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 NEIC aporta información sobre diferentes aspectos que recogen las sondas que registrar eventos sísmicos en todo el mundo.

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 <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.

Se añaden algunos comentarios procedentes del primer análisis exploratorio en cuanto a la forma de los registros.

* **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, vemos que hay 10 columnas de tipo numérico (float64) y 13 columnas de tipo objeto. A priori, podemos ver también que el dataset tiene una gran cantidad de valores nulos que deberá ser tratada posteriormente.

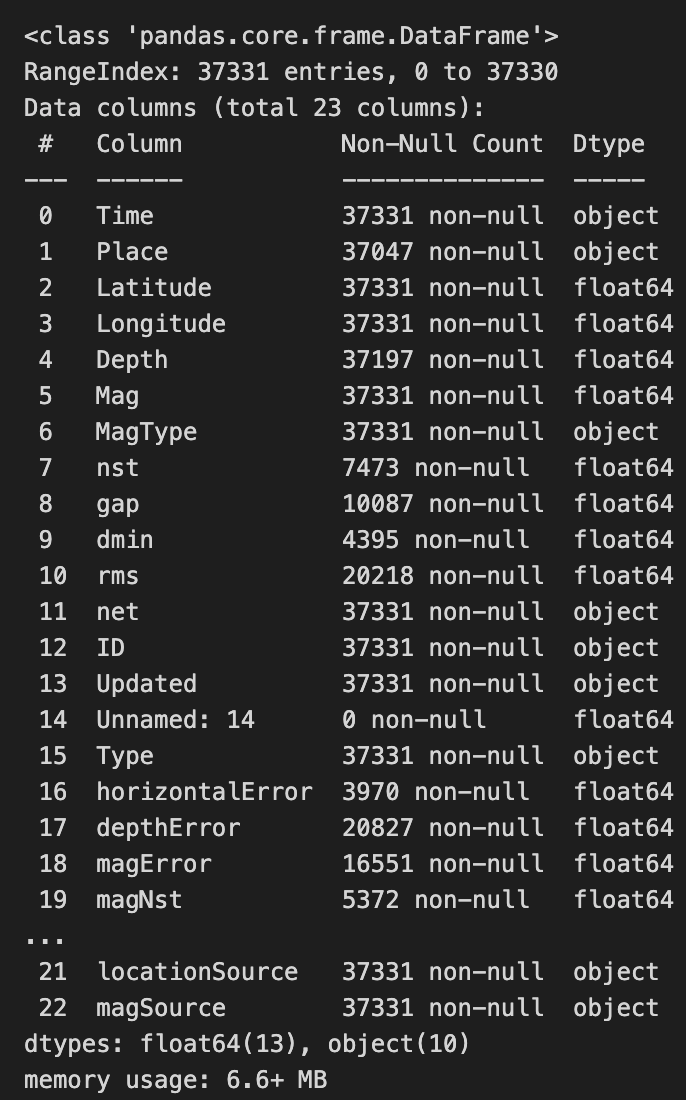
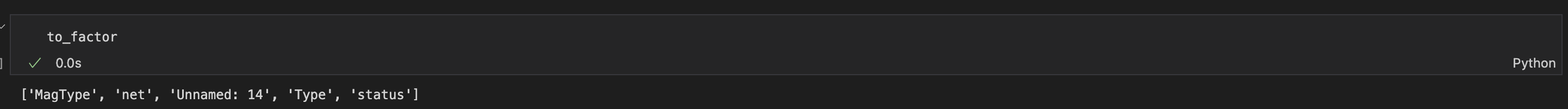


Fig. 1: Información sobre del DF

Después, voy a pasar a tipo factor aquellas columnas que se hayan importado como tipo numérico y sean categórica, de forma que podremos realizar un análisis exploratorio correcto.

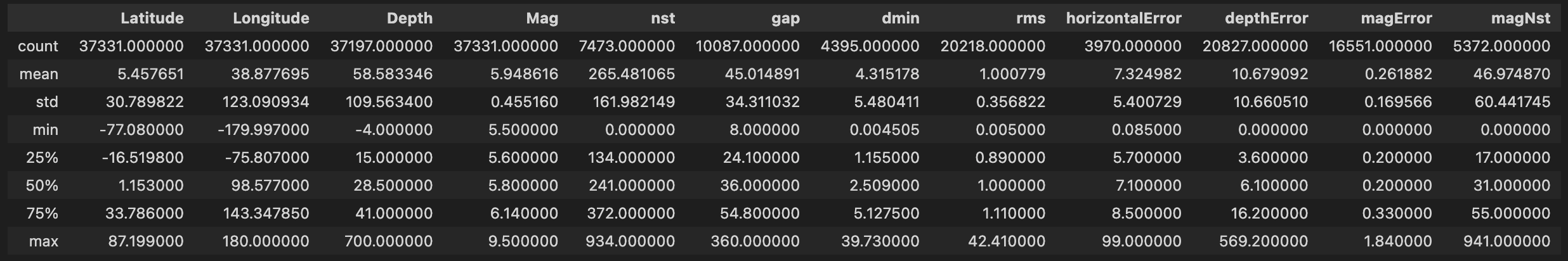
He decidido pasar a categóricas las columnas que tuvieran hasta 20 valores diferentes, para poder llegar a columnas como ‘MagType’ que tiene 18 valores diferentes.



## Descriptivo de Variables

Al ejecutar una primera descripción para sacar las medidas de centralidad de las variables, vemos que en general los valores no exceden de medidas naturalmente lógicas. Por ejemplo, vemos que el rango de la longitud está entre + 180, - 180 y que el de la latitud está en +90, -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. Lo trataremos posteriormente.



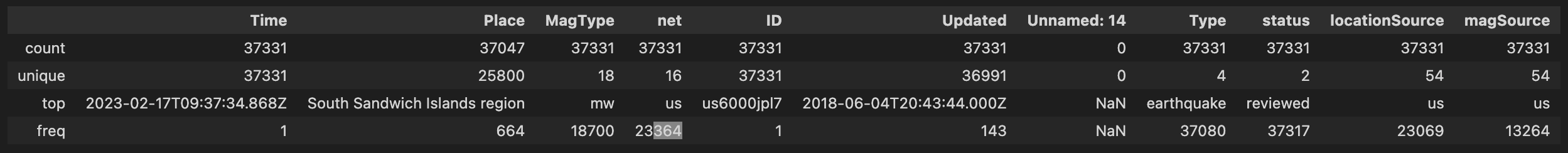
Los valores que arrojan en cuanto a **Magnitud** y **Profundidad** indican que los terremotos registrados tienen valores muy extremos en las colas derechas, ya que se encuentran muy lejos del rango intercuartílico 1Q-3Q y de la media.

Vemos 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 veremos 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 media porque los valores extremos ubicados en la cola derecha hacen que esta se incremente, provocando así distribuciones asimétricas. Lo estudiaremos posteriormente en profundidad con las gráficas de distribución.

En cuanto a las variables no numéricas, podemos 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 podemos 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 podemos concluir que para la mayoría de categorías vemos que 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, he tenido que tomar varias decisiones iniciales para poder llevar a cabo el Análisis Visual de las Variables, puesto que en su formato original no era posible extraer información valiosa.

Se exponen a continuación los cambios llevados a cabo:

* **Time**: Variable originalmente de tipo string con un formato propio ( *YYYY-MM-DDT00:00:00.000Z* ). He optado por pasar este string a tipo datetime64, pues de esta forma son valores interpretables para realizar gráficos. He optado por eliminar los minutos y segundos de los eventos sísmicos, de forma que los resultados queden más compactados.

Para llevar a cabo estas transformaciones he utilizado el método regex, de forma que he podido extraer 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* ). He optado 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, me ha parecido interesante continuar manteniendo esta variable y ver cómo podía homogeneizarla. Va a resultar útil en las primeras fases del estudio, pues aporta información de forma mucho más rápida y comprensible que por coordenadas.

Para llevar a cabo las transformaciones, he utilizado el método regex. Primero, me he quedado sólo con la parte del string del país donde se produjo el terremoto, que se corresponde con las últimas palabras tras la coma. Después, he realizado una limpieza de esos valores para eliminar palabras referentes a la zona geográfica (norte, sur, este, oeste) del país donde se produjo el terremoto.

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

Tras analizar numéricamente las variables, he procedido 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.

Podemos apuntar que en cuanto a la Latitud, vemos que los valores suelen estar en torno a la media, siendo bastante raros en sus extremos. En el caso de la Longitud, no sucede así pues que podemos ver que los valores se concentran en la cola derecha. Esto hace que su media y su mediana estén bastante distantes y tenga una cierta 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. Podemos ver una gran presencia de outliers, pues hay valores muy extremos que llegan hasta los 700km. Es asimétrica positiva, como se puede ver 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 podemos ver una gran cantidad de outliers en la cola derecha, puesto que los valores máximos se encuentran en torno a los 9.5º. No vemos outliers hacia la cola izquierda, puesto que los terremotos registrados son de magnitudes bajas. 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 podemos ver en determinados valores. Estudiaremos la causa en profundidad más adelante.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

* **Otras variables menos interesantes que hablan de la calidad del dato:**
  + **Nst**: La media se encuentra muy cerca de la mediana, en torno a las 250 estaciones que registran cada seísmo, por lo que estamos ante una distribución bastante simétrica. Es, sin embargo, ligeramente asimétricamente positiva ya que observamos presencia de outliers en la cola derecha. Sin embargo, podemos 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 bastante positivo, pues cuanto más pequeño sea el valor del dato, mayor será la fiabilidad del registro del evento.

Imagen que contiene Gráfico

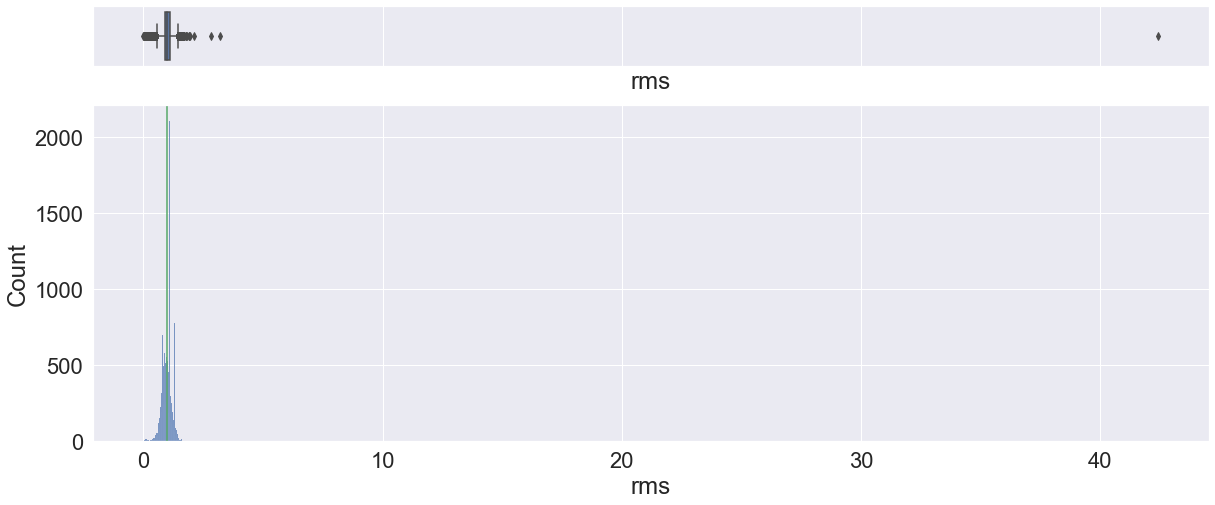
Descripción generada automáticamente

* + **Dmin**: Presenta una clara asimetría positiva con la presencia de outliers en la cola derecha. Este hecho también es positivo, puesto que vemos 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 podemos ver nada en la distribución, puesto que su asimetría es muy grande situándose en torno a 77.



* + **horizontalErorr – DepthError - MagError**: Observamos 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 porque hay dos valores muy extremos que trastocan completamente el análisis.

Independientemente de los outliers, es positivo ver que los errores son relativamente bajos. Las gráficas arrojan unos errores de localización de 0-20 km y de unos 0-50 km de profundidad. Esto nos habla de un dato sólido y bien construido. Sobre todo, es interesante verlo en el caso de la distribución de la profundidad, pues 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. Observamos 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

* **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 hemos podido prever 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 Seismológico Internacional. El resto de categorías apenas son relevantes.

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 esta totalmente desbalanceada, puesto que casi la totalidad de las entradas son revisadas por seres humanos, dejando las notificaciones automáticas muy restringidas.

Gráfico, Gráfico en cascada

Descripción generada automáticamente

* **Country**: El top10 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

* **Date\_earthquake:** Alrealizar una representación gráfica de la fecha en la que se produjo en seísmo, podemos apreciar un aumento claro a lo largo de los años. A priori, podría indicar un mayor aumento de la actividad sísmica pero podría haber otras razones para tumbar esta teoría, como el gran avance de la tecnología de detección sísmica y su aportación de más datos al modelo.

Gráfico, Histograma

Descripción generada automáticamente

* **Data\_updated**: Al modelar la fecha en la que se revisaron y actualizaron los datos, podemos ver una gran cantidad de revisiones en el año 2022. Esto nos podría indicar que estamos ante una gran calidad del dato, pues ha sido revisado de forma constantemente y está actualizado.

Imagen que contiene Forma

Descripción generada automáticamente

## Análisis de relaciones con la variable objetivo Magnitud

Una vez analizadas las distribuciones de las variables, es interesante analizar 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 con respecto a la variable objetivo por si solas, he decidido realizar una análisis en 2D gracias a un mapamundi.

El resultado nos hace 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, la magnitud de los seísmos podemos ver que también viene indicada por la ubicación de estos, pues son más fuertes en estas dos zonas de gran actividad sísmica. En otras zonas, como la placa africana o en el centro de Europa, los seísmos son mucho menos intensos.

Mapa de colores

Descripción generada automáticamente con confianza media

Aparte de la localización de los seísmos, he optado por dibujar gráficos de las variables más interesantes.

Magnitud según LocationSource:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Magitud 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 hemos podido obtener una serie de insights, que son los siguientes:

* 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 concrentra 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 también está fuertemente 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 también parece guardar relación con la magnitud del evento sísmico. Podemos ver 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. Esto quizá podría apuntar a un errores de lectura, fallos de los sistemas de gestión de la información o, por el contrario, publicación de datos históricos consolidados de forma automática.

## Análisis de Correlaciones

Una vez analizadas las relaciones con la variable objetivo, analizamos 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 reseñables 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.

Podemos 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 un error en profundidad está asociado a un mayor error de ubicación del sísmo.
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 sistemas utilizados menores errores de interpretación proporcionan.

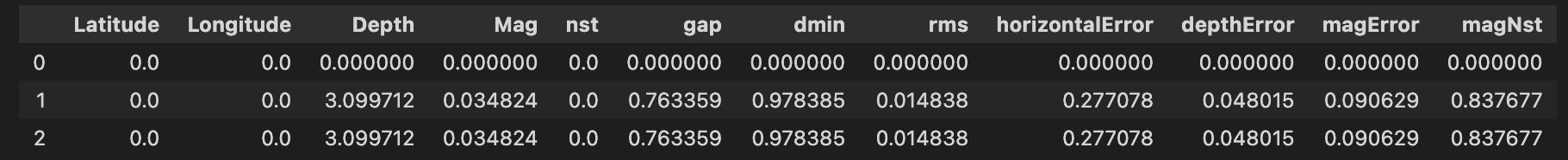
Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

## Análisis de Outliers

Dando por concluido el estudio de correlaciones, pasamos a analizar la **incidencia de valores atípicos**. Tras el análisis matemático, podemos ver que estamos ante una baja incidencia de outliers que no es indicativa de la presencia de varias poblaciones dentro de la muestra. A su vez, vemos que 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 hemos ido haciendo previamente a través de los gráficos de distribución de las variables, pues sólo veíamos presencia de valores anómalos en la cola positiva, es decir, a la derecha de la media.



Veo necesaria en una primera aproximación utilizar técnicas para el tratamiento de outliers, que centrarán la distribución y harán desaparecer los valores extremos.

Así pues, he decidido utilizar una técnica de **winsorización** sobre el percentil 1%. He podido hacer más pruebas (sobre percentiles 2,5%, 3% y 5% pero los mejores resultados los he conseguido con este 1%).

Gracias a esta técnica de winsorización, ha sido 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 podemos observar 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

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente Gráfico

Descripción generada automáticamente Gráfico

Descripción generada automáticamente Gráfico, Histograma

Descripción generada automáticamente Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente Gráfico

Descripción generada automáticamente Gráfico

Descripción generada automáticamente

Quedaremos atentos de cómo se comporta el modelo, pues podemos ver que quizá los valores han quedado demasiado centrados en la media y los modelos que desarrollemos no son lo suficientemente sensibles ante valores extremos.

## Análisis de Valores Missings

Tras aplicar esta técnica, he optado por continuar el análisis exploratorio atendiendo a los valores missings.

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

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 será 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

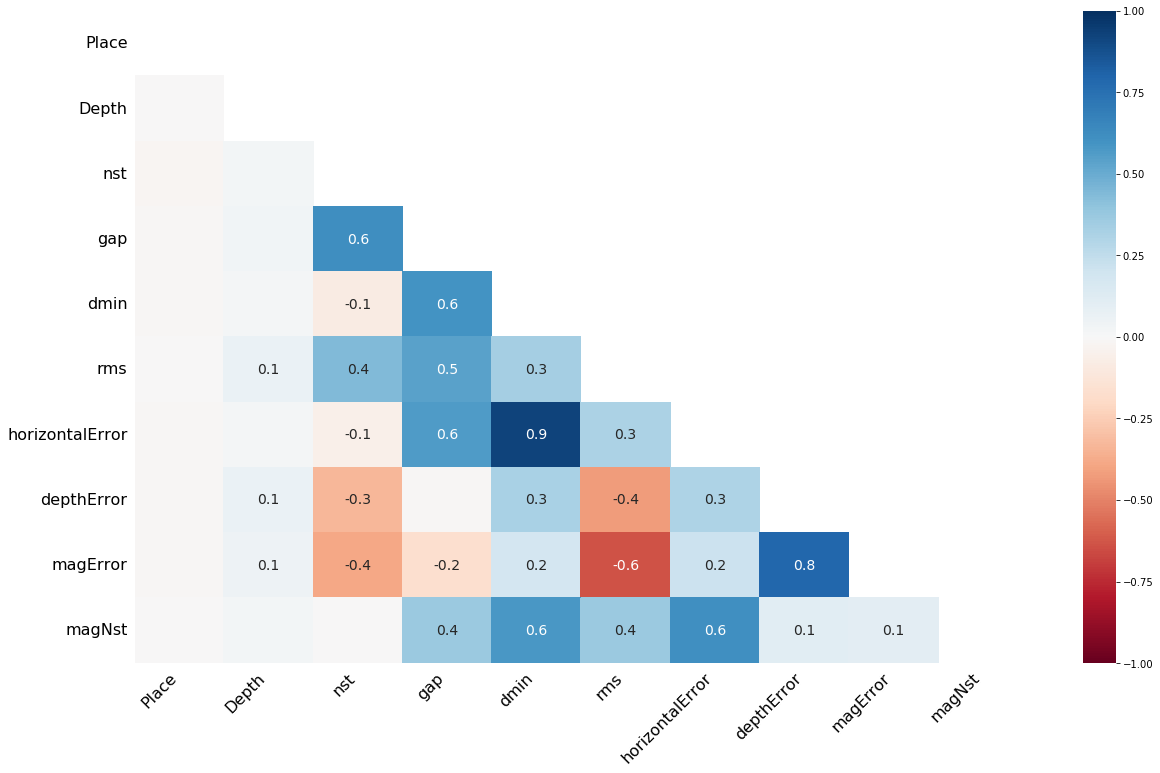
Descripción generada automáticamente

Se estudia a continuación la correlación de los valores missings, **observando algunos patrones entre columnas**. Por ejemplo, podemos 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, observamos 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 vemos otras columnas con patrones de missing relacionados que presentan valores de 0,6.

Gráfico

Descripción generada automáticamente



**Concluimos entonces que podemos observar una correlación general en los patrones de missings.** Se estudiarán diferentes métodos de imputación y tratamiento de los valores missings en apartados posteriores. Además, se eliminarán aquellas columnas con una incidencia demasiado elevada como para poder reconstruir datos con la calidad suficiente como para tratarse en apartados posteriores de modelización.

Había determinadas columnas que tenían demasiados missings como para imputar los valores, ya que me estaría inventando los valores para entre 45-90% de los casos y no tiene sentido. Tampoco podía reconstruir los valores a raíz de otras columnas, ya que las columnas con missings están altamente correlacionadas.

La única columna que presentaba missings era Depth con un 0,35%, por lo que la imputé a través de los knn vecinos más próximos.

Por otra parte, intenté utilizar la API de Geopy para intentar poner los valores de country que estaban missing. Tampoco eran muchos, eran sobre 280, por lo que intenté utilizar Geolocator de Geopy para extraer el país ya que ningún otro método de imputación tenía sentido (ni moda, ni nada). Entonces sólo pudo encontrar unos 80 valores, porque el resto al haberse producido fuera de los límites de los países, por ejemplo en zonas de mar, no conseguía determinar el país.

Acabé descartando la variable porque bien es cierto que la información sobre la ubicación ya está recogida en la latitud + longitud y esta es redundante y menos acertada.

Texto

Descripción generada automáticamente con confianza baja

Después he hecho un pandas profile report.

* Correlación -> bajas correlaciones entre variables numéricas, pero algunas correlaciones altas en las variables no numéricas (categóricas y tal). Todas las correlaciones son positivas. No llegan a ser demasiado altas a excepción de dos variables que presentan cierta multicolinearidad (magsource-net , locationsource-net). Quizá no son las variables que más nos podrían interesar a priori porque son variables que recogen la información de los sucesos (redes que los detectan, fuentes de información,…) y no métricas del evento sísmico en sí.
* No hay valores missing, esto es positivo. Al haber descartado todas las variables que presentaban mucha falta e información tenemos finalmente un dataset de datos muy poblados y que nos permitirá aplicar técnicas de machine learning mejor.
* Al habernos cargados los outliers con técnicas de winsorización en variables como Depth y la variable target ‘mag’ hemos hecho que colapse la cola derecha de la distribución. Esto ha hecho que baje significativamente el valor de los extremos y creo que cuando se apliquen modelos esto hará que sean más robustos ante valores extremos. Lo malo es que creo que va a centrar demasiado lel modelo y va a hacer que los cálculos no lleguen a calcular seísmos de grandes magnitudes.

Después de esto, separo ya entre la variable target y el resto.

Hago una tabla V de Crammer para ver la asociación entre variables.

Aplicar Cramer's V en el dataframe es especialmente crítico para el desarrollo de modelos de aprendizaje automático. Esta medida estadística nos permite identificar y cuantificar relaciones entre variables categóricas y la variable objetivo de la magnitud. Esto facilita la selección de las características más relevantes para la predicción, pues al comprender mejor estas asociaciones, es posible desarrollar modelos más precisos y eficientes.

En última instancia, mejorará así la capacidad del modelo para pronosticar la magnitud de los terremotos con mayor exactitud.

Se implementan dos variables nuevas aleatorias para que hagan de “punto de corte” con respecto a las variables del dataset. Esto nos sirve para saber si la relación entre las variables es fuerte o es más inútil que la de los aleatorios.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Vemos que una de las más importantes es la fecha en la que se produjo el terremoto, seguida muy de cerca por el contribuidor de la información.

Otras variables muy relacionadas con la magnitud son el algoritmo utilizado para medir la magnitud y su estatus de revisado o no revisado. La de magtype y net precisamente será porque la proporción de valores está muy centrada en varios valores muy concretos? No lo sé, esto me lo he inventado un poco.

Aparte de la fecha, el resto de variables más influyentes en la magnitud no aportan realmente información sobre el terremoto, si no sobre el registro de la información. Es un poco decepcionante.

Las variables que menos aportan son, precisamente, las relacionadas con el evento sísmico (profundidad, latitud, longitud y profundidad).

El punto positivo es que todas están por encima de las dos variables aleatorias, por lo que todas son buenas para poder modelar los modelos de machine learning.

Gráficos de variables categóricas frente a la Magnitud:

Transformación de variables:

Después de revisar las relaciones entre las variables, pruebo a hacer algunas transformaciones a ver si así la

En general, volvemos a observar una baja relación lineal 2 a 2 presente en las variables. Esto implica que las variables continuas no van a tener una fuerte influencia en la variabilidad de los modelos de regresión y que será más fuerte la relación con respecto a las variables categóricas.

Por otra parte, viendo el aspecto positivo, podemos ver que hay una ausencia general de correlación que evitará posteriormente en la fase de modelado los problemas de colinealidad en los modelos que desarrollemos.

He probado a transformar las variables a través de VCramer y coef. De correlación de Pearson con las transformaciones habituales ( logarítmica, exponencial, y raíces cuadradas y cuartas).

Finalmente, tras elegir los mejores resultados de cada una de las imputaciones (Cramerr y coficiente de Pearson) lo he puesto todo en una tabla de correlación para ver si sube con alguna de las transformaciones.

La verdad es que no suben una maravilla, apenas unas centésimas. Pero bueno, parece que mejoran con las mejores transformaciones de Pearson. Sobre todo la Depth, que sube la que más.

Captura de pantalla de computadora

Descripción generada automáticamente

Finalmente pues, me quedo con las siguientes transformaciones:

latitude\_log\_co -> Hará la distribución más simétrica y bajará la asimetría.

longitude\_cuarta\_c -> Magnificará las diferencias entre las variables, por lo que hará al modelo a entender mejor las diferencias que estarán más exageradas.

Depth\_sqrt\_co -> Es lógico puesto que esta variable era asimétrica positiva, tenía una cola derecha muy larga. La función cuadrática hará que la función se parezca más a la normal y estabilizará la varianza.

# Análisis Descriptivo

<https://es.wikipedia.org/wiki/Cinturón_alpino>

<https://es.wikipedia.org/wiki/Cinturón_de_fuego_del_Pac%C3%ADfico>

<https://github.com/rachittoshniwal/machineLearning/blob/master/Winsorizing%20outliers.ipynb>