Anexos

Trabajo Fin de Máster

21 de septiembre de 2023

# Anexo Análisis Descriptivo

## Análisis de Variables

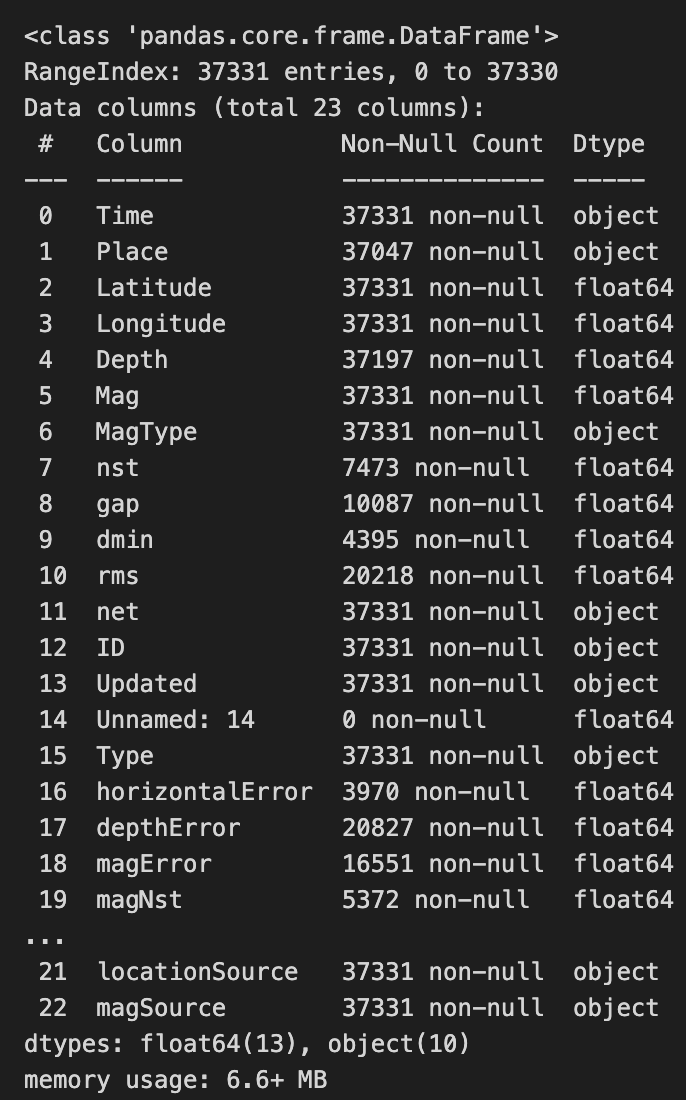
En este primer capítulo, se realiza una explicación extensa sobre la información que aportan las variables que componen el dataset escogido para el análisis (*Significant Earthquakes, 1965-2016* ). Este conjunto de datos, recopilado y mantenido a través del Centro Nacional de Información de Terremotos (NEIC, por sus siglas en inglés), aporta información sobre diferentes mediciones tomadas por las sondas que registran eventos sísmicos en todo el mundo, así como información sobre la forma de registro y errores en las mediciones.

Para ampliar la información que aportan estas variables, he acudido al Servicio Geológico de los Estados Unidos (USGS, por sus siglas en inglés) a través de la siguiente página web: <https://earthquake.usgs.gov/data/comcat>. En ella, se recopila la información de los parámetros de origen que componen el dataset, el cual está compuesto por 23 columnas.

Tras realizar un primer análisis exploratorio de las variables que conforman el dataset, se enumeran a continuación, las variables del dataset y una breve explicación de cada una de ellas:

* **Time**: Momento en el tiempo en el que el seísmo inicia su ruptura. Formato YYYY-MM-DDT00:00:00.000Z.
* **Place**: Lugar donde tiene lugar el seísmo. Columna tipo string que no sigue un patrón homogéneo. Parte de los registros indican primero los kilómetros hasta la población más cercana para luego especificar el condado/provincia y el país donde se produce el seísmo.
* **Latitude**: Latitud de la coordenada.
* **Longitude**: Longitud de la coordenada.
* **Depth**: Valor de la profundidad a la que un evento sísmico empieza su ruptura. Toma valores típicos entre 0-1000.
* **Mag**: Magnitud del evento sísmico medido en la escala Richter.
* **MagType**: Método/ Algoritmo utilizado para calcular la magnitud del evento. Algunos valores típicos son “Md”, “Ml”, “Ms”, “Mw”, “Me”, “Mi”, “Mb”, “MLg”.
* **nst**: Número de estaciones sísmicas utilizadas para determinar la ubicación de un único evento sísmico.
* **gap**: Mayor brecha azimutal entre estaciones azimutalmente adyacentes (en grados). Según indica la USGS, cuanto menor sea este número, más fiable será la posición horizontal calculada del terremoto.
* **dmin**: Distancia horizontal desde el epicentro del evento sísmico a la estación que realiza la lectura del evento sísmico. Es un valor medido en grados, por lo que para trasladarlo a km podemos hacer la siguiente aproximación 1º ≈ 111.2 km.
* **rms**: Tiempo de viaje raíz-cuadrado medio residual (RMS) medido en segundos. Proporciona una medida de ajuste de los tiempos de llegada de los eventos sísmicos con respecto a los tiempos de llegada previstos para esta ubicación.
* **net**: ID del contribuidor de la información, identificado a través de la red.
* **ID**: Identificador único del evento sísmico.
* **Updated**: Momento en el que el evento fue actualizado por última vez.
* **Unnamed: 14:** Columna con valores NaN. A primera vista, no tiene ningún valor.
* **Type:** Tipo de evento sísmico. Puede tomar los valores 'earthquake', 'volcanic eruption', 'nuclear explosion', 'explosion'.
* **horizontalError**: Error de la distancia (horizontal) medido en km. Los valores típicos oscilan entre 0-100 km.
* **depthError**: Error de la profundidad (vertical) medido en km. Los valores típicos oscilan entre 0-100 km.
* **magError**: Grado de incertidumbre del reporte de la magnitud del evento, expresado en valores de 0-100.
* **magNst**: Número total de estaciones sísmica utilizadas para calcular la magnitud del terremoto.
* **status**: Revisión de eventos sísmicos, variable que puede tomar el valor 'reviewed' si el evento sísmico ha sido verificado por una persona antes de ser publicado o 'automatic' en caso de que haya sido publicado por los procesos automáticos de los sistemas de detección sísmica.
* **locationSource**: Red que originó el reporte de la localización del evento sísmico. Es un string que puede tomar 54 valores diferentes. Estos tienen relación, por lo que parece a primera vista, con la ubicación donde se produjo el terremoto.
* **magSource**: Red que originó el reporte de la magnitud del evento sísmico. En el 71% de los registros se trata del mismo centro el que proporciona el valor de la ubicación y el de la magnitud del evento sísmico.

En total, hay 10 columnas de tipo numérico (*float64*) y 13 columnas de tipo objeto. A priori, se puede observar también que el dataset tiene una gran cantidad de valores nulos que deberá ser tratada posteriormente.



Una vez hecho esto, aquellas columnas que se hayan importado como tipo numérico y sean de tipo categórica, se convierten a tipo factor de forma que sea posible realizar un análisis exploratorio correcto.

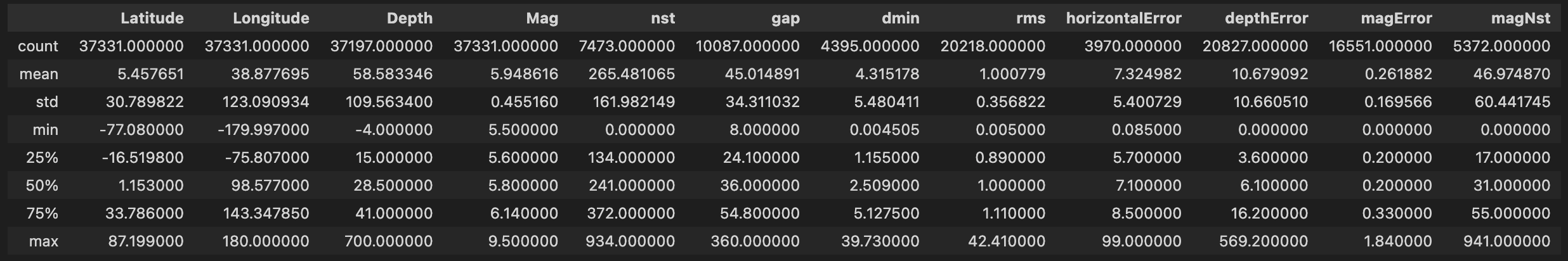
Se decide pasar a categóricas las columnas que tuvieran hasta 20 valores diferentes, para poder llegar a columnas como ‘MagType’ que tiene 18 valores diferentes.



## Descriptivo de Variables

Al ejecutar una primera descripción para sacar las medidas de centralidad de las variables, se puede ver que en general los valores no exceden de medidas naturalmente lógicas. Por ejemplo, en el rango de la longitud está entre +180 y -180, y que el de la latitud está entre +90 y -90. Por otra parte, valores como la profundidad, cuyos máximos son 700 km también está en rangos lógicos de la naturaleza.

El único valor extraño, que podría tratarse de un error, es el de la Depth mínima ya que es un número negativo. Esto indicaría que el terremoto se dio sobre el nivel del mar o que se debe a un error humano de interpretación. Esto se tratará posteriormente.



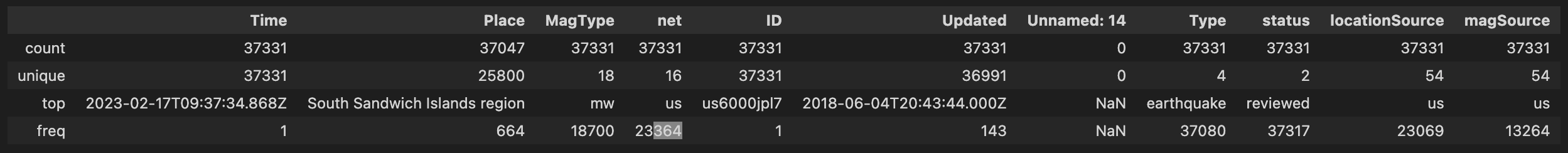
Los valores que arrojan en cuanto a **Magnitud** y **Profundidad** indican que los terremotos registrados tienen valores muy extremos en las colas derechas, ya que se encuentran muy lejos del rango intercuartílico 1Q-3Q y de la media.

Se aprecia también que esto se da en las variables que hablan del registro del evento sísmico, como **rms**, cuya media se sitúa en 1.00, pero su valor máximo es de 42.41, lo cual se trata sin duda de un valor atípico como se concluirá más tarde. Lo mismo sucede con otras variables como **MagNst**, **dmin** u **HorizontalError.**

El punto positivo es que podemos ver cómo los valores mínimos y el rango 1Q-3Q es muy compacto y uniforme en torno a la mediana. Sin duda, la media es superior a la mediana porque los valores extremos ubicados en la cola derecha hacen que esta se incremente, provocando así distribuciones asimétricas. Esto se estudiará posteriormente en profundidad con las gráficas de distribución.

En cuanto a las variables no numéricas, se puede ver a priori que algunas de las categorías están claramente dominadas por un registro. Es el caso del Tipo de Terremoto, que es en su gran mayoría ‘earthquake’ o en el tipo de Revisión, que será en su gran mayoría ‘reviewed’. También se puede comprobar que el valor predominante en las variables **net, locationSource** y **magSource** es ‘us’. Esto indicaría que uno de los centros que contribuye de forma más activa al registro de seísmos es el de Estados Unidos, ya que es el que más datos aporta.

Finalmente, es posible concluir que para la mayoría de categorías los datos están completamente desbalanceados. Los valores predominantes de **net, type, status** o **locationSource** suponen entre el 63% y el 99,9% de los valores de sus categorías.



## Cambios de formatos iniciales

Al importar el dataset original, se deben tomar varias decisiones iniciales para poder llevar a cabo el Análisis Visual de las Variables, puesto que en su formato original no es posible extraer información valiosa.

* **Time**: Variable originalmente de tipo string con un formato propio ( *YYYY-MM-DDT00:00:00.000Z* ). Se opta por pasar este string a tipo datetime64, pues de esta forma son valores interpretables para realizar gráficos.

Para llevar a cabo estas transformaciones se utiliza el método regex, de forma que extrae fácilmente el año, mes, día y hora para después juntarlo en un único string y convertirlo a dataetime64.

* **Updated**: Variable originalmente de tipo string con un formato propio ( *YYYY-MM-DDT00:00:00.000Z* ). Se opta igualmente por utilizar el mismo método que con la variable Time, ya que se tratan de formatos de origen idénticos.
* **Place**: Variable de tipo string muy poco heterogénea. A pesar de que la combinación de Latitud y Longitud aportan la misma información más precisa, parece interesante continuar manteniendo esta variable y ver cómo se puede homogeneizar. Va a resultar útil en las primeras fases del estudio, pues aporta información de forma mucho más rápida y comprensible que por coordenadas.

words\_ls = ['near', 'off', 'the', 'coast', 'of', 'border', 'region', 'central', 'southeast', 'southern', 'eastern', 'east', 'East', 'nothwestern', 'northeastern', 'western', 'southwest', 'south', 'South', 'north', 'North', 'Northern', 'west', 'western', 'West', 'northwest', 'southwest', 'southwest', 'southwestern']

Los resultados quedan almacenados en la variable **‘country’.**

## Análisis Visual de Variables

### Variables numéricas

Tras analizar numéricamente las variables, se procede a analizar sus gráficas de distribución.

* **Latitud - Longitud**: La distribución de la **Latitud** se asemeja a una distribución normal, sin presencia de outliers. La **Longitud** no tiene una distribución normal, pero no cuenta con outliers. Ambas son bastante simétricas y ligeramente negativas.

Así, se puede apuntar a que, en cuanto a la Latitud, los valores suelen estar en torno a la media, siendo bastante raros en sus extremos. En el caso de la Longitud, no sucede así puesto que los valores se concentran en la cola derecha. Esto hace que su media y su mediana estén bastante distantes y tenga una asimetría positiva.

Gráfico, Histograma

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

* **Depth**: Valores muy concentrados en un rango bajo de la distribución, pues están en torno a los 0-40 km de profundidad. Hay una gran presencia de outliers, puesto que hay valores muy extremos que llegan hasta los 700km. Es asimétrica positiva, como se verá más adelante con el coeficiente de asimetría, ya que está por encima de los 3 puntos.

Interfaz de usuario gráfica

Descripción generada automáticamente

* **Mag**: Concentración de terremotos en magnitudes bajas, muy concentrados en la zona de entre 5,7 – 6, que es donde se encuentran su media y mediana. También hay una gran cantidad de outliers en la cola derecha, puesto que los valores máximos se encuentran en torno a los 9.5º. No se aprecian outliers hacia la cola izquierda, puesto que los terremotos registrados son de magnitudes bajas y el dataset recoge aquellos a partir de 5.5º. Esto se confirma con el coeficiente de asimetría, ya que refleja una ligera asimetría positiva de 1,64.

Llama la atención la alta concentración en “picos” que aparecen en determinados valores. Esto se estudiará en profundidad más adelante.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

* **Otras variables que hablan de la calidad del dato:**
  + **Nst:** La media se encuentra muy cerca de la mediana, en torno a las 250 estaciones por registro. Es ligeramente asimétricamente positiva ya que se observa presencia de outliers en la cola derecha. Sin embargo, se puede decir que los valores están bastante centrados y se parece bastante a una distribución normal.

Gráfico, Gráfico de barras

Descripción generada automáticamente

* + **Gap:** Distribución parecida a la normal, pero con asimetría positiva debido a la aparición de muchos outliers en la cola derecha. Este hecho es favorable, puesto que cuanto más pequeño sea el valor del dato, mayor será la fiabilidad del registro del evento.

Imagen que contiene Gráfico

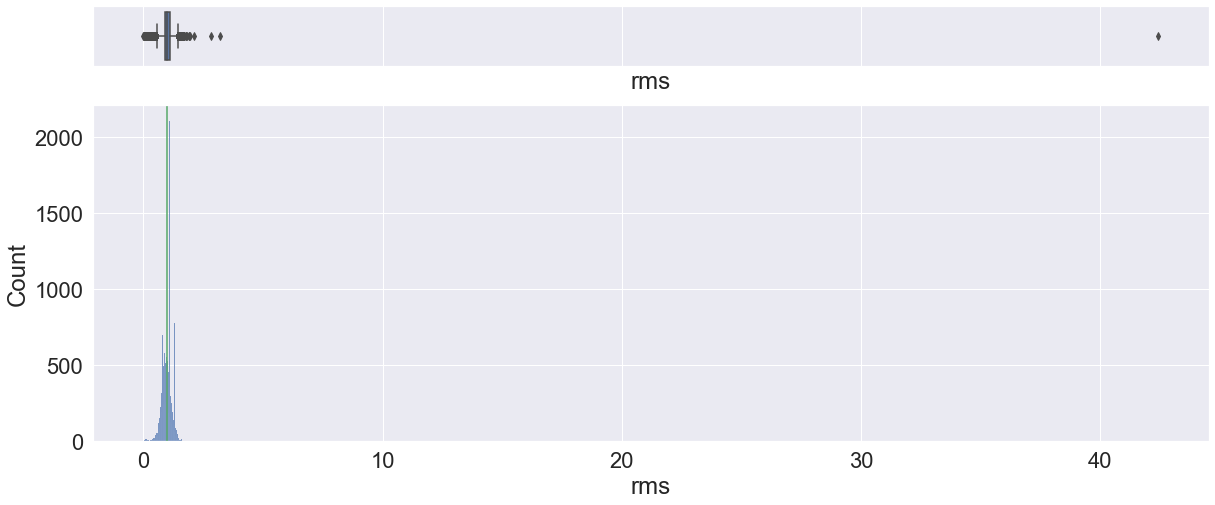
Descripción generada automáticamente

* + **Dmin:** También presenta una asimetría positiva con la presencia de outliers en la cola derecha. Este hecho también es positivo, puesto que los valores están en torno al 1Q de la distribución. Esto indica que las distancias habituales al epicentro del terremoto suelen oscilar entre 0 y 4 grados, siendo mejor la calidad del dato cuanto menos menores son las distancias al epicentro.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

* + **Rms:** Variable muy afectada por la presencia de un outlier de 42.41 siendo que su media está en 1.00. Apenas se puede ver nada en la distribución, puesto que su asimetría es muy grande situándose en torno a 77.



* + **horizontalErorr – DepthError - MagError:** Se observa una forma normal en cuanto al horizontalError, sin embargo, hay varios outliers que hacen que las distribuciones sean muy grandes y no estén centradas. Es más grave en el en caso de depthError puesto que hay dos valores muy extremos que trastocan completamente el análisis.

Independientemente de los outliers, es positivo ver que los errores sean relativamente bajos. Las gráficas arrojan unos errores de localización de 0-20 km y de unos 0-50 km de profundidad. Esto habla de un dato sólido y bien construido. Sobre todo, es interesante verlo en el caso de la distribución de la profundidad, ya que los valores están en torno al 1Q de la distribución.

El error de magnitud también arroja unos valores muy bajos, aunque no tenga una distribución normal. Se observa también la presencia de outliers en la cola derecha.

Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Imagen que contiene Tabla

Descripción generada automáticamente

### Variables categóricas y objeto

* **Magtype**: El valor más común es ‘mw’ con una clara diferencia. Este dato por sí sólo no aporta mucha información, pues se corresponde con el algoritmo que se utiliza para medir la magnitud del seísmo.

Imagen que contiene Teams

Descripción generada automáticamente

* **Net**: Tal y como se había previsto anteriormente, la mayoría de los seísmos son captados por redes americanas, puesto que los valores más repetidos son ‘us’. El segundo valor más común es el correspondiente al organismo del ISC-GEM, el Centro Sismológico Internacional. El resto de las categorías apenas son relevantes, puesto que apenas aportan registros.

Gráfico, Gráfico de barras

Descripción generada automáticamente

* **Type**: En cuanto al tipo de terremoto, la muestra se encuentra totalmente desbalanceada puesto que la amplia mayoría de registros es de tipo ‘earthquake’. La segunda categoría es la de explosiones nucleares, pero apenas representa un 0.006% de los registros.

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

* **Status**: La variable status también está totalmente desbalanceada, puesto que casi la totalidad de las entradas son revisadas por seres humanos, siendo los registros automáticos apenas 14 de los 37.331 registros del dataset.

Gráfico, Gráfico en cascada

Descripción generada automáticamente

* **Country**: El top 10 revela que la zona asiática del Mar Filipino es la más castigada por los terremotos, ya que las islas de Indonesia, Japón, Papúa, Filipinas, Vanuatu y las Islas Salomón se encuentran en esta región. Otras zonas con una gran actividad sísmica son el estado de Alaska (Estados Unidos) y Rusia.

Teams

Descripción generada automáticamente con confianza media

### Variables fecha

* **Date\_earthquake:** Alrealizar una representación gráfica de la fecha en la que se produjo el seísmo, se aprecia un claro trend de aumento. Comienza en 1900, con su valor más bajo y llega a valores máximos en la década de 2010. Se observa que no tiene una tendencia central, sino que se trata de una distribución multimodal con distintos picos de registros. A primera vista, se observan ciertos patrones de estacionalidad, por lo que se debe realizar un análisis en profundidad de esta variable para sacar más información en claro.

Gráfico, Histograma

Descripción generada automáticamente

* **Data\_updated**: Al modelar la fecha en la que se revisaron y actualizaron los datos, se ve un trend de valores a mantenerse cercanos a 0 y a concentrarse en torno al año 2022. Se observan también algunos patrones cíclicos de ciertos picos cada 1.5 - 2 años, mientras que entre estos picos hay valles cercanos a 0. Tal y como ocurre con la variable date\_earthquake, habría que realizar un estudio de Series Temporales para poder aumentar la información. A priori, el hecho de que las revisiones sean muy recientes, podría indicar que se tiene una gran calidad del dato ya que se encuentra actualizado.

Imagen que contiene Forma

Descripción generada automáticamente

## Análisis de relaciones con la variable objetivo Magnitud

Una vez analizadas las distribuciones de las variables, se analiza su relación con respecto a la **Variable Objetivo Magnitud**.

En un primer lugar, al no poder comprender bien la implicación de la Latitud y la Longitud por separado, se realiza un análisis conjunto en 2D a través de la librería scatter\_mapbox. El resultado deja ver, en primer lugar, que los registros de los seísmos siguen las placas tectónicas terrestres y que la mayor concentración de estos se encuentra en el Cinturón de Fuego del Pacífico y en el Cinturón Alpino.

Además, se modeliza teniendo en cuenta su magnitud, viendo como su aparición es **aparentemente aleatoria**. No se aprecia un patrón visual aparente, siendo la única observación apreciable la de que hay terremotos de mayor magnitud donde se observan más registros.

Mapa de colores

Descripción generada automáticamente con confianza media

Aparte de la localización de los seísmos, se opta por dibujar gráficos de las variables más interesantes:

Magnitude según LocationSource:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Magitude según MagType:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Magnitude según Type:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Magnitude según status:

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

Magnitude según MagSource:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gracias a estas representaciones se pueden obtener una serie de conclusiones:

* Los organismos gubernamentales que más seísmos detectan son, en primer lugar, el International Seismological Centre (ISC-GEM) seguido del Instituto Geológico Americano (USGS). Ambos organismos proporcionan información en un rango de magnitudes muy homogéneo, sin embargo, podemos ver que el USGS proporciona información con 1 decimal y el ISC-GEM con dos decimales. Podemos apreciarlo en el espaciado de la distribución, que se concentra en torno a valores con 1 decimal. Esto puede provocar un sesgo en la fase de modelización posterior.
* Los algoritmos que utilizan para calcular las magnitudes (variable MagType) también utiliza distintos niveles de decimales. El valor predominante de ‘mw’ es el único que utiliza dos decimales. Los otros algoritmos parecen utilizar solo 1 valor.
* El tipo de evento sísmico está relacionado con su magnitud. Las explosiones, tanto nucleares como no nucleares, generan terremotos de baja-media magnitud mientras que los eventos sísmicos provocan magnitudes en todo el rango de la escala Richter.
* El estado de una revisión guarda relación con la magnitud del evento sísmico. Se puede observar cómo los eventos de magnitud más alta no son revisados, sino que son publicados directamente por los automatismos de las redes de sistemas de detección.

Antes de dar por concluido el estudio de las variables input con respecto a la variable objetivo, se opta por **aplicar la técnica de la tabla V de Crammer**. Es interesante aplicar este método, puesto que permite identificar y cuantificar relaciones entre variables input y la variable objetivo de la magnitud. Esto facilita posteriormente la selección de las características más relevantes para la predicción, pues al comprender mejor estas asociaciones, es posible desarrollar modelos más precisos y eficientes.

Para establecer un “límite inferior” que marque la utilidad de las variables, se crean **dos variables nuevas aleatorias** y se incorporan a la gráfica. Esto sirve para saber si la relación entre las variables es fuerte es incluso más débil que una variable aleatoria. En caso de estar por debajo de este punto de corte, redundado a través de dos variables aleatorias, se estudiaría en profundidad para ver si pueden ser finalmente descartadas.

Gráfico, Gráfico de barras

Descripción generada automáticamente

El estudio de V de Crammer devuelve un gráfico de asociaciones débiles con respecto a la variable Magnitud, pues están por debajo de 0,2. La variable que tiene una mayor asociación con Magnitud es **date\_earthquake**, seguida muy de cerca por **net**. En tercer y cuarto lugar se tienen otras variables con una fuerte asociación: **MagType** y **Status**.

Aparte de la fecha, el resto de las variables más influyentes en la magnitud no aportan realmente información sobre el terremoto, si no sobre el registro de la información, por lo que se trata de información poco relevante.

Precisamente, **las variables que tienen una asociación más débil con respecto a su magnitud son las relacionadas con el evento sísmico en sí como Depth, Latitud, Longitud.**

Finalmente, el punto positivo que se extrae de este análisis es que todas las variables propias del dataset están por encima de las dos variables aleatorias. A priori, todas serían adecuadas para poder incorporarse a modelos de machine learning.

## Análisis de Correlaciones

Una vez analizadas las relaciones con la variable objetivo, se analizan las correlaciones a través del método numérico y del método visual.

El análisis arroja una correlación general muy baja entre las variables y con respecto a la variable objetivo. Los aspectos más destacables se exponen a continuación:

* **magNst – nst [ 0.7 ].** Correlación alta entre ambas variables, siendo esta relación positiva. Puede indicar que suelen utilizarse el mismo número de estaciones para construir valor de la magnitud que el de la localización del seísmo.
* **magError – magNst [ -0.4 ].** Correlación media entre el error de la magnitud y el número de estaciones sísmicas utilizadas para determinar su valor. Indica que están relacionadas de forma negativa, por lo que menor será el error cometido cuanto mayor sea el número de estaciones utilizadas para construir el registro.
* **Gap – depthError [ 0.4 ].** Correlación media entre la mayor brecha azimutal y el error en profundidad del terremoto. Al ser positiva, indica que ambos errores están relacionados y que cuanto mayor es uno, también es el otro.
* **Gap-nst** **[ -0.4 ]**. Correlación media entre la mayor brecha azimutal y el número de estaciones utilizados para determinar el seísmo. De nuevo, su naturaleza negativa indica que cuantas más estaciones se utilizan para determinar distintos aspectos de las naturalezas de los seísmos, menores son sus errores de lectura.
* **Mag-nst. [ 0.4 ].** Correlación media positiva. Indica que la magnitud está directamente relacionada con el número de estaciones que registran el seísmo. Quizá esta relación no sea orgánica, si no que ha sido el propio ser humano el que ha estaciones desplegado más estaciones en zonas activas sísmicamente ya que espera poder registrar una mayor cantidad de seísmos.

Se puede concluir el análisis con la reflexión de que **la correlación general es muy baja, a excepción de dos casuísticas que presentan correlaciones medias:**

1. **Correlaciones positivas entre los diferentes tipos de errores**, pues parece que el error en profundidad de la medición está asociado a un mayor error de ubicación del sismo.
2. **Correlaciones negativas entre errores y el número de estaciones utilizadas** para determinar las características de los eventos sísmicos, pues a un mayor número de estaciones de medida utilizadas, menores son los errores de interpretación proporcionan.

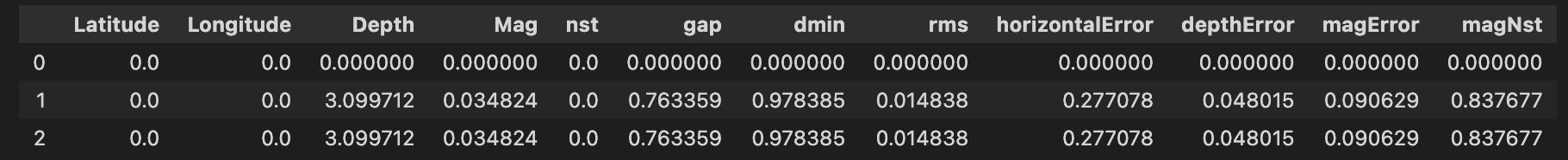
Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

## Análisis de Outliers

Dando por concluido el estudio de correlaciones, se procede a analizar la **incidencia de valores atípicos**. Tras el análisis matemático, se puede ver que hay una baja incidencia de outliers que no es indicativa de la presencia de varias poblaciones dentro de la muestra. A su vez, la mayor presencia de outliers se encuentra en la variable Depth, suponiendo el 3% de sus valores.

Otra apreciación interesante es que sólo tienen **outliers positivos**. Este hecho se complementa con las interpretaciones que se han ido realizando previamente a través de los gráficos de distribución de las variables, pues sólo se veía la presencia de valores anómalos en la cola positiva, es decir, a la derecha de la media.



Se considera necesario que en una primera aproximación se utilicen técnicas para el tratamiento de outliers, que centrarán la distribución y harán desaparecer los valores extremos.

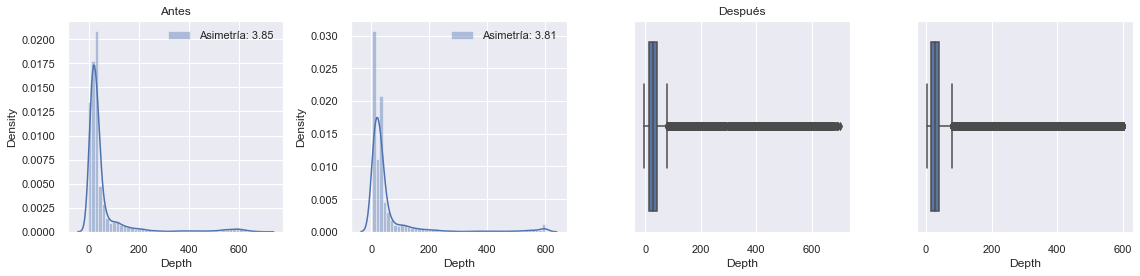
Así pues, se decide utilizar una técnica de **winsorización** **sobre el percentil 1%.** Ha arrojado los mejores datos de corrección de la asimetría, pues se ha estudiado también colapsar las colas sobre los percentiles 2,5%, 3% y 5%. Como se ha comentado previamente, los mejores resultados se consiguen con este percentil del 1%.

Gracias a esta técnica de winsorización, es posible bajar el valor de la asimetría en todas las variables, así como el porcentaje de outliers, que a excepción de Depth y magNst han bajado hasta 0.

Además, gráficamente se observa cómo las distribuciones se centran y la cola derecha es mucho menor asemejándose así a una distribución normal.

Imagen de la pantalla de un celular con letras y números

Descripción generada automáticamente con confianza baja



Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Captura de pantalla con fondo blanco

Descripción generada automáticamente con confianza media

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

Se seguirá observando cómo se comporta el modelo, ya que se puede ver que quizá los valores quedan demasiado centrados en la media y los modelos que se desarrollen no son lo suficientemente sensibles ante valores extremos.

## Análisis de Valores Missings

Tras aplicar esta técnica, se continúa el análisis exploratorio atendiendo a los valores missings.

La incidencia de valores missings es bastante elevada para algunas columnas, hecho que ha podido comprobar previamente en el Apartado 1.1.

Texto

Descripción generada automáticamente con confianza baja

Se presenta a continuación, un análisis visual de los valores missings de aquellas columnas en las que hay incidencia. Destacan las columnas **'nst', 'gap', 'dmin', 'horizontalError', 'magNst',** que presentan un alto número de valores faltantes, con porcentajes que oscilan entre el 79% y el 89% de datos ausentes. En el caso de ‘Unnamed: 14’ el valor aumenta hasta representar el 100% de los valores, por lo que esta columna es directamente descartada.

Por otro lado, columnas como **'Time', 'Latitude', 'Longitude', 'Mag', 'MagType', 'net', 'ID', 'Updated', 'Type', 'status', 'locationSource' y 'magSource'** no presentan valores missings, por lo que las convierte en atributos fiables para poder modelizar sobre los datos.

Icono

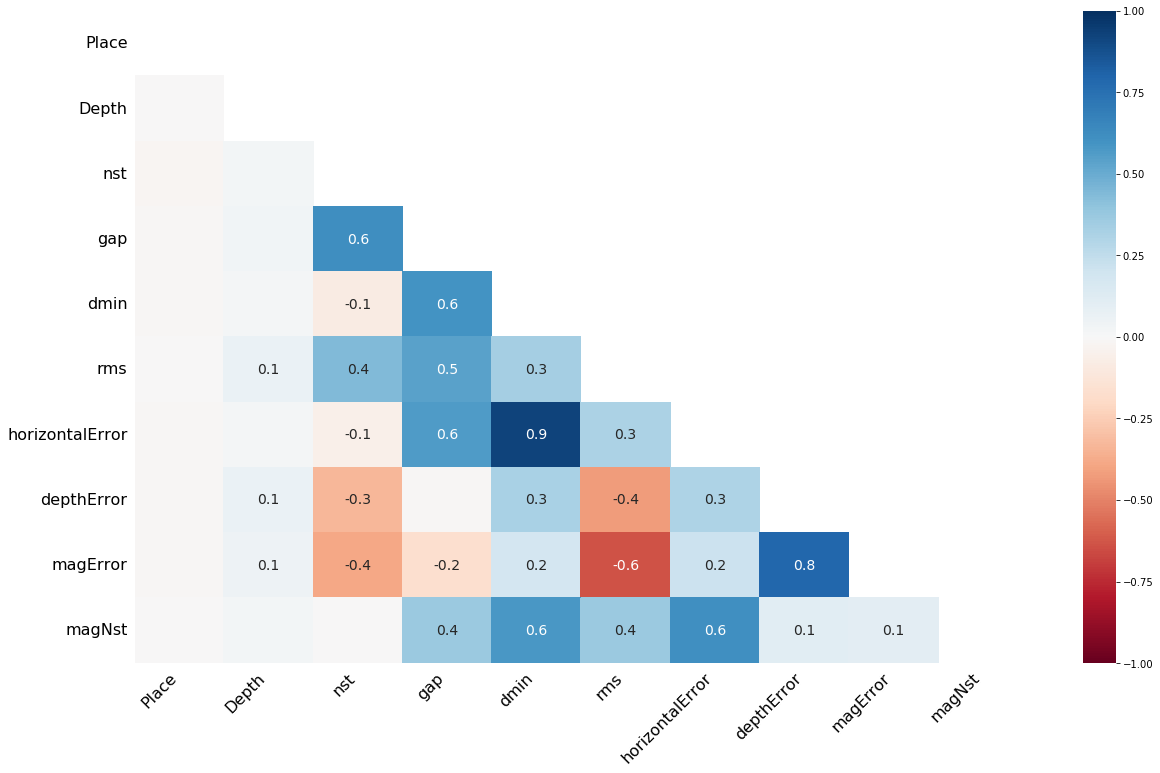
Descripción generada automáticamente

Se estudia a continuación, la correlación de los valores missings, **observando algunos patrones entre columnas**. Por ejemplo, se puede ver que las columnas depthError y magError mantienen un patrón con similitudes. Lo mismo ocurre con las columnas ‘horizontalError’ y ‘magNst’.

Aplicando un mapa de calor, se aprecia que la aparición de missings en las columnas horizontalError – dmin, magError – depthError están altamente correlacionadas ya que presentan valores de 0,8 – 0,9. También se tienen otras columnas con patrones de missing relacionados que presentan valores de 0,6.

Gráfico

Descripción generada automáticamente



**Se concluye así, que hay una correlación general en los patrones de missings.**

Se han estudiado diferentes métodos de imputación y tratamiento de los valores missings. En primer lugar, **se ha optado por eliminar aquellas columnas con una incidencia demasiado elevada** como para poder reconstruir datos con la calidad suficiente para tratarse en el apartado de modelización. Estas columnas tenían un porcentaje de 45 - 90% de valores missing y se correspondían con 'nst', 'gap', 'dmin', 'horizontalError', 'magError', 'magNst', 'rms', 'depthError'.

En un primer lugar, se intentaron reconstruir los datos a partir de otras columnas, pues existía una alta correlación entre algunas variables. Sin embargo, los resultados que arrojaba esta reconstrucción no eran satisfactorios, ya que los patrones de missings hacían imposible una reconstrucción fiel del dato.

Por otra parte, se ha aplicado la técnica de **imputación por KNN sobre la variable Depth** ya que presentaba un 0,35% de missings, lo cual es un porcentaje muy aceptable de reconstrucción.

Otra variable que presentaba valores missings era **country**, la variable string en la que se había almacenado previamente el país donde se produjo el evento sísmico. El porcentaje de missings que presentaba era de 0,76%, por lo que **se optó por reconstruir la variable a través de sus coordenadas de Latitud y Longitud gracias a la API de Geopy.** Ningún otro método de imputación, como el de imputar por la moda, tenía sentido por lo que se optó por la reconstrucción. El resultado de esta no fue adecuado, puesto que muchos eventos sísmicos se originaban fuera de los límites físicos de países, como en el mar. Esto devolvía valores ‘Unknown’ que la API Nominatim de Geopy no sabía interpretar.

Se acabó desechando la variable, pues la información ya estaba presente en el dataset a través de la combinación de Latitud + Longitud.

# Transformación de variables

## Análisis preliminar con PandasProfile Report

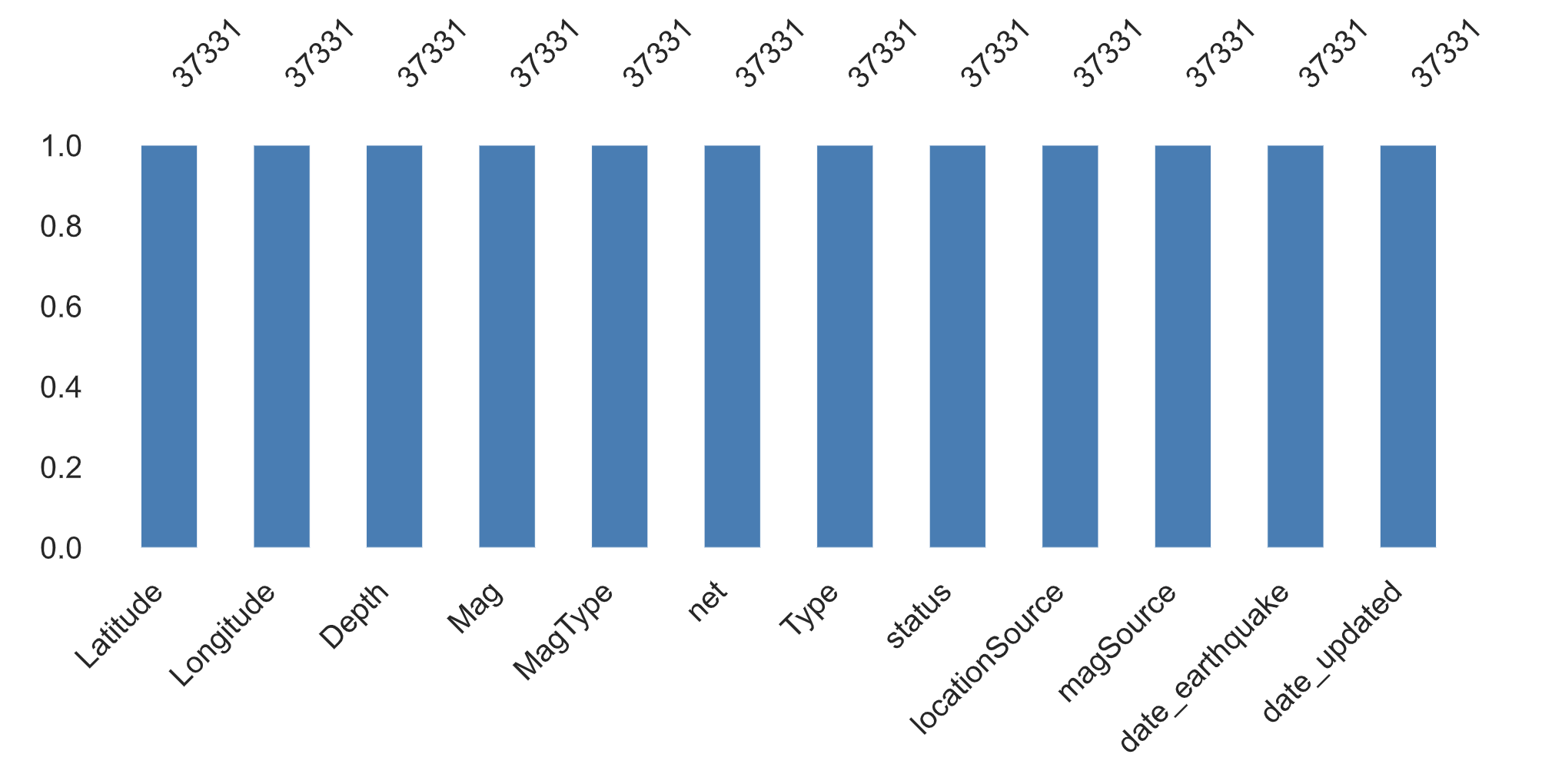
Antes de comenzar con la siguiente etapa del análisis, **se procede a realizar una revisión a través del reporte semiautomático de Pandas Profile Report.** Se observan varios puntos que indican que la primera fase de depuración ha sido llevaba a cabo con éxito.

* **Correlación:** Se siguen observando una correlación general baja con algunas excepciones, tal y como se ha visto durante el estudio preliminar. Las relaciones más altas se observan con variables no numéricas que arrojan información sobre la medición de los errores de los eventos sísmicos. Hay dos variables que presentan cierta multicolinealinedad (magSource-net y locationSource – net). No son variables que podrían ser de interés a priori, puesto que hablan de los errores de medición obtenidos y no métricas del evento sísmico en sí.

Gráfico

Descripción generada automáticamente

* **Missings**: Se aprecia una completa ausencia de valores missing, lo cual es muy positivo. Al haber descartado todas las variables que presentaban una alta falta de información y aplicar técnicas de imputación y reconstrucción de variables, se ha obtenido un dataset completo que permitirá aplicar técnicas de machine learning de forma adecuada.



* **Outliers**: Al haber aplicado técnicas de winsorización sobre variables que presentaban una alta presencia de valores atípicos, como Depth y la variable target Magnitud, se comprueba que se ha colapsado la cola derecha de la distribución, ya que los valores correspondientes a los extremos superiores han disminuido. Cuando se apliquen posteriormente modelos machine learning, esta centralidad hará que sean más robustos ante valores extremos de los datos a predecir. El aspecto negativo de aplicar estas técnicas es que pueden llegar a centrar demasiado la distribución y pueden hacer al modelo demasiado neutral e incapaz de predecir seísmos de grandes magnitudes.
* **Alta cardinalidad y Variables Desbalanceadas**: Se observan variables con alta cardinalidad, en concreto las variables de fecha y otras variables categóricas, que tienen hasta 54 variables diferentes. Esto indica que se tendrán que agrupar en diferentes categorías para encontrar un balance entre su capacidad para comprender las relaciones complejas y no sobre ajustarse a los datos.

Por otra parte, se observa que hay variables que están desbalanceadas, como net, type, status, location y magsource, por lo que se agruparán en categorías para generar una muestra más balanceada.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

## Transformación de variables numéricas

Con el objetivo de conseguir aumentar la linealidad de las variables input frente a la variable objetivo, se aplican diferentes transformaciones a las variables dependiendo de su naturaleza numérica o no.

En cuanto a las variables numéricas, se aplican las **transformaciones** **logarítmicas, exponencial, potencias y raíces cuadradas y cuarta.** Después, se analizan sus correlaciones con respecto a la variable Magnitud a través de una gráfica de correlaciones. Las transformaciones no mejoran apenas la correlación con respecto a la variable objetivo; lo hacen apenas unas centésimas.

Se escogen finalmente aquellas que presentan un mayor coeficiente de correlación. Sin embargo, esta baja correlación general con respecto a la variable objetivo augura unos modelos que no lograrán capturar los patrones subyacentes del dato y cuyo poder predictivo será muy limitado.

Captura de pantalla de computadora

Descripción generada automáticamente

Así, se decide continuar con las siguientes transformaciones, ya que han mejorado unas centésimas la correlación con respecto a la variable target.

* **latitude\_log\_co** -> La mejor transformación a aplicar es la logarítmica, lo que hará la distribución más simétrica.
* **longitude\_cuarta\_c** -> La mejor transformación a aplicar es la exposición a la cuarta. Exagerará la diferencia entre los valores, por lo que hará al modelo entender mejor las las diferencias de una distribución.
* **Depth\_sqrt\_co** -> La mejor transformación a aplicar es la raíz cuadrada. Es lógico que funcione bien con esta variable, puesto que tenía una marcada asimetría positiva y una cola derecha muy larga. La función cuadrática hará que la función se parezca más a la normal y estabilizará la varianza.

## Transformación de variables categóricas

En primer lugar, tras observar el diagrama de correlaciones entre las variables, se observa la colinealidad de las variables **net, locationSource, magSource y y status**. Presentan correlaciones muy altas y cercanas a 1, lo que puede llevar a problemas de multicolinealidad y, por tanto, a modelos débiles.

Se realiza un análisis de Crammer para variables categóricas y los resultados de este análisis reflejan una alta correlación entre net-locationSource y net-magSource. No presenta una alta correlación la asociación locationSource-magSource, por lo que se decide eliminar la variable **net.**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Después, se resuelve el problema de la alta cardinalidad que se ha observado desde el Análisis Descriptivo de Variables para las tipo object. Se opta por agrupar las categorías menos representadas, para que al pasarlas a dummies posteriormente, el modelo no tenga tantas variables. Se agrupan de la siguiente manera:

* **Type:** La categoría predominante de esta variable es earthquake, con casi el 99% de registros de tipo ‘earthquake’. Se ha agrupado por tanto en dos categorías, de forma que queda la categoría principal de ‘earthquake’ y la secundaria de OSA (Other Seismic Activities) agrupando explosiones nucleares o volcánicas. **De esta forma, Type pasa de 4 a 2 categorías.**
* **MagType:** La categoría predominante es ‘mw’, con casi el 50% de la representación muestral. Se decide hacer una nueva categoría agrupando al resto de las 17 categorías minoritarias cuyos registros independientemente son muy pocos. Así pues, se agrupan los registros en ‘mw’ y ‘OMT’ (Other MagTypes). La muestra queda así muy balanceada, en torno al 51%-49%. **De esta forma, MagType pasa de 18 a 2 categorías.**
* **magSource:** Hay varias categorías con una amplia representación, que son ‘us’ con el 36,4% de los registros, ‘iscgem’ con el 34,5% y ‘hrv’ con el 22%. El resto de categorías están muy poco representadas, por lo que se agrupan en ‘OMS’ (Other Mag Sources). De esta forma, **magSource pasa de tener 54 categorías a 4.**
* **locationSource:** Como sucede con magSource, hay dos categorías muy representadas que son ‘us’ e ‘iscgem’ con el 61,8% y el 34,5% respectivamente. Se respetan estas dos categorías y se crea una nueva para agrupar al resto de categorías que no aportan datos, que es ‘OLS’ (Other Location Sources). De esta forma, **locationSource pasa de tener 54 categorías a 3.**

**Imagen que contiene Texto

Descripción generada automáticamente**

## Selección de mejores transformaciones

Finalmente, para concluir con esta fase y pasar a la de modelizado, se analiza si las transformaciones llevadas a cabo han mejorado la asociación con respecto a la variable objetivo y son por tanto estadísticamente significativas. Para asegurar su correcta interpretación, se introducen dos variables aleatorias que funcionan como punto de corte del análisis.

En general, **el análisis muestra que las transformaciones de las variables afectan positivamente a la asociación y significancia con respecto a Magnitud.** Todas las variables transformadas (con sufijo ‘tr’) encabezan el ranking. Por ejemplo, MagType pasa de tener un coeficiente de 0,13 a 0,21. Por otra parte las transformaciones de las variables numéricas también mejoran, puesto que Depth pasa de 0.02 en su forma original a 0.06 con su transformación tomando raíces cuadradas.

**El aspecto negativo del ranking implica que las relaciones de todas las variables con respecto a Magnitud son muy débiles, en torno al 0.1.** Una de las variables que más información aporta sobre el sismo y más relacionada está con su magnitud es la fecha. Por el contrario, otras variables con la latitud y longitud del evento sísmico no tienen apenas relevancia para predecir la magnitud.

Algunas de las variables ni siquiera pasan el punto de corte de las variables aleatorias, como en el caso de Latitude\_log\_cr.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Finalmente, se hace un drop de las variables que aportan menos información al modelo, bien sea en su forma original o transformada.

Pantalla de computadora con números

Descripción generada automáticamente con confianza media

# Generación de modelos de regresión

## Ajuste Backwards de variables de regresión

Para ajustar los modelos de regresión, se trabaja con dos librerías y filosofías:

1. **Matriz de diseño (sci-kit learn):** Consiste en una amplia librería para aprendizaje de modelos de machine learning. Esta se caracteriza por ser muy rápida y ser capaz de implementar una amplia variedad de métricas e interfaces de remuestreo que podemos utilizar en este caso, para la validación cruzada repetida y GridSearch. Trabajamos con una matriz de diseño explícita, por lo que para obtener dicha matriz, vamos a extender a k-1 dummies.
2. **Interfaz fórmula (stat-models):** Proporciona un summary, constrastes, parámetros del modelo y de las variables para obtener información estadística. Así, se tiene el imput de predictores y el de la variable target final. El modelo se especifica mediante una fórmula

Con el fin de unir las dos filosofías, se utiliza la librería **pasty**, la cual hace de nexo entre ellas.

En primer lugar, se realiza un modelo completo de referencia con todos los predictores, para obtener de esta forma, el R2 máximo que se podría alcanzar con un modelo complejo que contase con todos los predictores. A raíz de ello, se van eliminando variables del modelo intentando que su R2 no baje en train. Es decir, intentar cumplir el principio de parsimonia, por el cual la solución más simple es la más probable y tener así modelos que no estén sobreajustados.

El objetivo es intentar ir eliminando efectos que no hagan que la capacidad predictiva del modelo baje para llegar a un modelo más simple que tenga la misma capacidad predictiva del modelo en train que el test.

El modelo de referencia que se crea tiene un R2 de 0.11. Esto es realmente desfavorable puesto que la varianza explicada por el modelo es muy inferior a la de la no explicada, ya que está explicando un 11% de la variabilidad del modelo. También se obtienen los parámetros AIC y BIC que sirven para comparar los modelos que se vayan prediciendo.

Además, se pueden ver los contrastes de cada uno de los parámetros que componen el modelo de referencia. Llama especialmente la atención el de ‘status’ y el de la ‘Depth’, puesto que no son 0. De hecho, status, está por encima de 0.05, por lo que se rechaza la hipótesis nula y se establece que el parámetro no es capaz de aportar para establecer la pendiente de la regresión logística.

Texto, Tabla

Descripción generada automáticamente con confianza media

Una vez que se ha determinado el punto de partida, se establece un modelo backwards para ir eliminando variables que no hagan que el R2 disminuya de forma efectiva. Para facilitar el proceso, se realiza además, un gráfico de importancia de la variable con respecto a la objetivo a través de la librería relativeImp, que ordena las variables en función de su aporte al R2 del modelo.

Gráfico

Descripción generada automáticamente

Tal y como se esperaba, las variables que se han conservado eran las que más aportaban al modelo: Type\_tr, MagType\_tr, magSource\_tr, locationSource\_tr, date\_earthquake, date\_updated.

Las variables que se eliminan son por tanto Depth\_sqrt\_cr, Latitude\_log\_cr, Longitude y status. Al eliminar estas 4 variables se aprecia cómo la complejidad del modelo baja considerablemente, mientras que el R2 del modelo baja apenas 2 milésimas y el AIC sube muy ligeramente. Finalmente, todos los estadísticos de contraste del modelo final son aceptados y aportan al modelo.

Texto

Descripción generada automáticamente con confianza media

Por último, se hace una predicción con el modelo final sobre los datos de test. De esta forma, se puede ver que el R2 del modelo no baja del 0.11 que era lo que se obtenía con los datos de train. Esto indica que el descarte de variables ha sido óptimo y que el modelo no está sobreajustado.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

## Modelos de regresión lineal

Teniendo en cuenta que la variable objetivo es continua, se crean diversos modelos de regresión lineal.

**Regresión Lineal**

En primer lugar, se comienza por aplicar un modelo básico de regresión lineal. Es interesante utilizar un modelo simple y con una alta interpretabilidad, ya que nos proporciona un punto de partida para después implementar modelos más complejos.

Además, aporta los coeficientes estimados en el modelo, lo cual proporciona información directa sobre la relación lineal entre el set de variables predictoras y la variable objetivo. De esta forma, es posible comprender qué variables aportaban poco, o incluso restaban interpretabilidad al modelo.

Al ser un modelo basado en la linealidad de las variables y la independencia de los errores, es posible que no logre capturar relaciones no lineales entre las predictoras. El resultado tan pobre de un R2 de 0.11 es esperable dado el concepto de baja correlación entre las variables, lo que hace pensar que no se cumplen las suposiciones de linealidad con respecto a la variable Magnitud y de independencia de errores.

Al representar la gráfica de residuos frente a los valores predichos, se puede ver un claro patrón espacial y, por tanto, una clara dependencia entre ellos que hace que no se cumpla el principio de linealidad del modelo. Además, atendiendo a la normalidad de los residuos del modelo, se aprecia que estos no siguen una distribución normal (gráfico Q-Q plot).

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Los resultados finales para este modelo son los siguientes:

* R2: 0.11015468422443597
* MAE: 0.3163985231614282
* MSE: 0.17934144351165066

Como era de esperar, sólo el 11% de la variabilidad puede explicarse con el modelo de regresión lineal. Además, el error absoluto (MAE) del modelo es de 0.316, lo que indica que el al predecir un seísmo se va en 0.316 grados en la escala de Richter.

**Ridge**

Tras probar con la regresión lineal, se opta por probar el modelo Ridge.

Este modelo añade una regularización basada en agregar como penalización la suma de los cuadrados de los valores de los coeficientes a la función de coste del modelo. Además, es posible regular la fuerza de la penalización a través de una constante alfa, la cual se ha ido modificando en diferentes iteraciones del modelo. Este planteamiento parece adecuado, ya que tal y como ha ido viendo durante el análisis de correlaciones de las predictoras, estas guardan correlación entre ellas. Se elimina una de las variables que era prácticamente combinación lineal de dos variables, pero aun así se dejan algunas variables que estaban bastante correlacionadas.

Se utilizan alfas grandes en un principio, para hacer que el modelo se simplifique al reducir el valor de los coeficientes. Sin embargo, se observa que el mejor R2 y el menor error se obtiene con alfas pequeños, en torno a 0-0.25. Esto indica que la regularización no funciona bien y que no produce ningún efecto la regularización L2. La penalización agregada no es lo suficientemente fuerte como para afectar a la magnitud de los coeficientes, lo que puede sugerir que los datos seleccionados finalmente no están sobreajustados de forma excesiva.

Como cabe esperar con un alfa tan débil, los valores obtenidos son muy similares a la regresión lineal:

* R2: 0.11014547827294652
* MAE: 0.3164147502353758
* MSE: 0.1793432988999971

Disminuye un poco el R2 y el MAE se mantiene igual indicando que el modelo no funciona y que los resultados son peores que los del modelo de regresión lineal.

**Lasso**

Al obtener un peor resultado con la técnica de Ridge, se decide probar también con la técnica Lasso. Esta guarda algunas similitudes con la técnica Ridge, pero utiliza otro tipo de regularización que puede llevar coeficientes a cero y, por lo tanto, realizar selección de las variables que se incorporan al modelo.

A pesar de haber tratado el tema de la alta cardinalidad de las variables categóricas, y por tanto, de sus dummies, se opta por probar la regresión Lasso para comprobar si alguna de las variables es irrelevante para el modelo y funciona mejor agregando este tipo de regularización.

Para ello, se ajustan diferentes modelos cambiando el valor de su alfa. Se prueban modelos con valores más altos, que aumentan la penalización y conducen a más coeficientes de las variables a cero. Después, se trabaja con alfas más pequeños, para incorporar todas las variables al modelo.

Los mejores resultados, se obtienen cuando no se aplica regularización al modelo. Se exponen a continuación los valores al aplicar una regularización mínima del 0.25.

* R2: 0.09422256526092432
* MAE: 0.32060022974753116
* MSE: 0.18255243890878345

Se concluye que las regularizaciones L1 (Ridge) y L2 (Lasso) no funcionan bien como modelos.

**Decission Tree**

Una vez concluido el estudio de modelos de Regresión Lineal Simple, Ridge y Lasso, se estudian otros modelos de árboles. Primeramente, se aplica el algoritmo de Árbol de Decisión (Decission Tree) aplicando profundidades de árbol diferentes.

A diferencia de los modelos testados anteriormente, los árboles de decisión pueden identificar las relaciones no lineales entre las diferentes características lo cual parece interesante a introducir en el estudio.

Aun así, conviene ser prudente en el uso de estas técnicas ya que profundidades de árbol muy grandes pueden llevar a un gran sobreajuste en los datos de train. Al probarlos sobre el test pueden llevar a resultados muy pobres o pueden ser muy sensibles a pequeñas variaciones en los datos de training.

Es por ello que se prueban profundidades de árbol en un rango bajo (1-10), obteniéndose los mejores resultados, que los ofrecía la profundidad de 8.

Ha resultado que estos modelos son capaces de captar una mayor varianza del modelo, pues los resultados que arroja son mejores tanto en R2 como en MAE/MSE.

* R2: 0.1830283279040521
* MAE: 0.2964207842655434
* MSE: 0.16465432405419167

El R2 aumenta casi un 7% y el error absoluto en la predicción de la magnitud baja 0.02 grados. Se trata por tanto de un mejor modelo.

**Random Forest:**

Se utiliza también, otro tipo de algoritmos de árbol más avanzados que el Decission Tree. En primer lugar, se usa el algoritmo de Random Forest, ya que supone una evolución del Decission Tree al basarse en la construcción de múltiples árboles de decisión y construir un ensemble de árboles de decisión. Al entrenarse cada árbol en un subconjunto aleatorio de los datos de training y características, parece a priori interesante puesto que podría aportar diversidad en los modelos.

Para poder ajustar más el modelo, se utiliza la técnica de GridSearchCV, una técnica de búsqueda de hiperparámetros para encontrar la combinación óptima que maximiza, en este caso, el R2 del modelo. En el caso del RandomForest, se ajustan los parámetros principales del modelo, que son el número de árboles del bosque aleatorio (100, 200, 300), la profundidad máxima de cada árbol de decisión (10, 20) y el tamaño mínimo de las hojas (1,2,4).

Finalmente, los mejores valores se obtienen con los siguientes hiperparámetros:

Best Hyperparameters: {'max\_depth': 10, 'min\_samples\_leaf': 1, 'n\_estimators': 300}

* R2: 0.19306609482293524
* MAE: 0.29558844947628066
* MSE: 0.1626312897391804

Se mejoran ligeramente los resultados del DecissionTree, lo cual era esperable ya el RandomForest es un modelo generalmente más robusto y reduce el riesgo de sobreajuste de datos producidos en un único DecissionTree.

**GradientBoosting Trees**

Para finalizar con el estudio de técnicas de árbol, se propone utilizar la mejora del algoritmo de RandomForest: el GradientBooting Trees. Este algoritmo también utiliza la combinación de predicciones de varios modelos para obtener una mejor predicción, aunque en este caso se construye de forma secuencial corrigiendo errores cometidos por árboles anteriores. Este procedimiento se sigue repitiendo, es decir, que se va corrigiendo el error y repitiendo el proceso hasta que el rendimiento del modelo se estabiliza.

Es interesante también utilizar diferentes hiperparámetros clave, al igual que en el caso del RandomForest. En este caso se usan los estimadores de número de árboles débiles que se utilizan en el aprendizaje por conjuntos (100, 200, 300), la profundidad máxima de estos árboles de decisión débiles (5, 8, 10, 15, 20) y la tasa de aprendizaje (0.01, 0.05, 0.1).

El GradientBoosting Trees, al tratarse de una evolución del RandomForest y de un algoritmo que entrena secuencialmente de los datos ajustándose a los errores del árbol anterior, ofrece una ligera mejora de los parámetros.

Best Hyperparameters: {'learning\_rate': 0.05, 'max\_depth': 5, 'n\_estimators': 200}

* R2: 0.1935198754599219
* MAE: 0.2964841569470546
* MSE: 0.1625398337602231

En esta ligera mejora de los resultados, el R2 aumenta indicando que explica una mayor proporción de la variabilidad de los datos. Sin embargo, el MAE también aumenta muy ligeramente lo que indica que el modelo está cometiendo errores mayores en promedio, lo cual es negativo.

**KNN:**

Tras finalizar con las técnicas de árboles, se prueba con el algoritmo de KNN, de forma que se calcula el valor ponderado de los k vecinos más cercanos para un nuevo conjunto de datos. Conociendo los datos de los que se partía, no era un algoritmo pareciese interesante a priori, pero se decidió introducir en el estudio.

Para ello, se prueban diferentes hiperparámetros. Cabe destacar el uso del hiperparámetro principal, que es el del número de vecinos más cercanos que se consideran para hacer una predicción. Además, se ajusta también la ponderación de los pesos, de forma que se prueba una distribución uniforme y por distancia (siendo que los vecinos más próximos tienen un mayor peso).

Como punto de partida para el estudio de hiperparámetros, se toma un n\_neighbors de la raíz cuadrada de las observaciones, es decir, unos casi 194 observaciones. Sin embargo, se observa que el modelo se comporta mejor con un número de vecinos mucho más pequeño, en torno a 60.

Finalmente, los resultados con los mejores hiperparámetros son los siguientes:

Best Hyperparameters: {'n\_neighbors': 60, 'weights': 'uniform'}

* R2: 0.13857628664065957
* MAE: 0.3081592339627696
* MSE: 0.17361328928768063

Es un tipo de modelo cuyo R2 está muy por debajo de las técnicas de árbol, pero algo por encima que las técnicas de regresión lineal.

**XGCBoost:**

Por último, se utiliza el algoritmo de aprendizaje automático de XGCBoost, puesto que se utiliza ampliamente en problemas de regresión y supone una mejora a los algoritmos de bosque que probados anteriormente tanto en eficiencia como en precisión. Además, al igual que en algoritmos utilizados anteriormente, selecciona las características más importantes de modelo gracias a que utiliza funciones de pérdida personalizadas, lo que permite adaptar el algoritmo este problema específico.

Se prueba este algoritmo, además, porque es capaz de capturar interacciones no lineales entre características, lo que lo hace adecuado para este problema en particular ya que se trata de una problemática compleja.

Finalmente, cabe destacar que se utiliza de nuevo GridSearchCV para buscar los hiperparámetros adecuados para este modelo de regresión. La métrica que se optimiza durante el entrenamiento es la del MSE, pues mide el promedio de los cuadrados de las diferencias entre las predicciones y los valores reales, por lo que es adecuada para un problema de regresión. Como se menciona anteriormente, los hiperparámetros que se utilizan son el learning rate, que controla la distribución de cada árbol (0.05, 0.1, 0.15, 0.2), el número de árboles que se utilizan en el ensamble (100, 200, 300) y la profundidad máxima de cada árbol (de 0 a 5).

Best Hyperparameters: {'learning\_rate': 0.15, 'max\_depth': 7, 'n\_estimators': 50}

* R2: 0.19472919858630988
* MAE: 0.2963032810726654
* MSE: 0.16229610403404096