Análisis de Sismos

Certamen 1: Ciencia de Datos en la Terminal Linux

Tomas Saavedra

2025-11-01

Contents

List of Figures

List of Tables

1. Obtención de Datos

La fase inicial consistió en la adquisición, transferencia y preparación del conjunto de datos sísmicos para el análisis. Este proceso aseguró que los datos estuvieran accesibles dentro del entorno de análisis (`dsitcl`) y en un formato manejable.

1.1 Adquisición y Transferencia del Conjunto de Datos.

El conjunto de datos (*data.csv*), proporcionado por el USGS, fue descargado manualmente. Posteriormente, se utilizó el comando `csp` (secure copy) para transferir el archivo de forma segura al servidor.

```
# Transferencia del archivo local al servidor
scp 'data.csv' tsaavedra@146.83.193.254:~/
```

Una vez en el servidor, el archivo fue movido al directorio *DSITCL*. Este directorio está mapeado a `/data` dentro del contenedor Docker, garantizando el acceso a los datos desde el contenedor.

```
# Mover data.csv al directorio
mv 'data.csv' DSITCL/
```

Finalmente, para una correcta organización del proyecto, se creó un directorio de trabajo (*certamen1*) y el conjunto de datos se movió a esta ubicación

```
# Crear y organizar el directorio de trabajo
mkdir certamen1
mv 'data.csv' certamen1/
```

1.2 Verificación de la Estructura de Datos.

El primer paso del análisis fue inspeccionar la estructura del archivo. Para ello, se accedió al directorio de trabajo dentro del contenedor (`/data/certamen1`).

Para obtener los nombre de las columnas, se utilizó `csvcut -n` para extraer una lista robusta de todas las columnas del conjunto de datos. El resultado fue redirigido y guardado en `columns.txt`.

```
# Ingresar al directorio de trabajo
cd /data/certamen1/

# Guardar la lista de nombres de columnas
csvcut -n data.csv > columns.txt
```

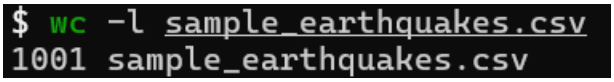
1.3 Creación de una Muestra Aleatoria.

Para optimizar el rendimiento del análisis exploratorio, se extrajo una muestra aleatoria de 1000 registros del conjunto de datos. Se construyó una *pipeline* que preserva la fila del encabezado (header) mientras baraja el cuerpo (body) del archivo.

```
# Comando para crear la muestra aleatoria de 1000 registros
cat data.csv | body shuf | head -n 1001 > sample_earthquakes.csv
```

Finalmente, se verificó el archivo de muestra usando `wc -l`, confirmando la presencia de 1001 líneas (1 encabezado + 1000 registros)

```
# Comando de verificación
wc -l sample_earthquakes.csv
```



2. Limpieza y Transformación

Esta sección detalla el proceso de depuración de la muestra de datos (`sample_earthquakes.csv`), preparándola para el análisis exploratorio y el modelamiento.

2.1 Identificación y Tratamiento de Valores Faltantes

El primer paso de la limpieza fue identificar las columnas que contenían datos nulos (faltantes). Se utilizó `csvstat --nulls`

```
# Identificar columnas con valores nulos
csvstat sample_earthquakes.csv --nulls
```

```
$ csvstat sample_earthquakes.csv --nulls
  1. time: False
  2. latitude: False
  3. longitude: False
  4. depth: False
  5. mag: True
  6. magType: True
  7. nst: True
  8. gap: True
  9. dmin: True
 10. rms: False
 11. net: False
 12. id: False
 13. updated: False
 14. place: False
 15. type: False
 16. horizontalError: True
 17. depthError: False
 18. magError: True
 19. magNst: True
 20. status: False
 21. locationSource: False
 22. magSource: False
```

Una vez identificadas las columnas con datos faltantes (como `mag`, `magType`, `nst`, etc.), se procedió a tratar estos registros. La estrategia utilizada fue eliminar todas las filas (registros) que estuvieran incompletas.

Para esto, se encadenaron múltiples comandos `csvgrep`, usando la opción `-i` (invertir) y la expresión regular `^$` (para campos vacíos) para conservar únicamente las filas completas.

```
# Eliminar filas con campos nulos en las columnas identificadas
cat sample_earthquakes.csv | csvgrep -c mag -i -r "^$" | \
csvgrep -c magType -i -r "^$" | csvgrep -c nst -i -r "^$" | \
csvgrep -c gap -i -r "^$" | csvgrep -c dmin -i -r "^$" | \
csvgrep -c horizontalError -i -r "^$" | csvgrep -c magError -i -r "^$" | \
csvgrep -c magNst -i -r "^$" > sample_no_missing.csv
```

El resultado de esta limpieza fue verificado usando `wc -l`. El archivo original de 1001 líneas se redujo a 753 líneas (752 registros + encabezado).

```
#Verificar la reducción de filas
wc -l sample_earthquakes.csv sample_no_missing.csv
```

```
$ wc -l sample_earthquakes.csv sample_no_missing.csv
  1001 sample_earthquakes.csv
   753 sample_no_missing.csv
  1754 total
```

2.2 Estandarización y Selección de Atributos

Posteriormente, se preparó el archivo para el análisis. Este proceso incluyó dos operaciones:

1. **Estandarización:** Todas las columnas de tipo texto se convirtió a minúscula `tr '[:upper:]' '[:lower:]'`.

```
#Identificar los tipod de datos de las columnas
sample_no_missing.csv --type
```

```
$ csvstat sample_no_missing.csv --type
  1. time: DateTime
  2. latitude: Number
  3. longitude: Number
  4. depth: Number
  5. mag: Number
  6. magType: Text
  7. nst: Number
  8. gap: Number
  9. dmin: Number
 10. rms: Number
 11. net: Text
 12. id: Text
 13. updated: DateTime
 14. place: Text
 15. type: Text
 16. horizontalError: Number
 17. depthError: Number
 18. magError: Number
 19. magNst: Number
 20. status: Text
 21. locationSource: Text
 22. magSource: Text
```

2. **Selección de Atributos:** Se eliminaron columnas irrelevantes para la clasificación, como identificadores (`id`), metadatos (`updated`) y descripciones textuales (`place`, `net`, etc.), usando `csvcut -C`.

Creando así un nuevo archivo llamado `earthquakes_clean.csv`

```
# Estandarizar a minúsculas y eliminar columnas irrelevantes
cat sample_no_missing.csv | tr '[:upper:]' '[:lower:]' | \
csvcut -C id,place,net,locationSource,magSource,updated > earthquakes_clean.csv
```

```
$ cat sample_no_missing.csv | tr '[:upper:]' '[:lower:]' | csvcut -C id,place,net,locationsource,magsource,updated > ear
thquakes_clean.csv

$ ls
columns.txt  data.csv  earthquakes_clean.csv  sample_earthquakes.csv  sample_no_missing.csv
```

Se utilizó `head -10` para visualizar las primeras 10 líneas y verificar así los cambios aplicados.

```
# Visualizar primeras 10 lineas
head -10 earthquakes_clean.csv | csvlook
```

```
$ head -10 earthquakes_clean.csv | csvlook
| time                     | latitude | longitude |    depth |  mag | magtype | nst | gap |   dmin |  rms | type       |
 horizontalerror | depterror | magerror | magnst | status    |
| ------------------------ | -------- | --------- | -------- | ---- | ------- | --- | --- | ------ | ---- | ---------- |
 ---------------- | ---------- | -------- | ------ | --------- |
| 2025-10-16t01:45:10.023z |  31.610… | -103.956… |   7.843… | 1.60 | ml      |  28 |  48 | 0.000… | 0.10 | earthquake |
         0.351… |     0.454… |   0.200… |     19 | reviewed  |
| 2025-09-24t19:57:19.550z |  37.533… | -118.953… |  12.260… | 1.85 | md      |  27 |  89 | 0.084… | 0.06 | earthquake |
         0.370… |     0.780… |   0.120… |     17 | reviewed  |
| 2025-10-05t04:35:52.590z |  38.835… | -122.815… |   1.950… | 1.11 | md      |  16 |  85 | 0.019… | 0.02 | earthquake |
         0.300… |     0.440… |   0.080… |     16 | automatic |
| 2025-10-11t09:05:26.670z |  37.235… | -119.534… |  35.070… | 2.51 | md      |  13 | 103 | 0.108… | 0.06 | earthquake |
         0.440… |     0.990… |   0.080… |      7 | automatic |
| 2025-09-18t21:49:10.817z |  32.125… | -102.196… |   6.421… | 1.30 | ml      |  17 |  65 | 0.000… | 0.20 | earthquake |
         0.891… |     1.613… |   0.200… |     13 | reviewed  |
| 2025-10-01t00:13:07.000z |  38.838… | -122.783… |   1.720… | 0.91 | md      |  10 |  79 | 0.006… | 0.02 | earthquake |
         0.960… |     1.970… |   0.210… |     11 | automatic |
| 2025-09-22t23:16:16.230z | -24.229… |  -67.551… | 190.456… | 4.20 | mb      |  13 |  96 | 1.396… | 0.38 | earthquake |
         9.330… |     7.326… |   0.214… |      6 | reviewed  |
| 2025-10-11t12:58:27.790z |  35.984… | -120.953… |   9.860… | 2.21 | md      |  43 |  52 | 0.066… | 0.12 | earthquake |
         0.270… |     0.780… |   0.230… |     35 | automatic |
| 2025-09-22t05:04:20.150z |  54.794… | -164.174… |  15.780… | 1.53 | ml      |   4 | 311 | 0.035… | 0.32 | earthquake |
         2.820… |     2.410… |   0.139… |      4 | reviewed  |
```

2.3 Generación de Estadísticas Descriptivas

Finalmente, se generaron estadísticas descriptivas para la variable `mag` (magnitud), agrupadas por `magtype`. Para esta tarea, se utilizó la herramienta `rush` para ejecutar un *one-liner* de R con funciones de `tidyverse`.

```
# Calcular estadísticas agrupadas usando rush y R
rush run --tidyverse "
    group_by(df, magtype) %>%
    summarise(
        media_mag = mean(mag),
        mediana_mag = median(mag),
        min_mag = min(mag),
        max_mag = max(mag),
        rango_mag = max_mag - min_mag,
        desv_est_mag = sd(mag),
        varianza_mag = var(mag)
    )" earthquakes_clean.csv | csvlook
```

Table 1: Estadísticas Descriptivas de Magnitud (mag) por Tipo (magType)

magtype	media_mag	mediana_mag	min_mag	max_mag	rango_mag	desv_est_mag	varianza_mag
mb	4.663	4.60	4.00	5.60	1.600	0.352	0.124
nd	1.289	1.09	-0.06	3.88	3.940	0.730	0.533
ml	1.113	1.14	-1.24	3.86	5.100	0.783	0.612
mwr	4.500	4.50	4.50	4.50	0.00		
mww	5.422	5.20	5.00	6.70	1.700	0.531	0.282

3. Análisis y Visualización

En esta sección, se realiza un análisis exploratorio de los datos. El análisis se centra en las distribuciones de magnitud, profundidad y ubicación geográfica, cruzadas por las variables categóricas tipo de magnitud, tipo de terremoto y status.

3.1 Preparación del Archivo de Visualización

Para facilitar el análisis temporal, se creó un archivo `earthquakes_anio.csv`. Este archivo se generó a partir de `earthquakes_clean.csv`, añadiendo una nueva columna `year` extrayendo los primeros cuatro caracteres de la columna `time`. Se usó `rush run` con la función `substr` para esta transformación.

```
# Crear archivo de visualización con la columna "year"
rush run --tidyverse "mutate(df, year = as.numeric(substr(time, 1, 4)))" earthquakes_clean.csv > earthq
```

Nota sobre el Análisis Temporal (por Año):

Posteriormente, se realizó una exploración preliminar de esta nueva variable `year` (ver Gráfico de Anexo 1, `scatter_mag_por_anio.png`). Esta exploración reveló que **todos los registros en la muestra de datos (`earthquakes_anio.csv`) pertenecen al mismo año (2025)**.

Dado que no existe variación en la variable temporal (es un valor constante), no es posible realizar un análisis de tendencia o una comparación interanual. Por lo tanto, el análisis cruzado "por año" fue omitido de las siguientes subsecciones por carecer de valor estadístico.

3.2 Análisis de Magnitud (mag)

Se exploró la variable `mag` para entender su distribución general y su relación con `magtype`, `type` y `status`.

Histograma General de Magnitud

Se generó un histograma para observar la distribución de todas las magnitudes en la muestra.

```
# Comando para generar el histograma de magnitud
rush plot --x mag --geom histogram --title "Distribucion de Magnitudes Sismicas" earthquakes_anio.csv >
```

Gráfico 1: Distribución de Magnitudes Sísmicas

Interpretación: Este histograma muestra la frecuencia (Eje Y) de los sismos según su magnitud (Eje X). La gran mayoría de los eventos sísmicos en la muestra son de baja magnitud.

- Se observa una alta concentración de sismos con magnitudes entre 0 y 2.5.

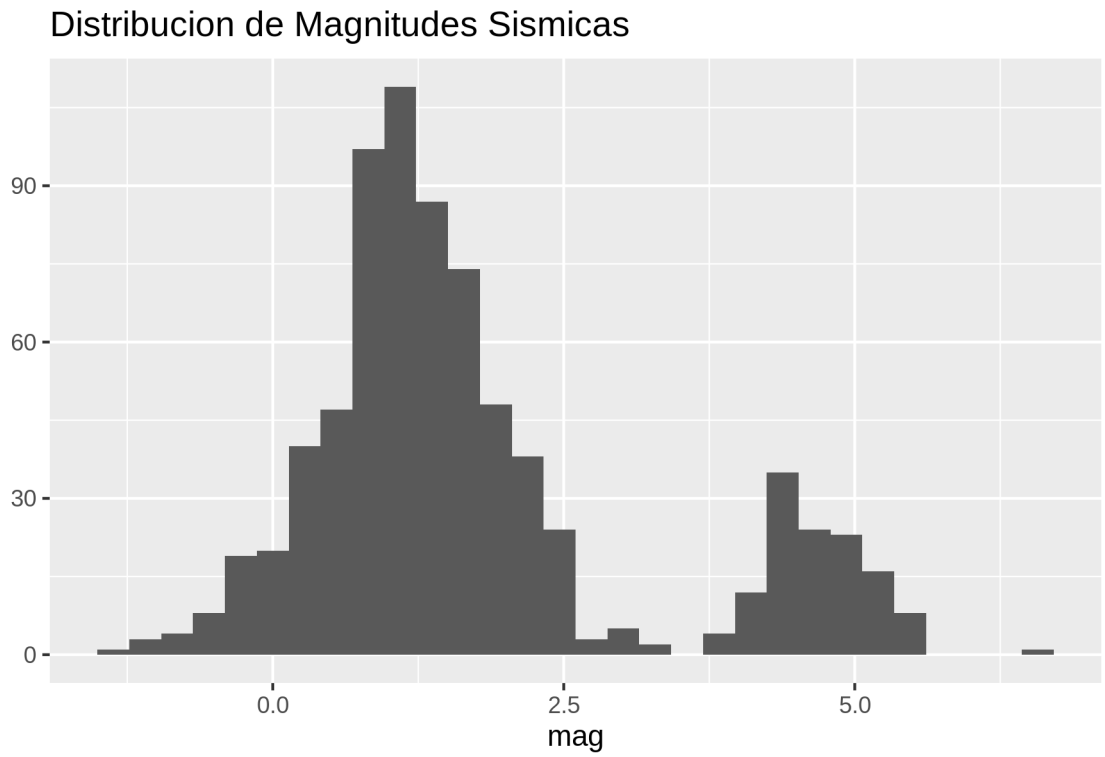


Figure 1: Histograma de Magnitud

- La distribución está sesgada hacia la derecha, lo que significa que a medida que la magnitud aumenta, la cantidad de sismos disminuye.
- Los eventos con magnitudes superiores a 5 son muy poco frecuentes en este conjunto de datos.

Magnitud por Tipo de Magnitud (magType) Se utilizó un gráfico de densidad para comparar las distribuciones de `mag` entre los diferentes `magtype`.

```
# Comando para generar el gráfico de densidad
rush plot --x mag --color magtype --geom density --title "Distribucion de Magnitud por Tipo (magType)" 
```

Gráfico 2: Distribución de Magnitud por Tipo (magType)

Interpretación: Este gráfico compara la distribución de la magnitud (Eje X) para cada categoría de la columna `magtype`. El tipo de magnitud parece estar fuertemente correlacionado con el rango de magnitud medido:

- `ml` y `md`: Estos tipos de magnitud están asociados a sismos de baja magnitud. Ambas distribuciones tienen mayor densidad en magnitudes inferiores a 2.5.
- `mb` y `mww`: Estos tipos de magnitud están asociados a los sismos de mayor intensidad, con sus distribuciones centradas en valores superiores a 3.8 aproximadamente.

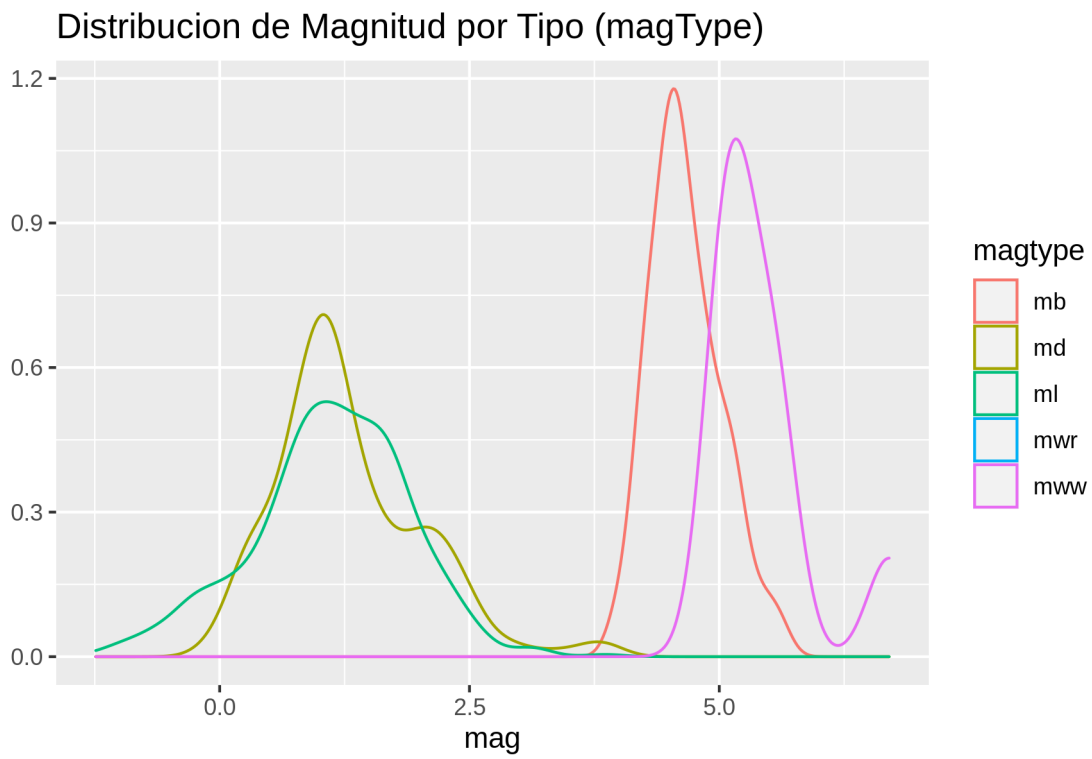


Figure 2: Densidad de Magnitud por magType

Magnitud por Tipo de Evento (type)

Se generó un boxplot para comparar la magnitud de los `earthquake` frente a otros eventos.

```
# Comando para generar el boxplot
rush plot --x type --y mag --geom boxplot --title "Magnitud por Tipo de Evento" earthquakes_anio.csv > l
```

Gráfico 3: Magnitud por Tipo de Evento

Interpretación: El tipo de evento está diferenciado por su magnitud.

- earthquake (terremoto): Presenta la mayor variabilidad en magnitud. La caja se sitúa en magnitudes bajas, pero presenta una gran cantidad de valores atípicos. Esto indica que, si bien la mayoría de los terremotos son leves, este tipo de evento es el único responsable de los sismos de alta magnitud en la muestra
- explosion (explosión): Magnitudes moderadas. El rango intercurtílico es acotado (poca dispersión), con pocos atípicos
- quarry blasy (explosión de cantera): Magnitudes bajas. La mediana está cerca de 1.3 y la caja es estrecha (variabilidad mínima), practicamente sin valores extremos. Indica detonaciones controladas.

Los eventos no tectónicos (explosiones y canteras) se concentran en magnitudes bajas a moderadas con poca dispersión, mientras que los terremotos son los únicos que alcanzan magnitudes altas y muestran gran variabilidad.

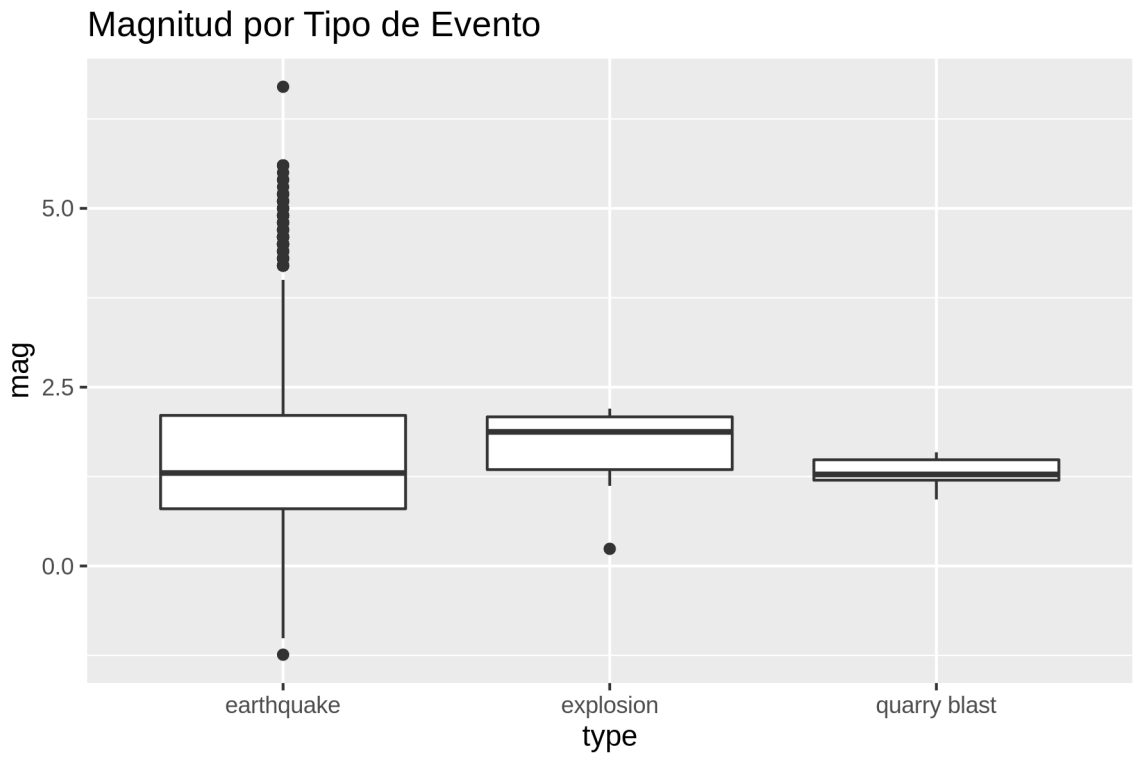


Figure 3: Boxplot de Magnitud por Tipo de Evento.

Magnitud por Estado (status)

Finalmente, se usó un boxplot para verificar si el estado se relaciona con la magnitud.

```
# Comando para generar el boxplot
rush plot --x status --y mag --geom boxplot --title "Magnitud por Estado del Evento" earthquakes_anio.cs
```

Gráfico 4: Magnitud por Estado del Evento

Interpretación: El estado del registro se relaciona fuertemente con la magnitud.

- automatic (automático): Presenta una caja estrecha con pocos valores atípicos. Indica que el procesamiento automático cubre mayormente eventos pequeños y rutinarios, con magnitudes consistentes.
- reviewed (revisado): Mucha mayor dispersión, hay numerosos atípicos que llegan sobre 5. Indica que los registros revisados incluyen tanto eventos prqueños como eventos moderados a grandes.

Los eventos de gran magnitud son casi siempre revisados. El flujo de calidad parece priorizar la revisión manual de eventos más significativos, mientras que los pequeños quedan validados por el sistema automático.

3.3 Análisis de Profundidad (depth)

De manera similar a la magnitud, se exploró la variable `depth` (profundidad).

Histograma General de Profundidad

Se generó un histograma para observar la distribución general de las profundidades.

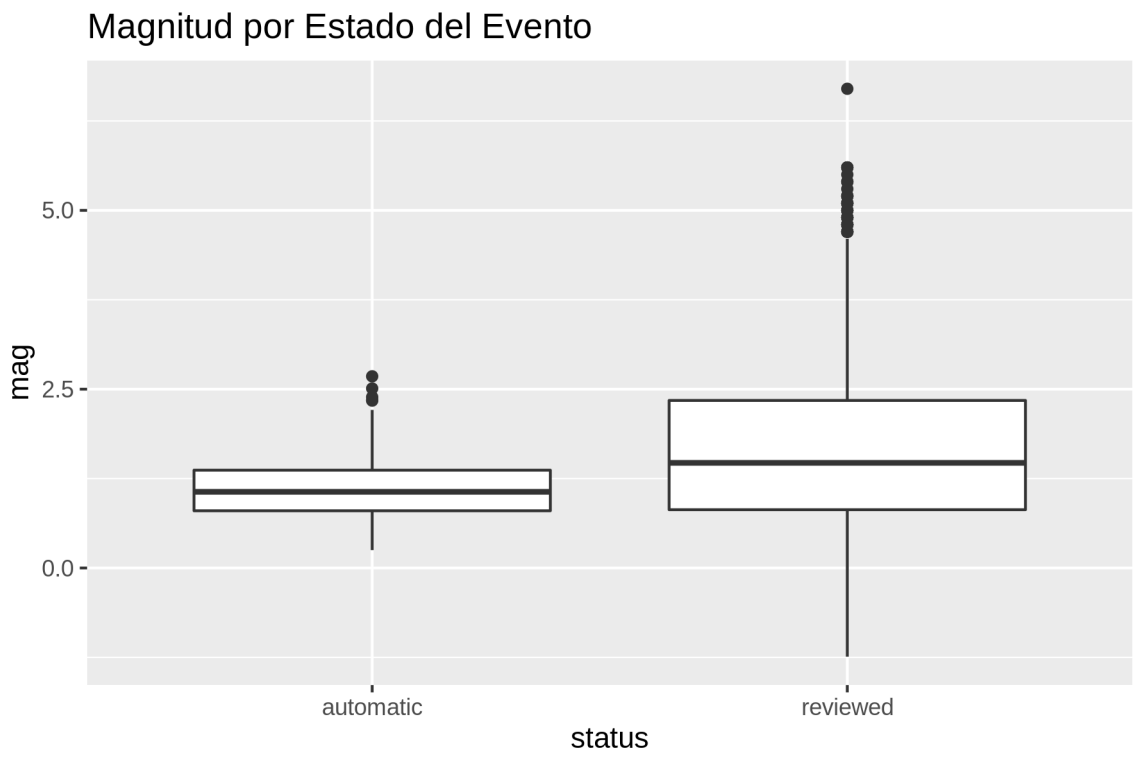


Figure 4: Boxplot de Magnitud por Status.

```
# Comando para generar el histograma de profundidad
rush plot --x depth --geom histogram --title "Distribucion de Profundidades Sismicas" earthquakes_anio.
```

Gráfico 5: Distribución de Profundidades Sísmicas

Interpretación: Muestra la frecuencia (Eje Y) de los sismos según su profundidad (`depth`) en kilómetros (Eje X). La mayoría de los eventos sísmicos en la muestra son superficiales.

- La distribución está fuertemente sesgada a la derecha.
- Se observa una alta frecuencia para sismos con profundidades muy bajas, la mayoría ocurriendo a menos de 50 km.

Profundidad por Tipo de Magnitud (magType)

Se utilizó un gráfico de densidad para comparar las distribuciones de `depth` entre los diferentes `magtype`.

```
# Comando para generar el gráfico de densidad
rush plot --x depth --color magtype --geom density --title "Distribucion de Profundidad por Tipo (magTy
```

Gráfico 6: Distribución de Profundidad por Tipo (magType)

Interpretación: Todas las curvas se concentran en profundidades pequeñas y luego caen rápido; a mayor profundidad, la densidad es baja. Las líneas muestran densidades, así que comparan la forma de la distribución por tipo de magnitud.

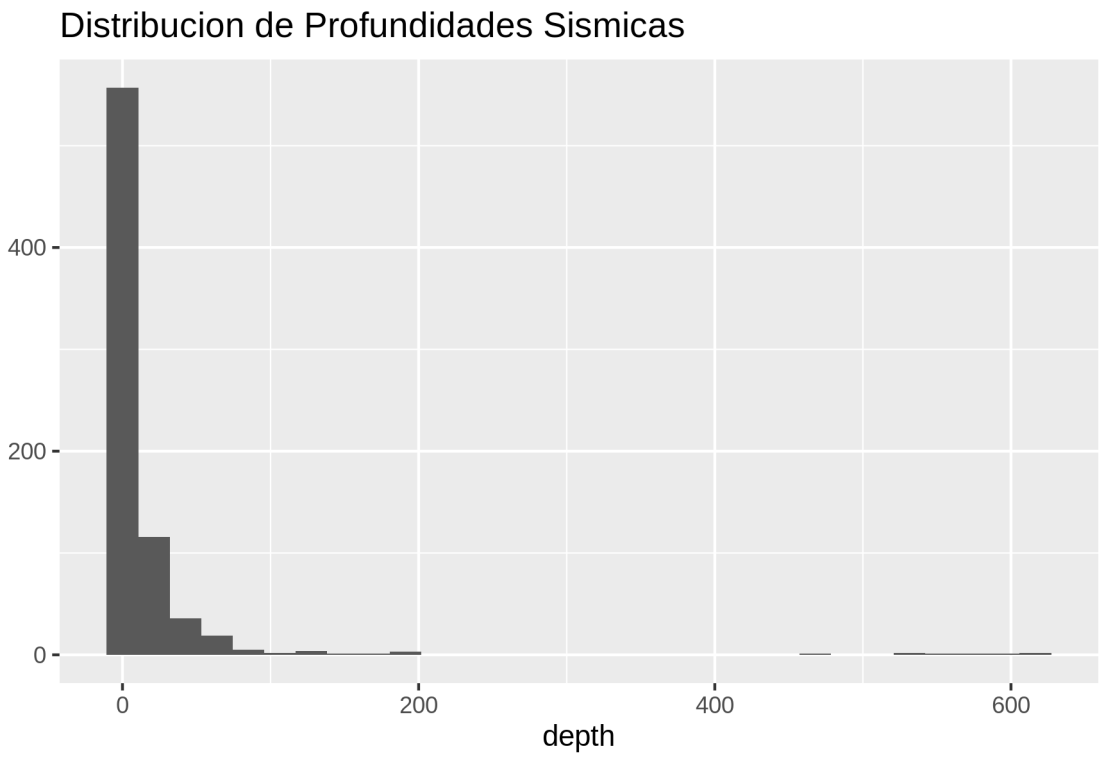


Figure 5: Histograma de Profundidad.

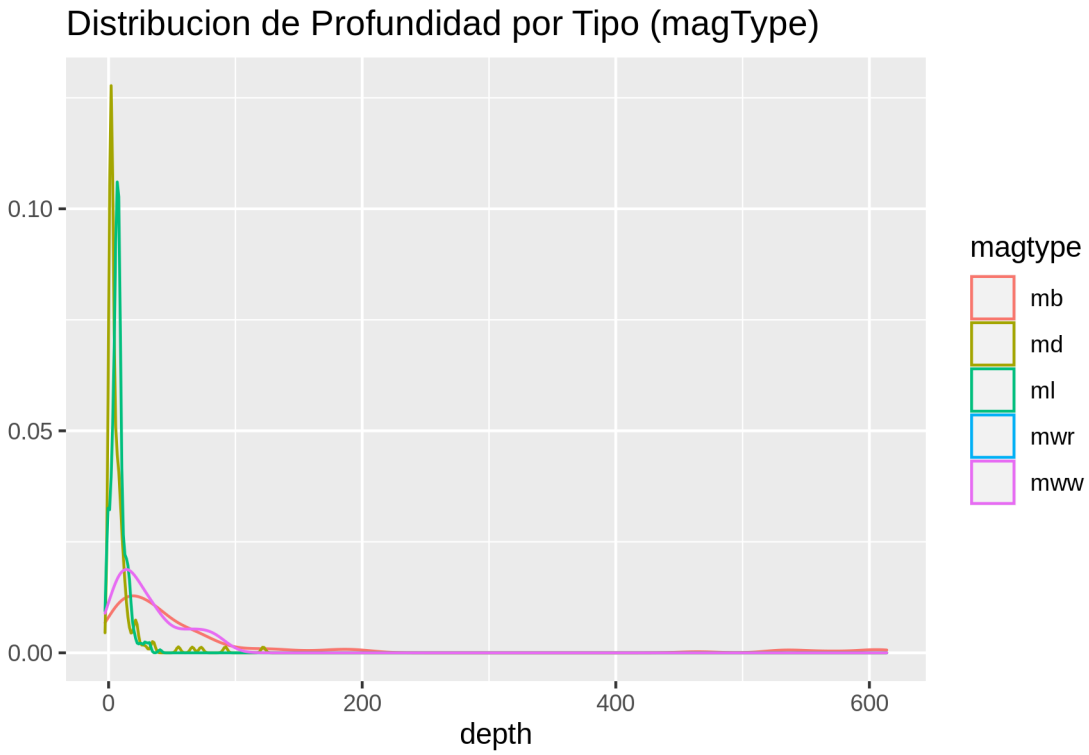


Figure 6: Densidad de Profundidad por magType.

Profundidad por Tipo de Evento (type)

Se usó un boxplot para comparar la profundidad entre los tipos de evento.

```
# Comando para generar el boxplot
rush plot --x type --y depth --geom boxplot --title "Profundidad por Tipo de Evento" earthquakes_anio.cs
```

Gráfico 7: Profundidad por Tipo de Evento

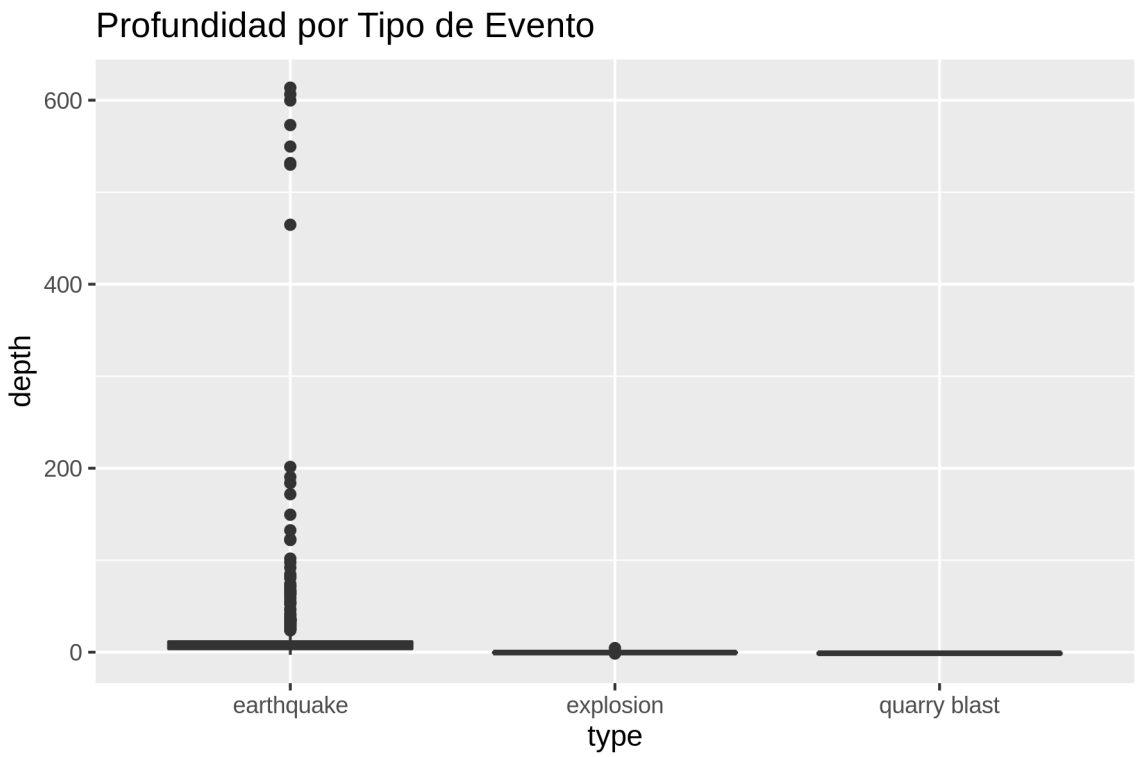


Figure 7: Boxplot de Profundidad por Tipo de Evento.

Interpretación: Los eventos que no son terremotos son exclusivamente superficiales.

- earthquake (terremoto): Aparecen en todo el rango de profundidad. Hay muchos muy cerca de la superficie, pero también hay intermedios y muy profundos.
- explosion (explosión): Están casi en la superficie. Los puntos se quedan pegados al 0 km.
- quarry blast (explosión de cantera):Igual que las explosiones: muy superficiales, prácticamente en 0 km.

Profundidad por Estado (status)

Se usó un boxplot para comparar la profundidad con el estado de revisión.

```
# Comando para generar el boxplot
rush plot --x status --y depth --geom boxplot --title "Profundidad por Estado del Evento" earthquakes_a
```

Gráfico 8: Profundidad por Estado del Evento

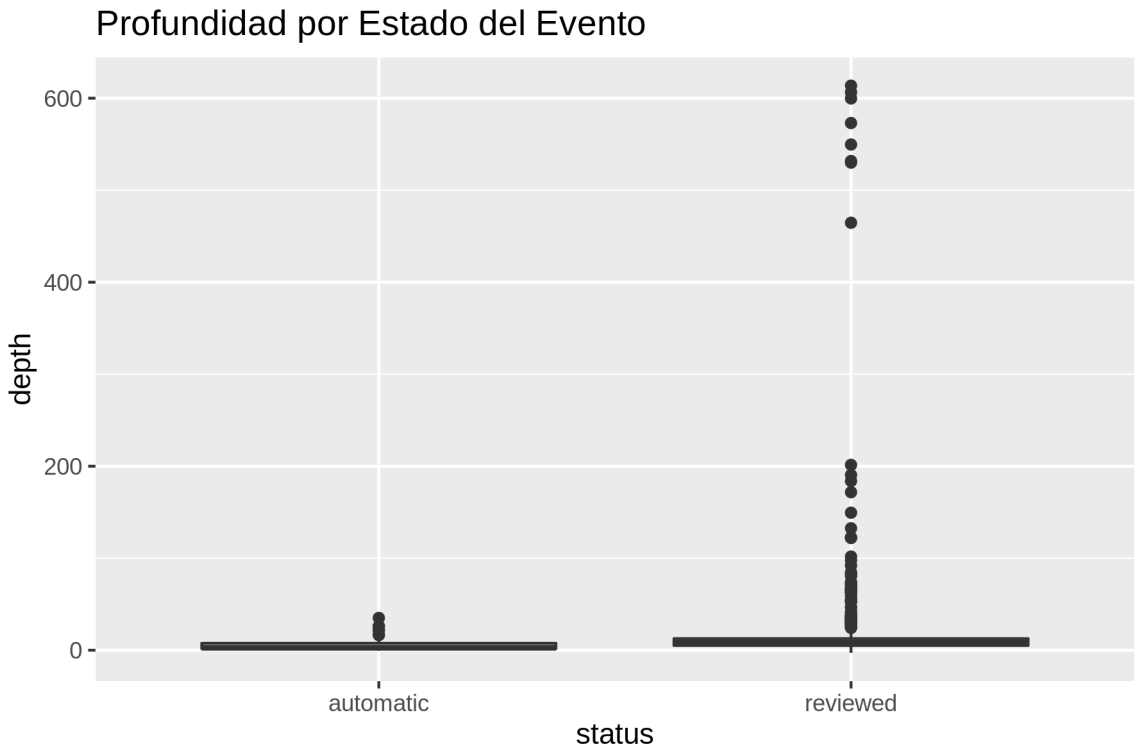


Figure 8: Boxplot de Profundidad por Status.

Interpretación: A medida que la profundidad aumenta, crece la probabilidad de revisión: los eventos profundos rara vez se quedan solo en automático.

- automatic (automático): Profundidades muy concenctradas, con algunos casos aislados hasta pocas decenas de km.
- reviewed (revisado): Distribución amplia, muchos superficiales pero también una cola larga de outliers que llega a cientos de km.

3.4 Análisis de Regiones Geográficas

Finalmente, se analizó la distribución geográfica (`longitude` y `latitude`) de los sismos mediante diagramas de dispersión.

Mapa de Sismos por Magnitud

```
# Comando para generar el mapa de dispersión coloreado por magnitud
rush plot --x longitude --y latitude --color mag --geom point --title "Mapa de Sismos por Magnitud" ear
```

Gráfico 9: Mapa de Sismos por Magnitud

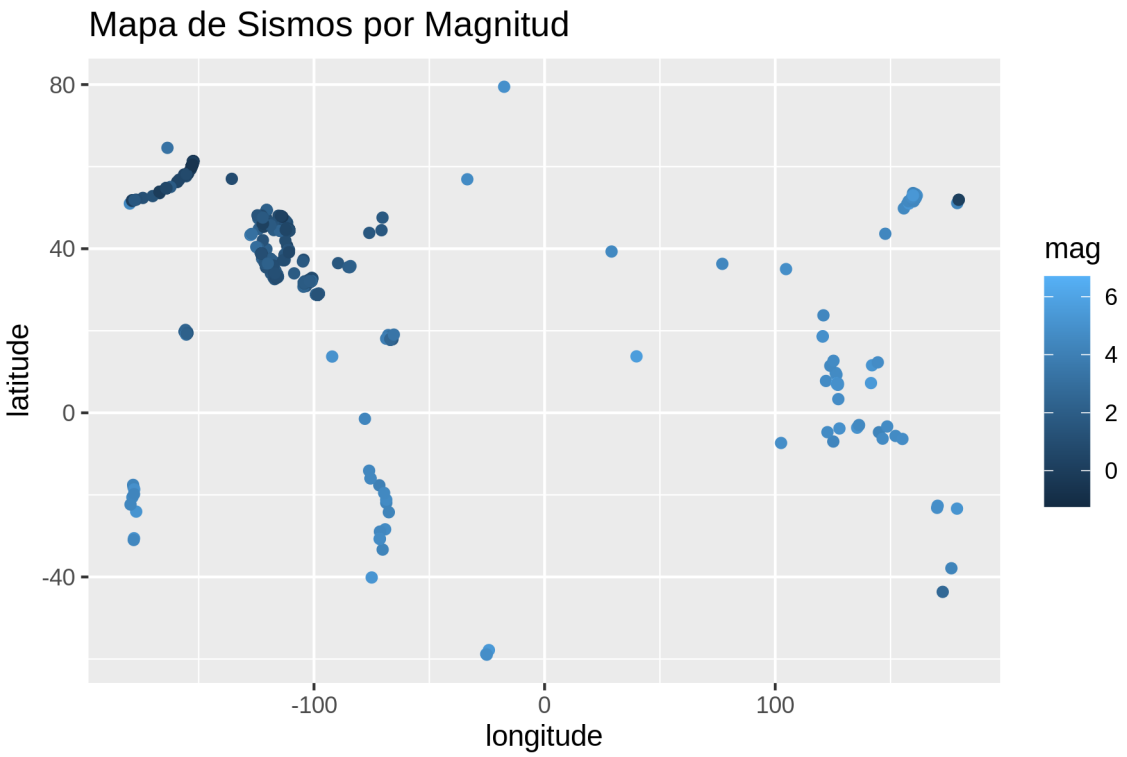


Figure 9: Mapa de Sismos por Magnitud.

Interpretación: En este gráfico el eje X representa la longitud y el eje Y la latidud. Cada punto es un sismo, y su color indica la magnitud. La actividad sísmica se concentra en zonas geográficas muy específicas.

Se observan cúmulos claros de sismos, particularmente en la parte izquierda del gráfico. También hay una agrupación notable en latitudes más altas.

El color de los puntos (donde los tonos más claros indican mayor magnitud) se mezcla con los puntos oscuros (baja magnitud). Sin embargo, se puede apreciar que los eventos de mayor magnitud también ocurren dentro de estas zonas de alta actividad, sugiriendo que las mismas fallas geográficas responsables del gran volumen de sismos menores también son responsables de los sismos mayores.

Mapa de Sismos por Tipo de Evento

```
# Comando para generar el mapa de dispersión coloreado por tipo
rush plot --x longitude --y latitude --color type --geom point --title "Mapa de Sismos por Tipo de Even
```

Gráfico 10: Mapa de Sismos por Tipo de Evento

Interpretación: En este gráfico se puede observar cómo se distribuyen los tipos de eventos sísmicos.

- earthquake (terremoto): Ocurren en cúmulos geográficos claros, consistentes con fallas tectónicas como la costa oeste de EE.UU.
- explosion (explosión): Forman un clúster muy localizado en el interior del país. Puede ser debido a actividad humana (operaciones industriales/voladuras controladas o un sitio de pruebas).
- quarry blast (explosión de cantera): Prácticamente no se ven.

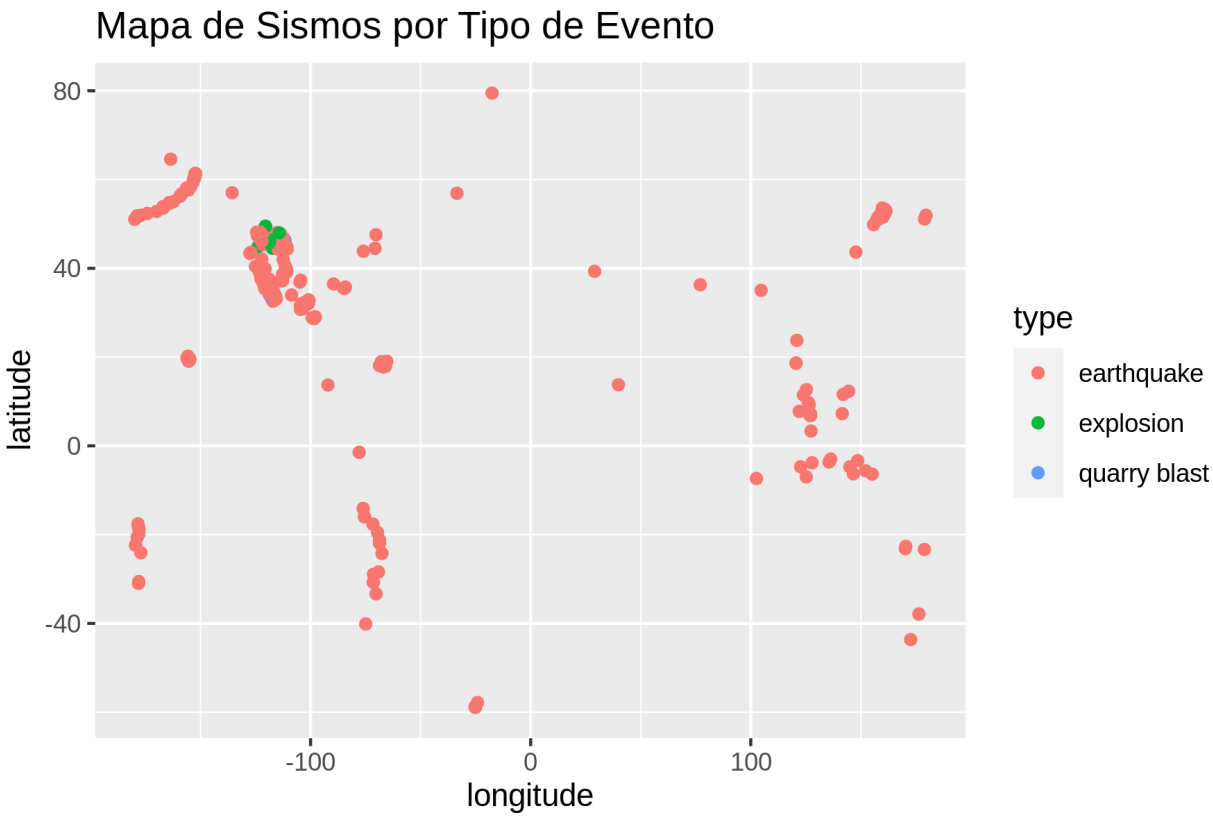


Figure 10: Mapa de Sismos por Tipo de Evento.

Mapa de Sismos por Estado

```
# Comando para generar el mapa de dispersión coloreado por status
rush plot --x longitude --y latitude --color status --geom point --title "Mapa de Sismos por Estado" ea
```

Gráfico 11: Mapa de Sismos por Estado

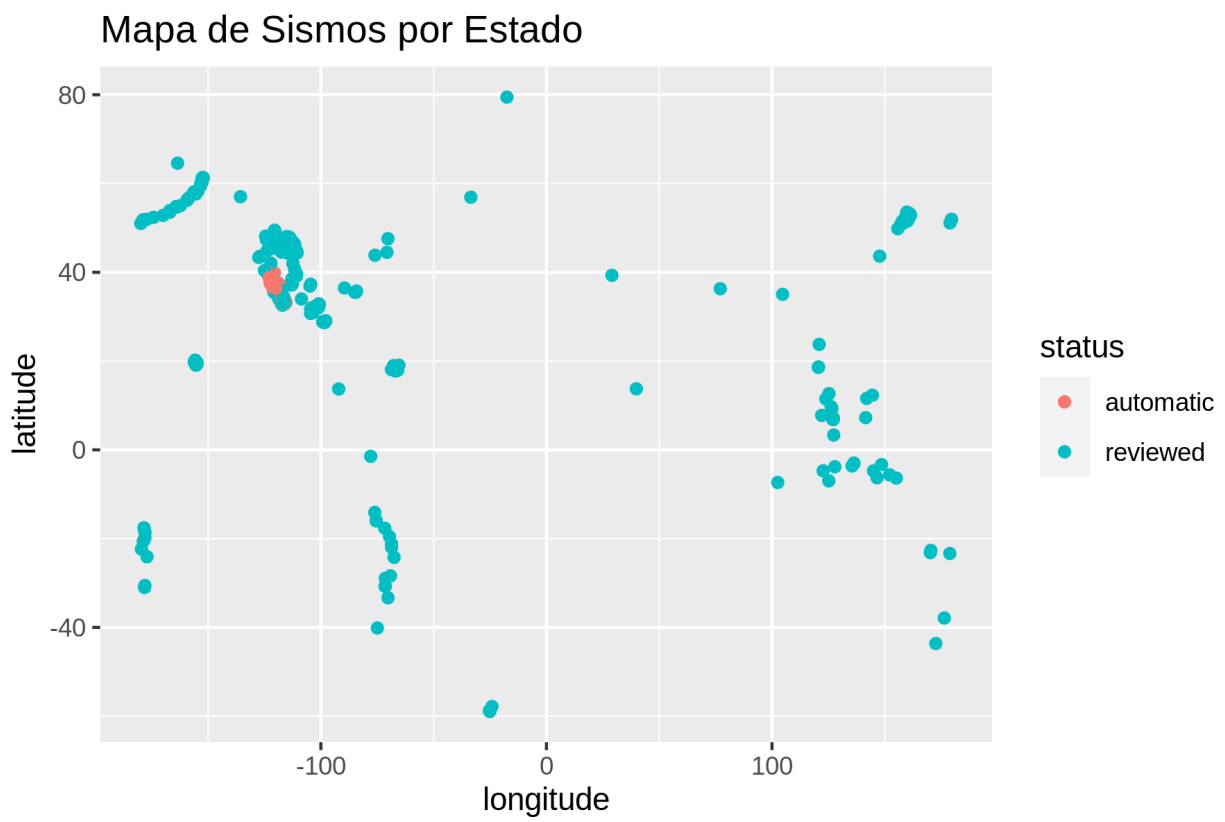


Figure 11: Mapa de Sismos por Status.

Interpretación:

- automatic (automático): Aparecen como un clúster muy concentrado en el interior del oeste, indicando detecciones locales. Prácticamente no hay automáticos fuera de ese foco.
- reviewed (revisado): Son la gran mayoría y están distribuidos por todas las zonas sísmicas. Sueles ser los eventos más significativos.

4. Modelamiento

La cuarta fase consistió en construir modelos de clasificación para predecir la variable `type` (tipo de evento).

4.1 Metodología de Modelamiento

Se implementó un enfoque donde `vw` se utiliza en su modo de regresión para predecir las etiquetas de clase, las cuales fueron convertidas manualmente de texto a números.

4.2 Preparación de Datos para VW

[cite_start]El primer paso fue crear un archivo CSV base para el modelamiento (`earthquakes_model.csv`), conservando la columna `id` y eliminando otras columnas irrelevantes.

```
# Crear el archivo CSV base para modelamiento
cat sample_no_missing.csv | tr '[:upper:]' '[:lower:]' | \
csvcut -C place,net,locationSource,magSource,updated > earthquakes_model.csv
```

A continuación, se identificaron las clases únicas en la columna `type`.

```
# Identificar las clases únicas en 'type'
csvcut -c type earthquakes_model.csv | header -d | sort | uniq -c
```

```
$ csvcut -c type earthquakes_model.csv | header -d | sort | uniq -c
    731 earthquake
     10 explosion
     11 quarry blast
```

Dado que `vw` (en modo regresión) requiere etiquetas numéricas, las 3 clases de texto se convirtieron a enteros (1, 2, 3) usando `sed`.

```
# Convertir etiquetas de texto a números
cat earthquakes_model.csv | sed 's/earthquake/1/' | \
sed 's/explosion/2/' | sed 's/quarry blast/3/' > earthquakes_numeric.csv
```

Este archivo numérico se convirtió al formato `.vw` usando `csv2vw`.

```
# Convertir el CSV numérico al formato .vw
csv2vw earthquakes_numeric.csv --label type > earthquakes.vw
```

Finalmente, los datos se barajaron (`shuf`) y se dividieron (usando `split`) en un conjunto de entrenamiento y uno de prueba.

```
# Barajar (shuf) el archivo .vw
shuf earthquakes.vw > earthquakes.shuffled.vw

# Dividir (split) en 5 partes (20% c/u)
split -d -n r/5 earthquakes.shuffled.vw vw-part-

#Crear el set de prueba (20%)
mv vw-part-00 vw-test.vw

# Crear el set de entrenamiento (80%)
cat vw-part-01 vw-part-02 vw-part-03 vw-part-04 > vw-train.vw

# Limpiar los archivos temporales
rm vw-part-* earthquakes.shuffled.vw
```

4.3 Ejecución y Evaluación de 5 Modelos

Se entrenaron 5 modelos variando los hiperparámetros de `vw`. La evaluación se realizó calculando la **Precisión (Accuracy)** en el set de prueba, para lo cual las predicciones de regresión (decimales) fueron redondeadas al entero más cercano usando `awk`.

El siguiente script (`03_model.sh`) automatiza el entrenamiento, prueba y evaluación de los 5 modelos, guardando los resultados en `model_accuracy.txt`.

(Nota: Solo mostraré los comandos de evaluación aquí por brevedad, el script completo se adjunta en los entregables)

```
# --- Modelo 1: Simple ---
vw -d vw-train.vw -f modelo1.vw --quiet
vw -d vw-test.vw -i modelo1.vw -t -p predictions1.txt --quiet
paste predictions1.txt <(cut -d ' ' -f 1 vw-test.vw) | \
awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo 1 (Simple) - Precision: " (correct/NR)*

# --- Modelo 2: Passes=10 ---
vw -d vw-train.vw -f modelo2.vw --passes 10 --cache_file modelo.cache --quiet
vw -d vw-test.vw -i modelo2.vw -t -p predictions2.txt --quiet
paste predictions2.txt <(cut -d ' ' -f 1 vw-test.vw) | \
awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo 2 (Passes=10) - Precision: " (correct/NR

# --- Modelo 3: Quadratic ---
vw -d vw-train.vw -f modelo3.vw --passes 10 --cache_file modelo.cache --quadratic :: --quiet
vw -d vw-test.vw -i modelo3.vw -t -p predictions3.txt --quiet
paste predictions3.txt <(cut -d ' ' -f 1 vw-test.vw) | \
awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo 3 (Passes=10, Quadratic) - Precision: "

# --- Modelo 4: Red Neuronal ---
vw -d vw-train.vw -f modelo4.vw --passes 10 --cache_file modelo.cache --nn 3 --quiet
vw -d vw-test.vw -i modelo4.vw -t -p predictions4.txt --quiet
paste predictions4.txt <(cut -d ' ' -f 1 vw-test.vw) | \
awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo 4 (Passes=10, NN=3) - Precision: " (cor

# --- Modelo 5: L2 ---
vw -d vw-train.vw -f modelo5.vw --passes 10 --cache_file modelo.cache --l2 0.000005 --quiet
vw -d vw-test.vw -i modelo5.vw -t -p predictions5.txt --quiet
paste predictions5.txt <(cut -d ' ' -f 1 vw-test.vw) | \
awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo 5 (Passes=10, L2) - Precision: " (corre
```

4.4 Resultados y Conclusiones del Modelamiento

La siguiente tabla resume las métricas de evaluación (Precisión) obtenidas por los 5 modelos en el conjunto de prueba.

Tabla 2: Tabla de Métricas de Evaluación de Modelos VW

Table 2: Resultados de Precisión (Accuracy) de los 5 modelos VW entrenados.

Modelo	Parámetros	Precisión (Accuracy)
Modelo 1	Simple (por defecto)	98.0132%

Modelo	Parámetros	Precisión (Accuracy)
Modelo 2	(Passes=10)	98.6755%
Modelo3	(Passes=10, Quadratic)	90.7285%
Modelo 4	(Passes=10, NN=3)	96.6887%
Modelo5	(Passes=10, L2)	98.6755%

Interpretación: Los Modelos 2 y 5 fueron los mejores, ambos con una precisión idéntica del 98.6755%.

- Modelo Base (Modelo 1): El modelo lineal simple de `vw` ya era extremadamente bueno (98.01%) para separar estos datos.
- Mejora (Modelo 2): Añadir `--passes 10` (ver los datos 10 veces) le dio al modelo la oportunidad de ajustarse mejor, resultando en una ligera (pero positiva) mejora.
- Empeoramiento (Modelo 3): Añadir interacciones cuadráticas (`--quadratic ::`) empeoró drásticamente el rendimiento (90.72%). Esto sugiere que el modelo se "sobreajustó", aprendiendo ruido en lugar de patrones útiles.
- Alternativa (Modelo 4): La red neuronal (`--nn 3`) fue menos precisa (96.68%) que el modelo lineal simple. Para este problema, un modelo lineal parece ser superior.
- Empate (Modelo 5): Añadir regularización L2 (`--l2`) (una técnica para prevenir el sobreajuste) no cambió el resultado del Modelo 2. Esto indica que nuestro Modelo 2 (con 10 pasadas) ya era muy robusto y no estaba sobreajustándose.

El Modelo 2 (con `--passes 10`) es el mejor modelo en términos de simplicidad y rendimiento.

Anexos

Anexo 1: Salida de Terminal (Tabla 1 - Estadísticas Descriptivas)

Esta captura respalda la **Tabla 1** (Sección 2.3), mostrando la salida directa del comando `rush ... | csvlook` ejecutado en la terminal.

```
$ rush run --tidyverse "
    group_by(df, magtype) %>%
    summarise(
        media_mag = mean(mag),
        mediana_mag = median(mag),
        min_mag = min(mag),
        max_mag = max(mag),
        rango_mag = max_mag - min_mag,
        desv_est_mag = sd(mag),
        varianza_mag = var(mag)
    )" earthquakes_clean.csv | csvlook

| magtype | media_mag | mediana_mag | min_mag | max_mag | rango_mag | desv_est_mag | varianza_mag |
| ------- | --------- | ----------- | ------- | ------- | --------- | ------------ | ------------ |
| mb      |    4.663… |        4.60 |    4.00 |    5.60 |    1.600… |       0.352… |       0.124… |
| md      |    1.289… |        1.09 |   -0.06 |    3.88 |    3.940… |       0.730… |       0.533… |
| ml      |    1.113… |        1.14 |   -1.24 |    3.86 |    5.100… |       0.783… |       0.612… |
| mwr     |    4.500… |        4.50 |    4.50 |    4.50 |    0.000… |              |              |
| mww     |    5.422… |        5.20 |    5.00 |    6.70 |    1.700… |       0.531… |       0.282… |
```

Figure 12: Salida de terminal para la Tabla 1.

Anexo 2: Análisis Temporal Descartado

Como se mencionó en la Sección 3.1, se exploró la variable `year`, pero se descubrió que todos los datos pertenecían a 2025. El siguiente gráfico de dispersión muestra todos los puntos de datos apilados en una sola línea vertical, confirmando la falta de variación temporal y justificando la omisión de un análisis de tendencia.

```
# Comando que generó el gráfico de dispersión por año
rush plot --x year --y mag --geom point --title "Magnitud (mag) por Ano" earthquakes_anio.csv > scatter
```

Gráfico 12: Verificación de Distribución por Año

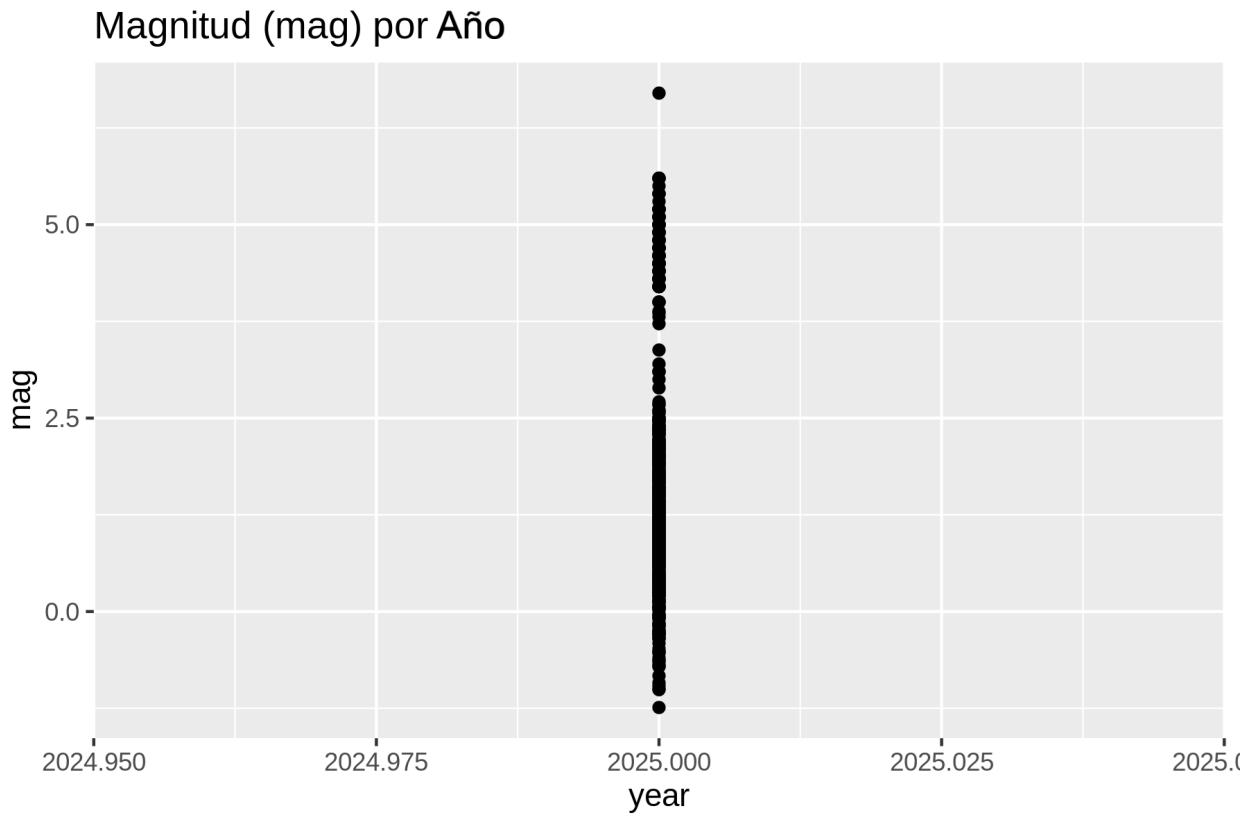


Figure 13: Todos los puntos de datos se agrupan en el año 2025.

Anexo 3: Verificación de Precisión de Modelos VW

Este anexo proporciona las salidas de terminal capturadas durante la evaluación de los 5 modelos de Vowpal Wabbit.

```
paste predictions1.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END
```

Modelo 1 (Simple) Salida de Terminal (Modelo 1):

```
$ paste predictions1.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo
 1 (Simple) - Precision: " (correct/NR)*100 "%" }'
Modelo 1 (Simple) - Precision: 98.0132%
```

```
paste predictions2.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END
```

Modelo 2 (Passes=10) Salida de Terminal (Modelo 2):

```
$ paste predictions2.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo
 2 (Passes=10) - Precision: " (correct/NR)*100 "%" }'
Modelo 2 (Passes=10) - Precision: 98.6755%
```

```
paste predictions3.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END
```

Modelo 3 (Passes=10, Quadratic) Salida de Terminal (Modelo 3):

```
$ paste predictions3.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo
 3 (Passes=10, Quadratic) - Precision: " (correct/NR)*100 "%" }'
Modelo 3 (Passes=10, Quadratic) - Precision: 90.7285%
```

```
paste predictions4.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END
```

Modelo 4 (Passes=10, NN=3) Salida de Terminal (Modelo 4):

```
$ paste predictions4.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo
 4 (Passes=10, NN=3) - Precision: " (correct/NR)*100 "%" }'
Modelo 4 (Passes=10, NN=3) - Precision: 96.6887%
```

```
paste predictions5.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END
```

Modelo 5 (Passes=10, L2) Salida de Terminal (Modelo 5):

```
$ paste predictions5.txt <(cut -d ' ' -f 1 vw-test.vw) | awk '{ if (int($1 + 0.5) == $2) correct++ } END { print "Modelo
 5 (Passes=10, L2) - Precision: " (correct/NR)*100 "%" }'
Modelo 5 (Passes=10, L2) - Precision: 98.6755%
```