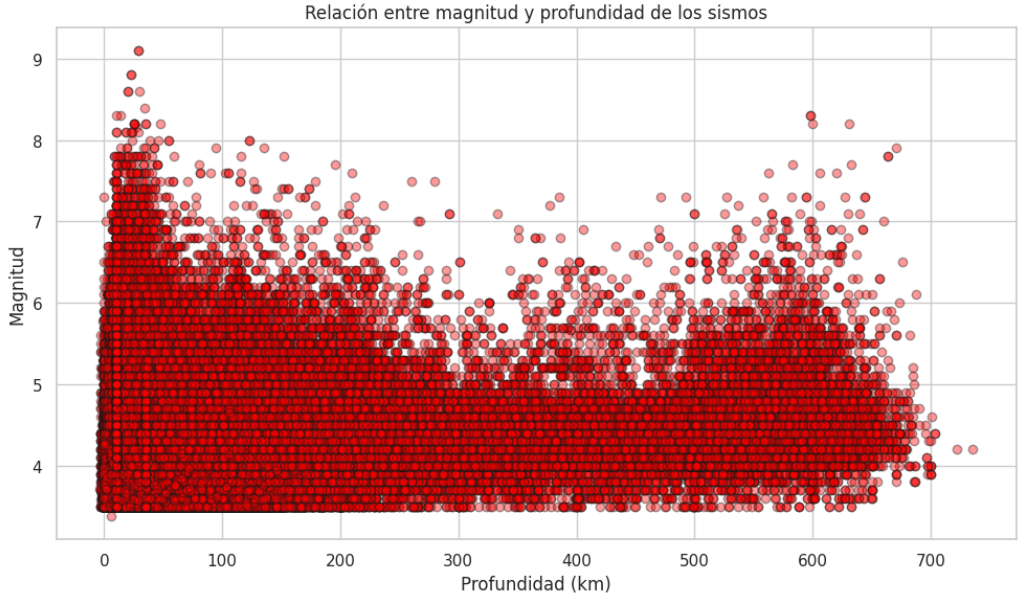


Figura 5: Relación entre magnitud y profundidad de los sismos.

- **Mapas de calor de ubicación por año:** Se elaboraron mapas de calor que muestran la densidad espacial de los sismos para distintos años seleccionados. Estos mapas evidencian cómo cambia la actividad sísmica en diferentes regiones a lo largo del tiempo.

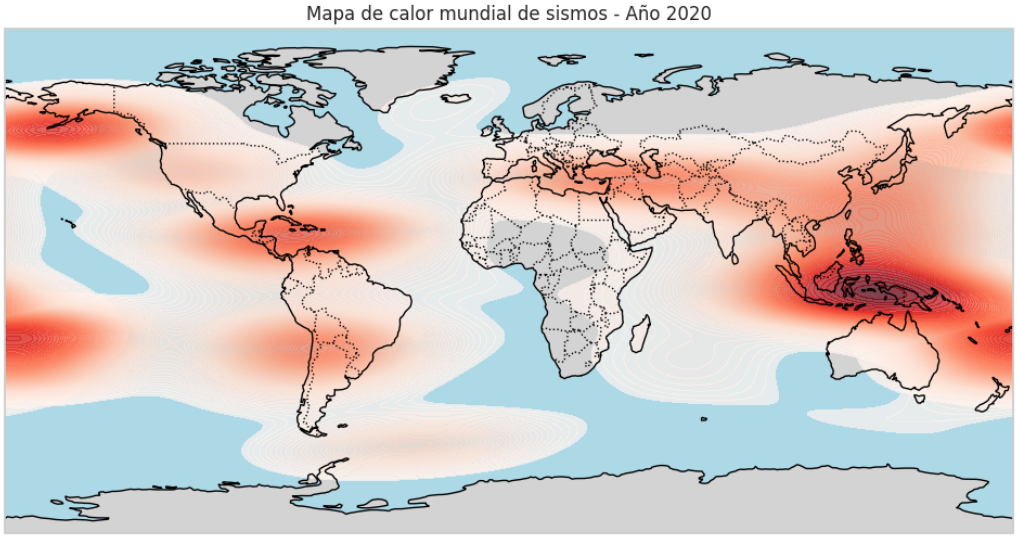


Figura 6: Mapa de calor de eventos sísmicos para el año 2020.

3.2. ¿Qué reflejan los patrones de tendencia?

Se identifican relaciones y posibles correlaciones:

- **Patrones de tendencia en los datos sísmicos:**

  Al analizar la magnitud promedio anual y la cantidad de eventos sísmicos a lo largo de los años, se observan ciertos patrones relevantes:

  - La magnitud promedio de los sismos muestra una tendencia relativamente estable, con ligeras fluctuaciones que podrían reflejar variaciones naturales en la actividad sísmica anual.
  - El número de eventos sísmicos por año presenta una tendencia creciente en los últimos años, lo cual puede deberse a mejoras en la tecnología de detección y reporte, así como a una mayor actividad sísmica en ciertas regiones.
  - Estos patrones permiten identificar periodos de mayor o menor actividad sísmica y ayudan a entender la dinámica tectónica subyacente.

  Las gráficas correspondientes se encuentran a continuación, donde se visualizan estas tendencias temporales que aportan una visión clara sobre el comportamiento de los eventos sísmicos en el tiempo.

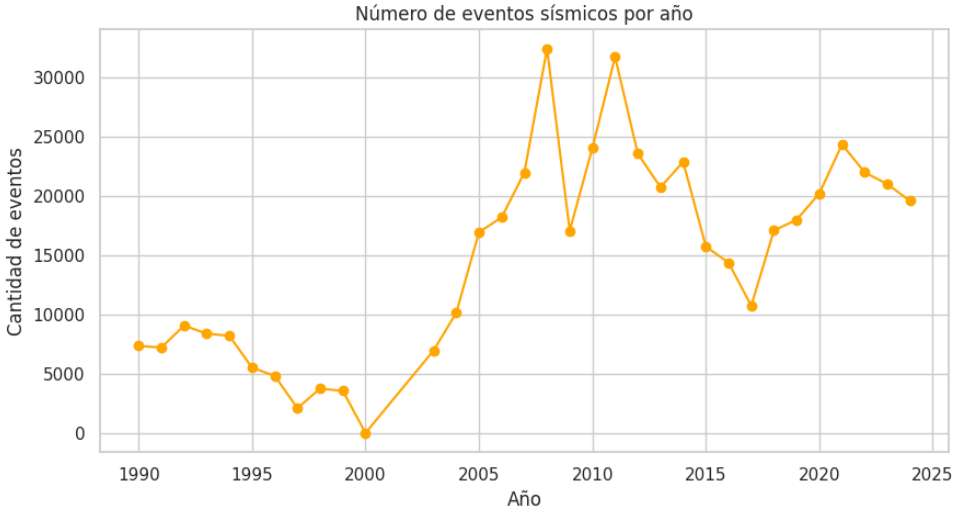


Figura 7: Evolución anual del número de eventos sísmicos

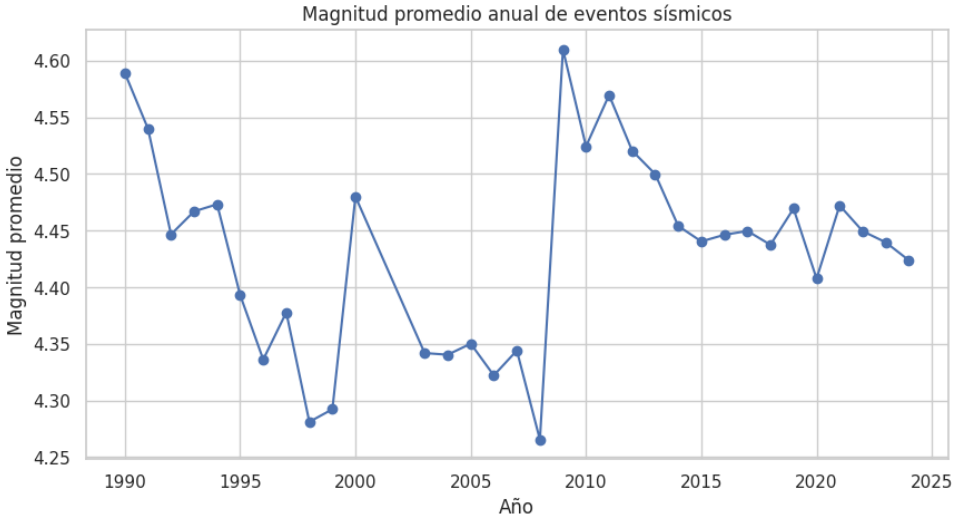


Figura 8: Magnitud promedio anual de los eventos sísmicos.

- Alta concentración de eventos en regiones del Cinturón de Fuego del Pacífico.

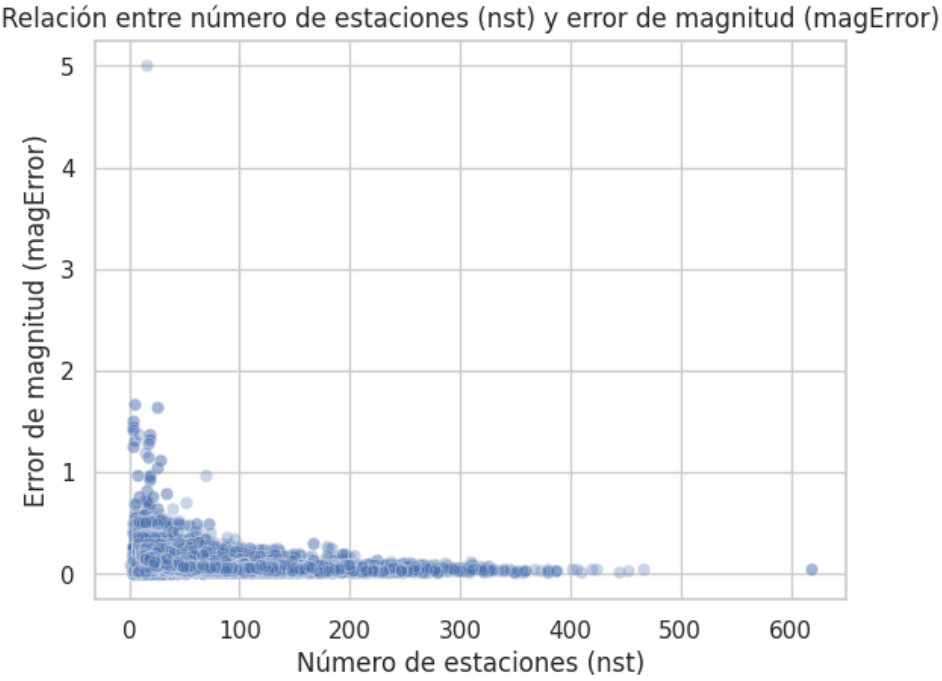
4. Preguntas de Investigación e Hipótesis

Con base en el análisis previo, se plantean las siguientes preguntas y sus respectivas hipótesis:

4.1. Pregunta 1: ¿Existe relación entre el número de estaciones y el error en la magnitud?

Hipótesis 1: A mayor número de estaciones que detectan un sismo (`nst`), menor será el error en la estimación de su magnitud (`magError`).

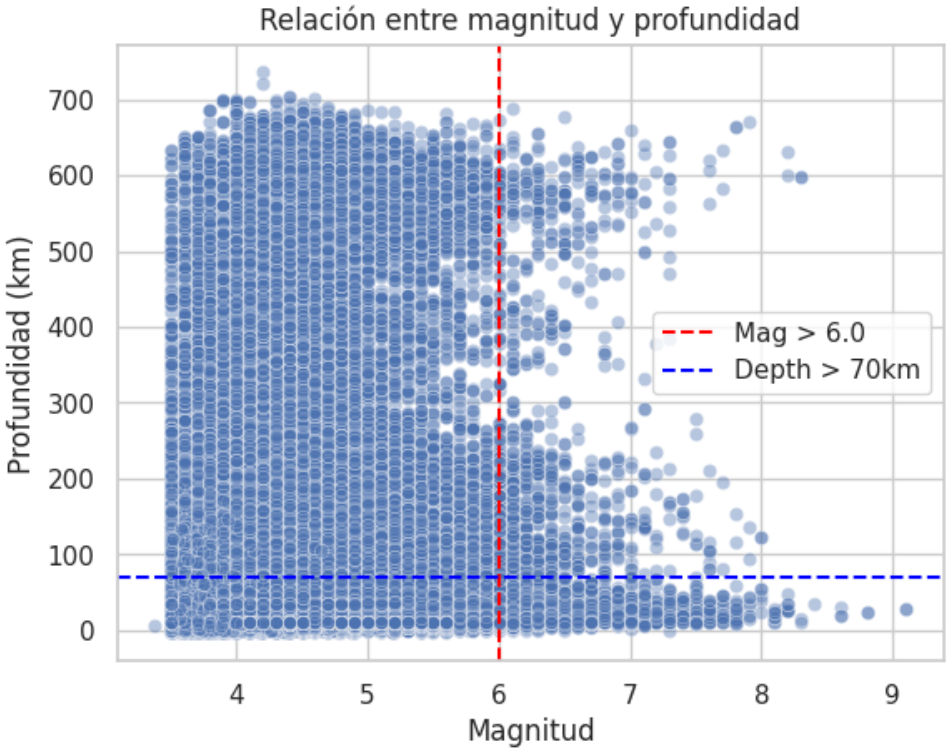
Esta hipótesis parte del supuesto de que contar con más estaciones sismológicas mejora la precisión del cálculo. Para evaluar esta relación, se elaboró un gráfico de dispersión donde se observa cómo varía el error de magnitud según el número de estaciones involucradas en la detección. Se espera una tendencia descendente si la hipótesis es válida.



4.2. Pregunta 2: ¿Los sismos de mayor magnitud ocurren a mayor profundidad?

Hipótesis 2: Los sismos con magnitudes superiores a 6.0 tienden a ocurrir a profundidades mayores a 70 km.

Esta hipótesis busca explorar la relación entre la energía liberada por un sismo y su ubicación en el subsuelo. Un gráfico de dispersión entre magnitud y profundidad permite observar si hay una agrupación de eventos fuertes (mag >6.0) en zonas profundas (depth >70 km), lo cual podría indicar un patrón geológico relevante.



4.3. Pregunta 3: ¿La mayoría de los eventos sísmicos ocurren en zonas del Cinturón de Fuego del Pacífico?

Hipótesis 3: Existe una alta concentración de eventos sísmicos en regiones que forman parte del Cinturón de Fuego del Pacífico, como Japón, Fiji, Chile y Alaska.

El Cinturón de Fuego del Pacífico es conocido por su intensa actividad sísmica y volcánica. Para validar esta hipótesis, se generó un mapa de calor global basado en las coordenadas de los eventos registrados. Se espera observar concentraciones significativas en las zonas del cinturón, confirmando su actividad predominante.