

# Analisi dei Terremoti e delle faglie nel mondo (1970-2025)

## Introduzione

In questo lavoro di gruppo, abbiamo deciso di analizzare un dataset preso da USGS.gov, che contiene varie informazioni sui terremoti avvenuti in tutti il mondo dal 1970 al 2025. Il dataset è stato selezionato in modo da avere solo terremoti di magnitudo superiore a 5.0 della scala Richter. Ci siamo limitati a questa magnitudo in quanto il sito consente di scaricare fino a 20.000 eventi alla volta, quindi per coprire tutta la serie storica è stato necessario unificare tutte le richieste creando un unico dataset con 87798 osservazioni. Inoltre, dal sito globalquakemodel.org abbiamo preso la mappa di tutte le faglie attive nel mondo. In questo modo abbiamo potuto confrontare le posizioni dei terremoti rispetto alle faglie. I dati sono stati analizzati tenendo conto sia della loro posizione geografica e di altre caratteristiche presenti nel dataset oltre che della loro frequenza nel corso del tempo. Queste analisi sono state effettuate sia attraverso l'utilizzo di mappe, sia attraverso l'utilizzo di grafici al fine di descrivere al meglio tutte le variabili quantitative e qualitative presenti.

## Librerie

Per semplicità si è deciso di presentare tutte le librerie utilizzate nell'analisi e nell'elaborazione dei dati contenute nel seguente blocco di codice.

```
library(readr)
library(sf)
library(ggplot2)
library(dplyr)
library(maps)
library(tidyr)
library(RColorBrewer)
library(stringr)
library(rnaturalearth)
library(rnaturalearthdata)
library(ggrepel)
library(lubridate)
library(gridExtra)
library(tidyr)
library(corrplot)
library(ggcorrplot)
library(GGally)
library(pander)
```

## Dataset Faglie

Il dataset contiene la mappa di tutte le faglie attive presenti nel mondo, ed è possibile importarlo tramite il seguente codice:

```
faglie <- st_read("gem_active_faults.shp")

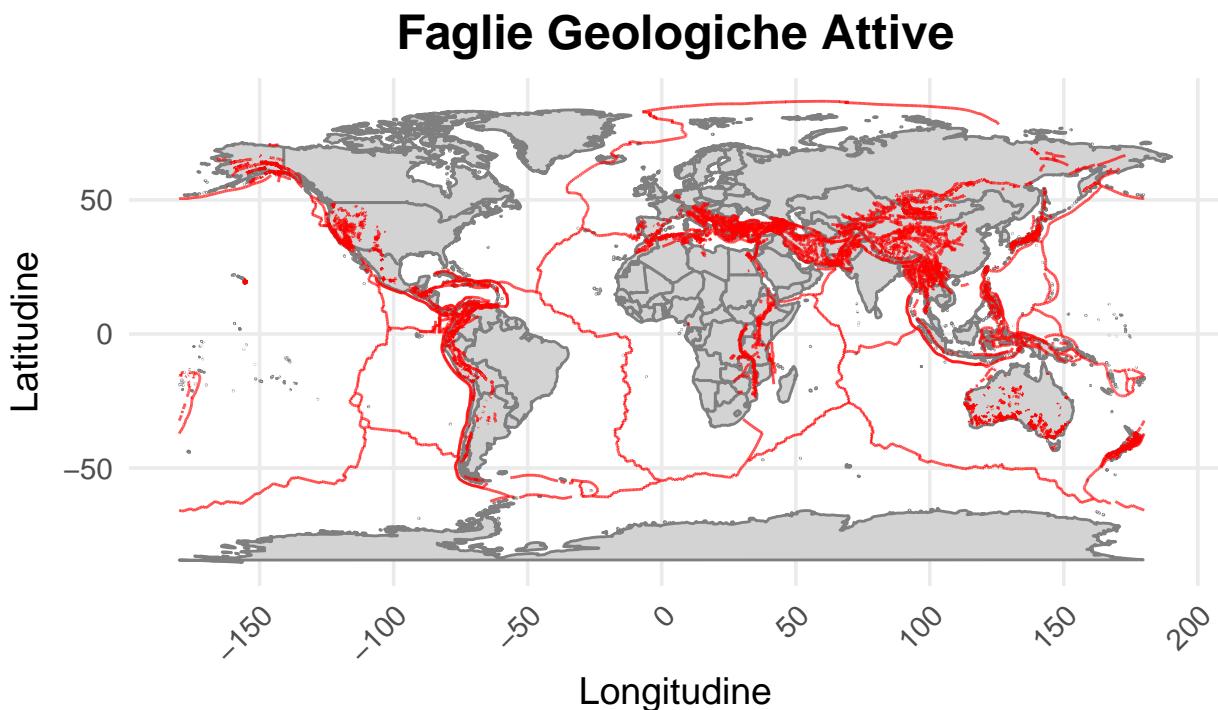
## Reading layer `gem_active_faults' from data source
##   `C:\Users\danie\Desktop\Master AI\R\Progetto_R-\gem_active_faults.shp'
##   using driver `ESRI Shapefile'
```

```

## Simple feature collection with 16195 features and 26 fields
## Geometry type: LINESTRING
## Dimension: XY
## Bounding box: xmin: -180 ymin: -66.163 xmax: 180 ymax: 86.805
## CRS: NA

ggplot() +
  # Aggiungi la mappa del mondo
  borders("world", colour = "gray50", fill = "lightgray") +
  # Aggiungi le faglie
  geom_sf(data = faglie, color = "red", size = 1, alpha = 0.7) +
  labs(title = "Faglie Geologiche Attive", x = "Longitudine", y = "Latitudine",
       color = "Magnitudo") +
  theme_minimal(base_size = 14) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    legend.position="right")

```



Nel grafico, viene mostrata la mappa del mondo in grigio, e in rosso, le faglie geologiche attive. Come si può vedere le faglie si distribuiscono principalmente lungo i confini delle placche tettoniche con una forte concentrazione nella parte sud-ovest dell'Asia a formare la catena dell' Himalaya. La presenza di faglie geologiche è di particolare interesse in quanto nei pressi di queste zone si osserva una maggiore probabilità che vi siano terremoti. Questo fenomeno verrà visualizzato in seguito, attraverso il dataset "Earthquake" in cui è presente la lista dei terremoti registrati dal 1970-2025.

## Dataset Earthquake:

I dati presi dal sito USGS.gov sono stati importati e filtrati con una magnitudo superiore a 5.0. Essendo la frequenza dei terremoti correlata alla magnitudo il sito consentiva solo di scaricare 20000 eventi per volta, producendo troppe richieste di dati all'ente fornitore. Inoltre, il dataset sull'intero fenomeno era piuttosto frammentato e si è deciso quindi di unire le informazioni distribuite su diversi anni in un unico blocco per avere una visione globale del fenomeno.

In R, i dati sono stati importati tramite il seguente codice:

```
terremoti <- read.csv("Earthquake_1970-2025.csv")
```

## Analisi Premilinare.

Il nostro dataset è composto da 22 variabili di varia natura. Per semplicità, mostriamo un breve estratto delle variabili più significative usate nell'analisi del fenomeno. Per una descrizione di tutte le variabili presenti nel database si rimanda a (<https://www.usgs.gov/programs/earthquake-hazards/magnitude-types>).

```
pander(head(terremoti %>%
               select(latitude, longitude, depth, mag, nst, type, time)))
```

latitude	longitude	depth	mag	nst	type	time
-20.18	-70.62	35	5	37	earthquake	2024-12-31T23:13:20.048Z
-6.668	150.6	10	5.1	61	earthquake	2024-12-31T20:09:41.043Z
-4.05	151.6	14.52	5	59	earthquake	2024-12-31T17:09:39.374Z
-17.66	168.2	66.61	5.1	121	earthquake	2024-12-30T05:49:02.808Z
-29.93	-72	10	5.5	127	earthquake	2024-12-30T05:41:06.678Z
-29.92	-72.06	10	5.5	178	earthquake	2024-12-30T05:40:49.261Z

Nella seguente tabella, le variabili indicano:

- **latitudine e longitudine:** rappresentano le coordinate geografiche di dove è avvenuto il terremoto.
- **depth:** La profondità, misurata in km rispetto alla superficie terrestre.
- **mag:** la magnitudo del terremoto in scala Richter.
- **nst:** numero di stazioni che hanno registrato il terremoto.
- **type:** l'origine del terremoto inteso come causa scatenante. (Terremoto terrestre, esplosione nucleare, collasso di miniere, frane ed esplosioni.)
- **time:** data e ora della rilevazione del fenomeno.

Con la funzione summary mostriamo le caratteristiche delle variabili quantitative menzionate sopra.

```
pander(summary(terremoti %>% select(latitude, longitude)))
```

latitude	longitude
Min. :-77.080	Min. :-180.00
1st Qu.:-19.215	1st Qu.: -72.48
Median : -3.692	Median : 102.12
Mean : 1.003	Mean : 40.61
3rd Qu.: 23.936	3rd Qu.: 143.03
Max. : 87.386	Max. : 180.00

```
pander(summary(terremoti %>% select(depth, mag, nst)))
```

depth	mag	nst
Min. : -3.00	Min. :5.000	Min. : 0.0
1st Qu.: 10.00	1st Qu.:5.100	1st Qu.: 69.0
Median : 33.00	Median :5.200	Median :122.0
Mean : 67.16	Mean :5.363	Mean :162.1
3rd Qu.: 57.00	3rd Qu.:5.500	3rd Qu.:219.0
Max. :700.00	Max. :9.100	Max. :934.0
NA	NA	NA's :58768

Ciò che è possibile notare è che per la variabile nst ci sono diversi dati mancanti (NA). Per la variabile depth, si può notare che la mediana e la media non coincidono, lasciando intendere in questa prima fase, la presenza di asimmetria nei dati. Maggiori conferme si hanno guardando la differenza tra il terzo quartile(3rd Qu) e il massimo(Max). Il terzo quartile della variabile mag, ci suggerisce che il 75% dei dati non supera il 5.5 di magnitudo.

```
glimpse(terremoti)
```

```
## Rows: 87,798
## Columns: 22
## $ time <chr> "2024-12-31T23:13:20.048Z", "2024-12-31T20:09:41.043Z"~
## $ latitude <dbl> -20.1826, -6.6675, -4.0500, -17.6555, -29.9272, -29.91~ 
## $ longitude <dbl> -70.6231, 150.5771, 151.6351, 168.2183, -71.9959, -72.~ 
## $ depth <dbl> 35.000, 10.000, 14.523, 66.612, 10.000, 10.000, 34.000~ 
## $ mag <dbl> 5.0, 5.1, 5.0, 5.1, 5.5, 5.5, 5.5, 5.1, 5.5, 5.6, 5.0, ~ 
## $ magType <chr> "mwr", "mb", "mb", "mww", "mww", "mb", "mww", "mww", "m~ 
## $ nst <int> 37, 61, 59, 121, 127, 178, 246, 47, 100, 72, 32, 43, 1~ 
## $ gap <dbl> 129, 46, 78, 99, 76, 74, 32, 116, 78, 86, 84, 90, 55, ~ 
## $ dmin <dbl> 0.470, 2.753, 0.450, 2.402, 0.658, 0.710, 4.341, 1.430~ 
## $ rms <dbl> 0.83, 0.74, 0.97, 0.47, 0.51, 0.54, 0.62, 0.75, 0.67, ~ 
## $ net <chr> "us", "us", "us", "us", "us", "us", "us", "us", "us", ~ 
## $ id <chr> "us6000pgri", "us6000pgqb", "us6000pgpd", "us6000pgf9"~ 
## $ updated <chr> "2025-01-22T05:39:23.037Z", "2025-01-13T17:04:41.040Z"~ 
## $ place <chr> "49 km W of Puerto, Chile", "124 km ESE of Kandrian, P~ 
## $ type <chr> "earthquake", "earthquake", "earthquake", "earthquake"~ 
## $ horizontalError <dbl> 3.29, 9.93, 3.77, 8.57, 3.75, 4.63, 8.74, 6.11, 11.64, ~ 
## $ depthError <dbl> 1.918, 1.340, 3.912, 6.575, 1.829, 1.838, 1.773, 1.900~ 
## $ magError <dbl> 0.061, 0.070, 0.077, 0.065, 0.058, 0.025, 0.054, 0.110~ 
## $ magNst <int> 26, 66, 53, 23, 29, 546, 33, 8, 14, 13, 30, 36, 18, 27~ 
## $ status <chr> "reviewed", "reviewed", "reviewed", "reviewed", "revie~ 
## $ locationSource <chr> "us", "us", "us", "us", "us", "us", "us", "us", ~ 
## $ magSource <chr> "us", "us", "us", "us", "us", "us", "us", "us", ~
```

## Analisi Quantitativa

Per iniziare l'esplorazione delle variabili quantitative presenti nel dataset, abbiamo deciso di usare l'istogramma e la funzione di densità di kernel relativi alla magnitudo, la quale rappresenta una delle variabili più importanti per l'analisi del fenomeno studiato. Nella creazione del grafico relativo alla densità di Kernel, si è scelto manualmente il parametro di bandwidth rispetto a quello ottenibile dalla funzione “bw.nrd0()”, poichè il valore restituito era inferiore allo step di magnitud presenti nei dati (0.1). Ciò causava una densità “frastagliata” data la natura discreta dei valori di magnitudo. Per questo motivo, il parametro scelto è di 0.05 cioè la metà dello “step” scelto per la magnitudo. Il grafico risulta abbastanza smooth rispetto all'istogramma.

```
hist = ggplot(terremoti, aes(x=mag)) +
  geom_histogram(fill="lightblue", color="white", bins=40) +
```

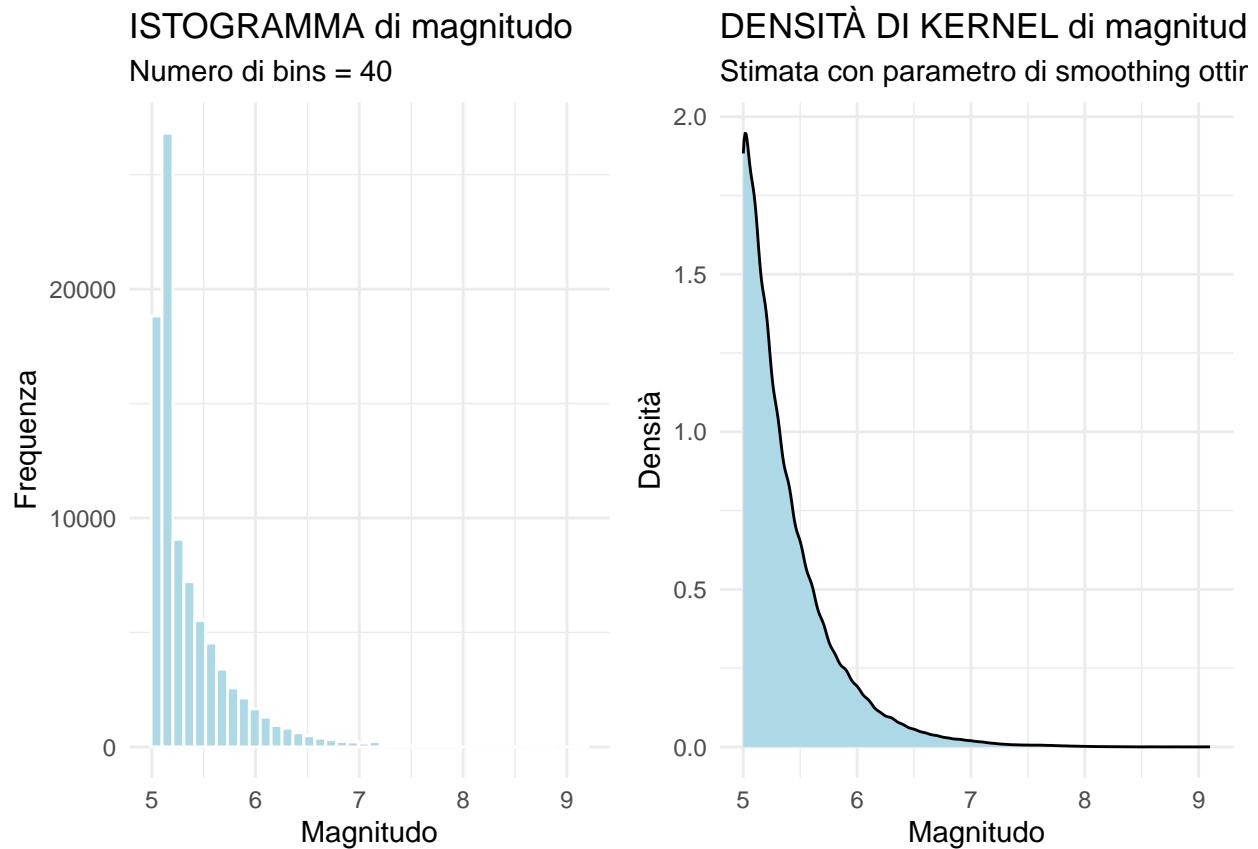
```

theme_minimal() +
labs(title="ISTOGRAMMA di magnitudo",
    subtitle="Numero di bins = 40",
    x="Magnitudo",
    y="Frequenza")

kernel = ggplot(terremoti, aes(x=mag)) +
  geom_density(fill="lightblue", bw=0.05) +
  theme_minimal() +
  labs(title="DENSITÀ DI KERNEL di magnitudo",
      subtitle="Stimata con parametro di smoothing ottimale",
      x="Magnitudo",
      y="Densità")

grid.arrange(hist, kernel, nrow=1)

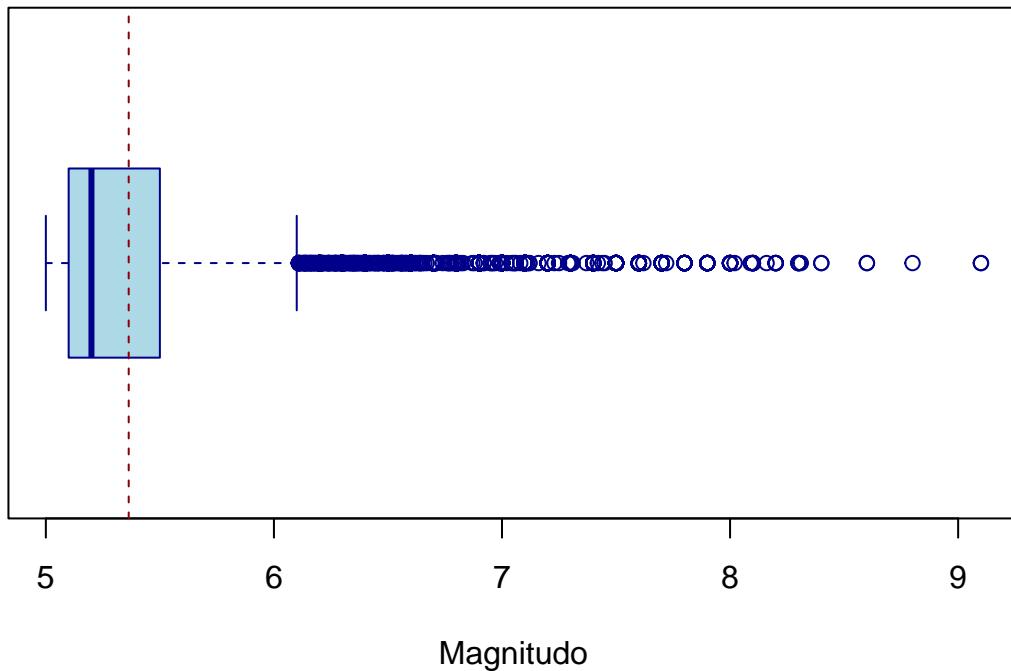
```



Come evidenziato già nell'analisi preliminare, anche dall'esame dei grafici soprastanti, si evince che la maggior parte dei terremoti presenta una magnitudo inferiore a 6. Tuttavia si nota una forte asimmetria positiva con una coda prolungata che evidenzia anche il verificarsi di terremoti con un valore di magnitudo molto superiore alla media, seppur con poca frequenza. Pertanto al fine di esplorare meglio questo andamento abbiamo deciso di riportare anche il box-plot della magnitudo, il quale conferma la presenza di outliers in corrispondenza di valori di magnitudo superiori a 6. Il grafico mostra come la quasi totalità dei valori sia compresa nell'intervallo 5 e 6. Anche la media, rappresentata dalla linea tratteggiata in rosso pari a 5.36 che è compresa in tale intervallo.

```
boxplot(terremoti$mag, col="lightblue", border="darkblue", horizontal=TRUE,
        main="BOXPLOT di magnitudo", xlab="Magnitudo")
abline(v=mean(terremoti$mag), lty=2, col="darkred")
```

## BOXPLOT di magnitudo



Esaminando le variabili presenti nel dataset e la loro descrizione abbiamo ritenuto che ci potesse essere una relazione significativa fra la profondità a cui avviene il fenomeno e la magnitudo dello stesso. Al fine di studiare tale relazione, abbiamo optato per una rappresentazione tramite box-plot e violin.plot, in quanto lo scatterplot risultava confusionario a causa della grossa quantità di dati. Inoltre, sempre per avere una migliore rappresentazione abbiamo raggruppati in fasce i livelli di magnitudo con un passo di 0.5.

```
# Filtra i dati e crea le fasce di magnitudo
DS3_filtered <- terremoti %>%
  filter(depth > 0) %>% # Rimuove profondità negative o zero
  mutate(mag_group = cut(mag, breaks = seq(5, 9, by = 0.5),
                         include.lowest = TRUE)) %>%
  filter(!is.na(mag_group)) %>% # Rimuove fasce di magnitudo senza dati
  group_by(mag_group) %>%
  filter(n() >= 10) %>% # Mantiene solo gruppi con almeno 10 osservazioni
  ungroup()

# Creiamo il violin plot
custom_colors <- c("#1F77B4", "#FF7FOE", "#2CA02C",
                     "#D62728", "#9467BD", "#8C564B",
                     "#E377C2", "#7F7F7F")

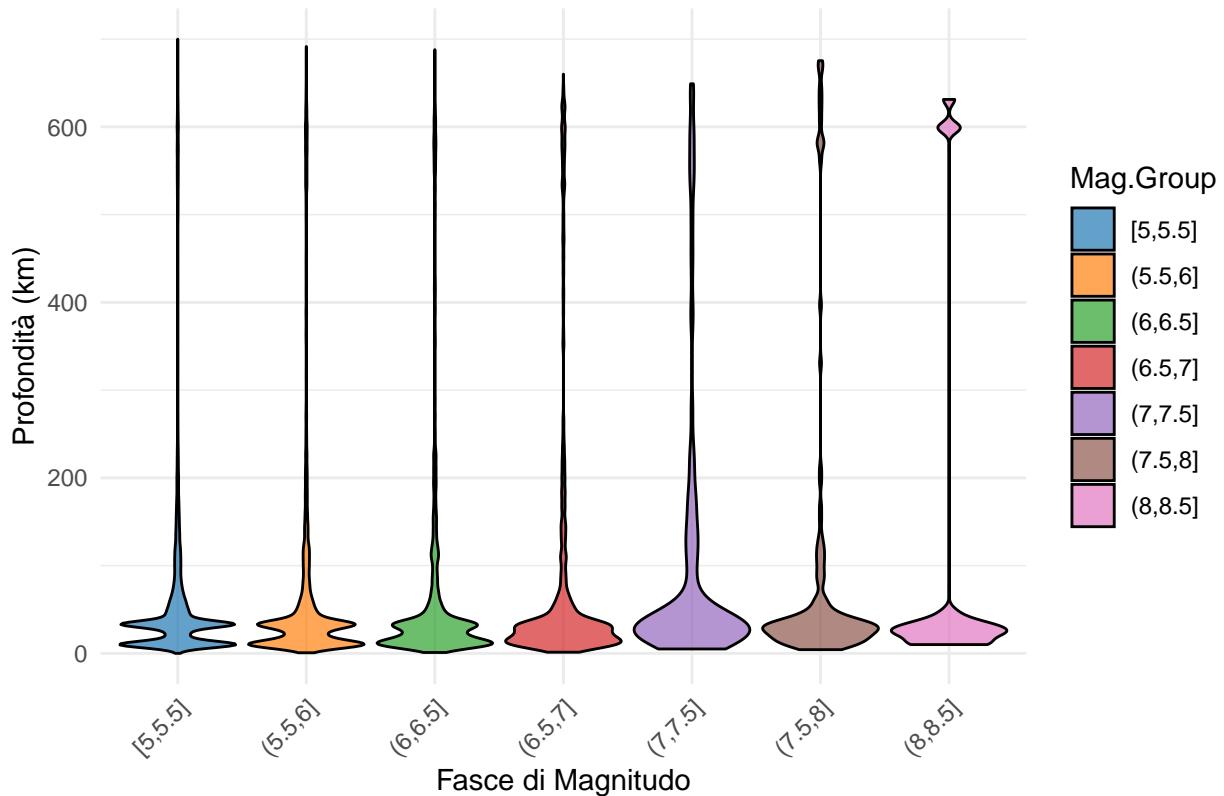
ggplot(DS3_filtered, aes(x = mag_group, y = depth, fill = mag_group)) +
```

```

geom_violin(scale = "width", alpha = 0.7, color = "black") + # Bordo nero
scale_fill_manual(values = custom_colors) + # Usa la palette personalizzata
labs(title = "Distribuzione della profondità per fasce di magnitudo",
x = "Fasce di Magnitudo",
y = "Profondità (km)",
fill="Mag.Group") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

```

Distribuzione della profondità per fasce di magnitudo



```

DS3_filtered <- terremoti %>%
  filter(depth < 300) %>% # Escludiamo solo outlier estremi
  mutate(mag_group = cut(mag, breaks = seq(5, 9, by = 0.5),
                        include.lowest = TRUE))

# Rimuovi le fasce di magnitudo che non contengono dati
DS3_filtered <- DS3_filtered %>%
  filter(!is.na(mag_group))

# Crea il grafico
library(ggsci)

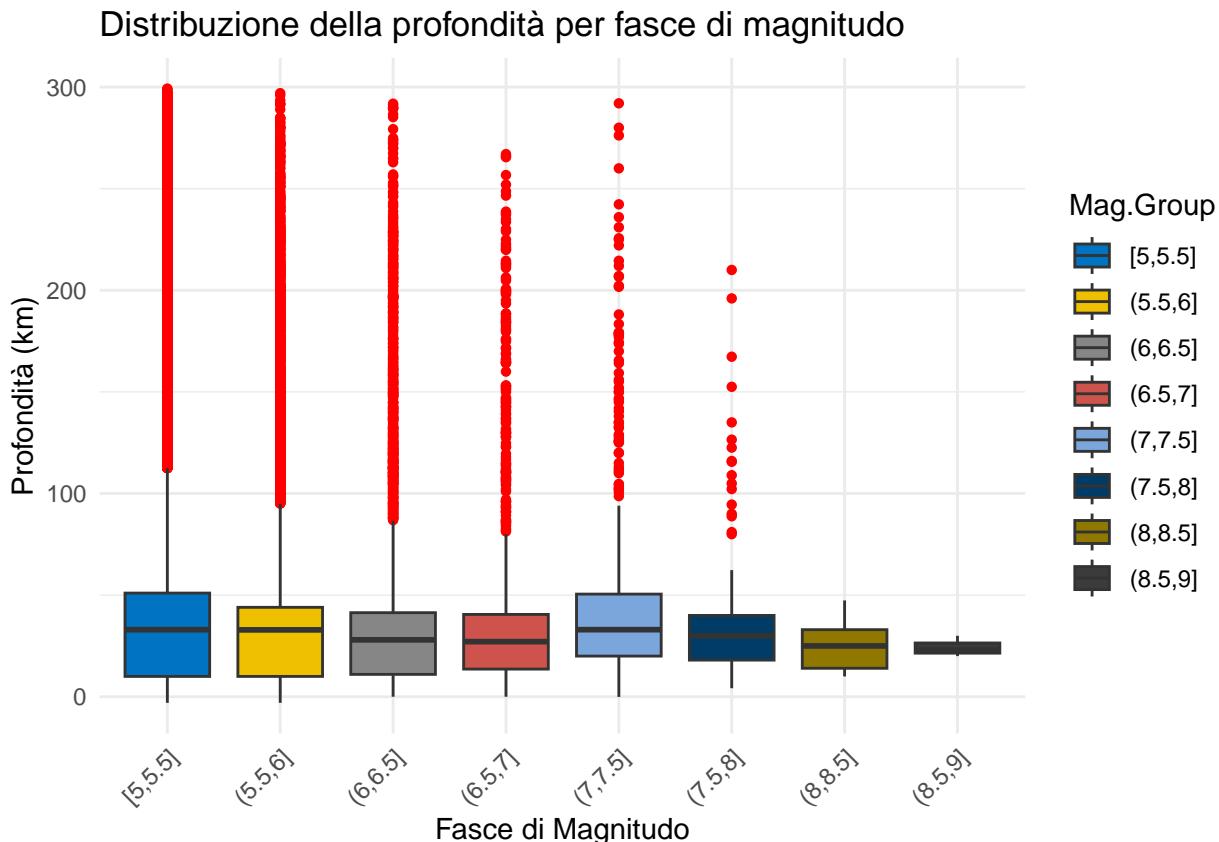
ggplot(DS3_filtered, aes(x = mag_group, y = depth, fill = mag_group)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  scale_fill_jco() + # Usa la palette "jco"
  labs(title = "Distribuzione della profondità per fasce di magnitudo",
       x = "Fasce di Magnitudo",
       y = "Profondità (km)")

```

```

y = "Profondità (km)",
fill="Mag.Group") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

```



Dai grafici soprastanti emerge che la maggior parte dei terremoti si verifica a profondità superficiali (tra 0-70 km), come ben evidenziato dalla posizione del box e dalla mediana. Questo riflette la natura della tettonica delle placche, dove l'attività sismica è più frequente lungo i margini delle placche. Sono presenti anche outliers a grandi profondità (addirittura tra 200 e 300 km), anche se questi eventi sono abbastanza rari.

Successivamente, al fine di esplorare l'esistenza di eventuali relazioni tra le variabili quantitative abbiamo deciso di esaminare la matrice di correlazione, creata utilizzando la funzione “cor” del pacchetto stats, che di default adotta il coefficiente di correlazione lineare di Pearson. Per una migliore interpretazione del valore dei coefficienti, si è deciso di riportare il correlation plot generato tramite il seguente codice.

```

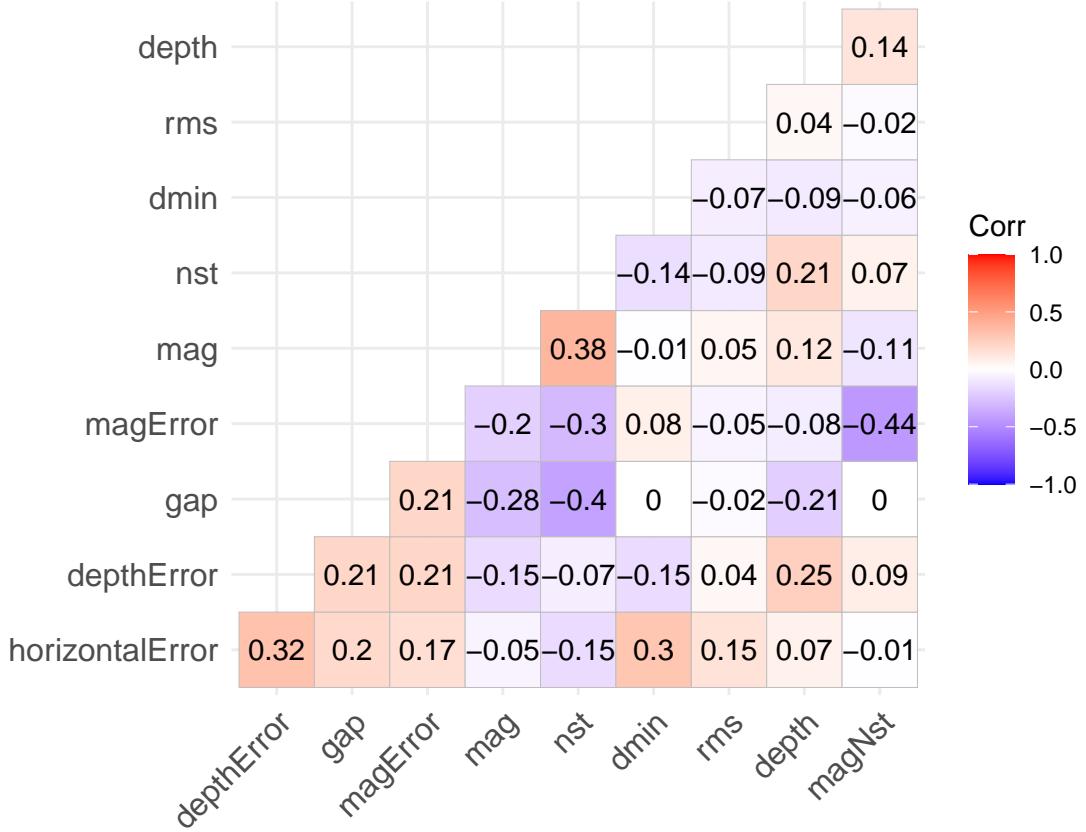
##seleziono le variabili per l'analisi
df_corr= terremoti%>%
  dplyr::select(depth, mag, nst, gap, dmin, rms, horizontalError, depthError,
               magError, magNst)

#### seleziono le variabili da pulire da NA
df_corr_clean <- df_corr %>%
  drop_na(depth, mag, nst, gap, dmin, rms, horizontalError,
          depthError, magError, magNst)

#calcolo la matrice di correlazione
df_f=cor(df_corr_clean)

```

```
ggcorrplot(df_f,
            hc.order = TRUE,
            type = "lower",
            lab = TRUE)
```



Dal correlation plot si evince che le correlazioni fra le variabili di maggiore interesse sono le seguenti:

- **Mag e nst:** Queste due variabili appartengono rispettivamente al livello di magnitudo e al numero di stazioni sismiche che hanno registrato il terremoto. Sono correlate positivamente in quanto un evento con alta magnitudo sarà rilevato da più stazioni sismiche presenti nel mondo.
- **nst e gap:** Con un maggior numero di stazioni (nst elevato), la copertura attorno all'epicentro è più uniforme, riducendo il gap. Al contrario, con meno stazioni la copertura è più concentrata, generando un gap più ampio.
- **Mag Error e magNst:** Queste due variabili appartengono rispettivamente al livello di magnitudo e al numero di stazioni sismiche che hanno partecipato alla determinazione della magnitudo. Per questo motivo osserviamo un coefficiente di correlazione negativo.
- **Horizontal Error e dmin:** L'errore orizzontale potrebbe essere correlato alla distanza minima (dmin) perché terremoti più lontani dalla rete di rilevamento (maggiore dmin) tendono ad avere una localizzazione meno precisa, aumentando l'errore orizzontale. Questo è dovuto alla geometria della rete sismica e alla riduzione della precisione con la distanza.
- **Horizontal Error e depth error:** L'errore orizzontale e l'errore di profondità sono spesso correlati perché entrambi dipendono dalla qualità e dalla distribuzione dei dati sismici. Una scarsa copertura strumentale o una geometria sfavorevole possono influenzare simultaneamente la precisione della localizzazione sia in orizzontale che in profondità.

Attraverso il seguente grafico possiamo dare una migliore interpretazione di quelli che sono i coefficienti di correlazione di alcune variabili già viste in precedenza, unite alla loro densità di kernel, mostrata sulla

diagonale principale. Nella parte triangolare inferiore della matrice possiamo identificare gli scatterplot unito ad un fit di un modello lineare.

```
# seleziono le variabili per l'analisi
df_corr= terremoti%>%
    dplyr::select(depth, mag, nst, gap, dmin, rms)

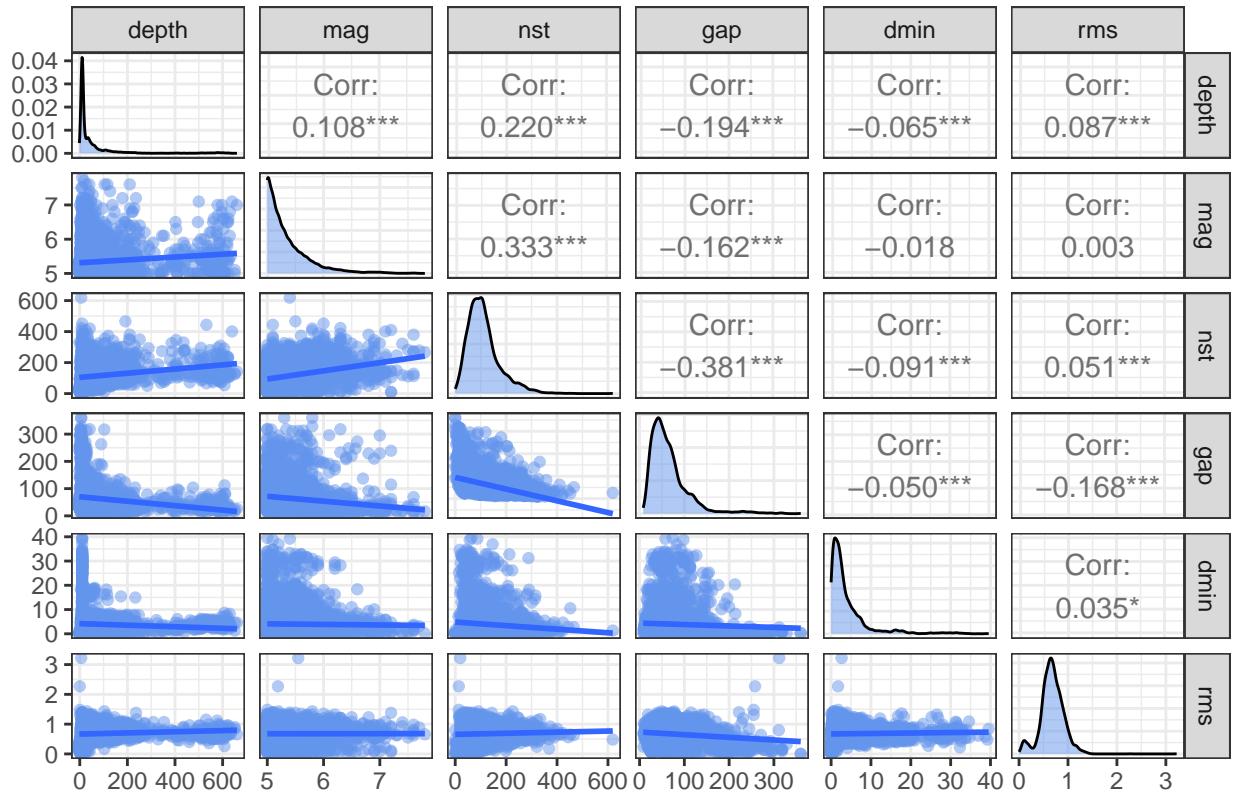
# seleziono le variabili da pulire da NA
df_corr_clean <- df_corr %>%
    drop_na(depth, mag, nst, gap, dmin, rms)

my_density <- function(data, mapping, ...){
  ggplot(data = data, mapping = mapping) +
  geom_density(alpha = 0.5,
  fill = "cornflowerblue", ...)
}

my_scatter <- function(data, mapping, ...){
  ggplot(data = data, mapping = mapping) +
  geom_point(alpha = 0.5,
  color = "cornflowerblue") +
  geom_smooth(method=lm,
  se=FALSE, ...)
}

ggpairs(df_corr_clean,
        lower=list(continuous = my_scatter),
        diag = list(continuous = my_density)) +
labs(title = "Matrice di correlazione tra le variabili") +
theme_bw()
```

## Matrice di correlazione tra le variabili

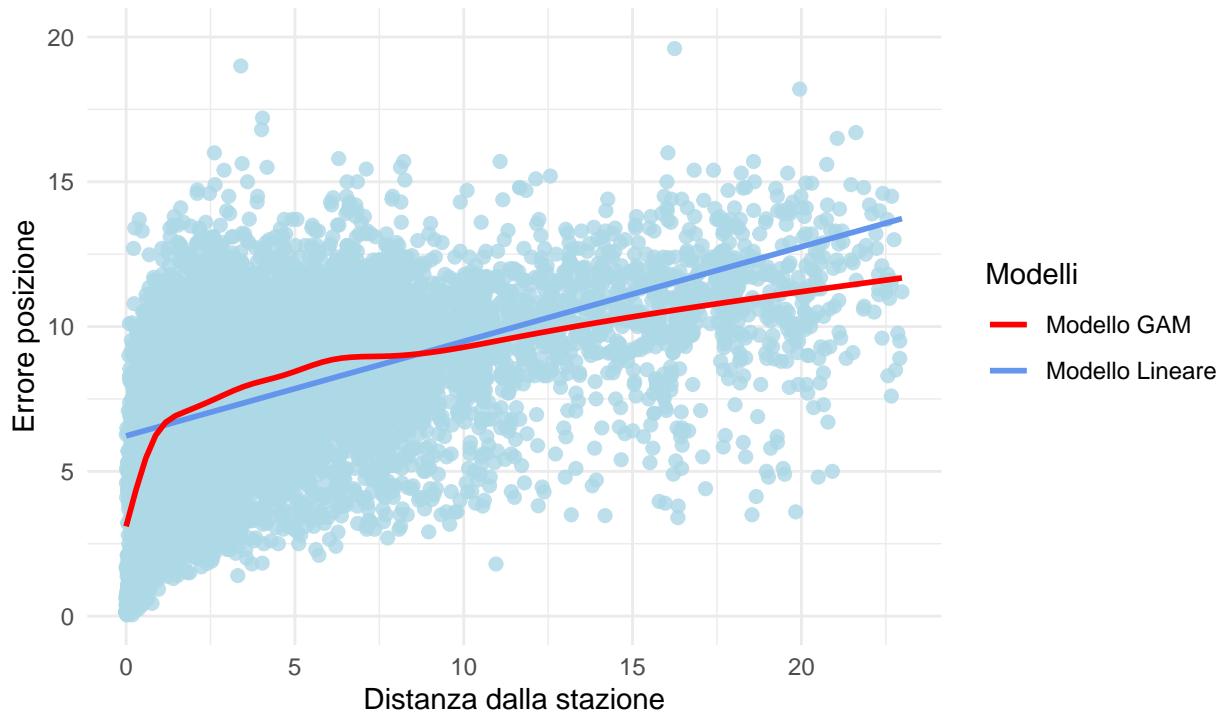


Tra le varie relazioni evidenziate sopra abbiamo deciso di concentrare l'attenzione sul legame tra le variabili dmin (distanza dalla stazione più vicina) e HorizontalError (errore sul rilevamento della posizione), che presentano una correlazione positiva pari a 0.30, come mostrato nel primo correlation plot. Pertanto al fine di approfondire questa relazione abbiamo deciso di rappresentare lo scatterplot delle due variabili e di confrontare l'adattamento di un modello lineare rispetto ad un modello GAM (Generalized Additive Model), il quale è il modello che si adatta meglio ai dati. In conclusione, pertanto, possiamo dire che la relazione tra queste due variabili risulta non lineare.

```
ggplot(terremoti, aes(x = dmin, y = horizontalError)) +
  geom_point(color = "lightblue", alpha = 0.8, size = 2) +
  geom_smooth(method = "lm", se = FALSE, aes(color = "Modello Lineare")) +
  geom_smooth(se = FALSE, aes(color = "Modello GAM")) +
  scale_color_manual(
    name = "Modelli",
    values = c("Modello Lineare" = "cornflowerblue", "Modello GAM" = "red"))
  ) +
  scale_y_continuous(limits = c(0, 20)) +
  scale_x_continuous(breaks = seq(0, 23, 5), limits = c(0, 23)) +
  theme_minimal() +
  labs(
    title = "Distanza dalla stazione più vicina vs Errore sul rilevamento della posizione",
    subtitle = "Modello Lineare vs Modello GAM",
    x = "Distanza dalla stazione",
    y = "Errore posizione")
  ) +
  theme(legend.position = "right")
```

## Distanza dalla stazione più vicina vs Errore sul rilevamento della posizione

Modello Lineare vs Modello GAM

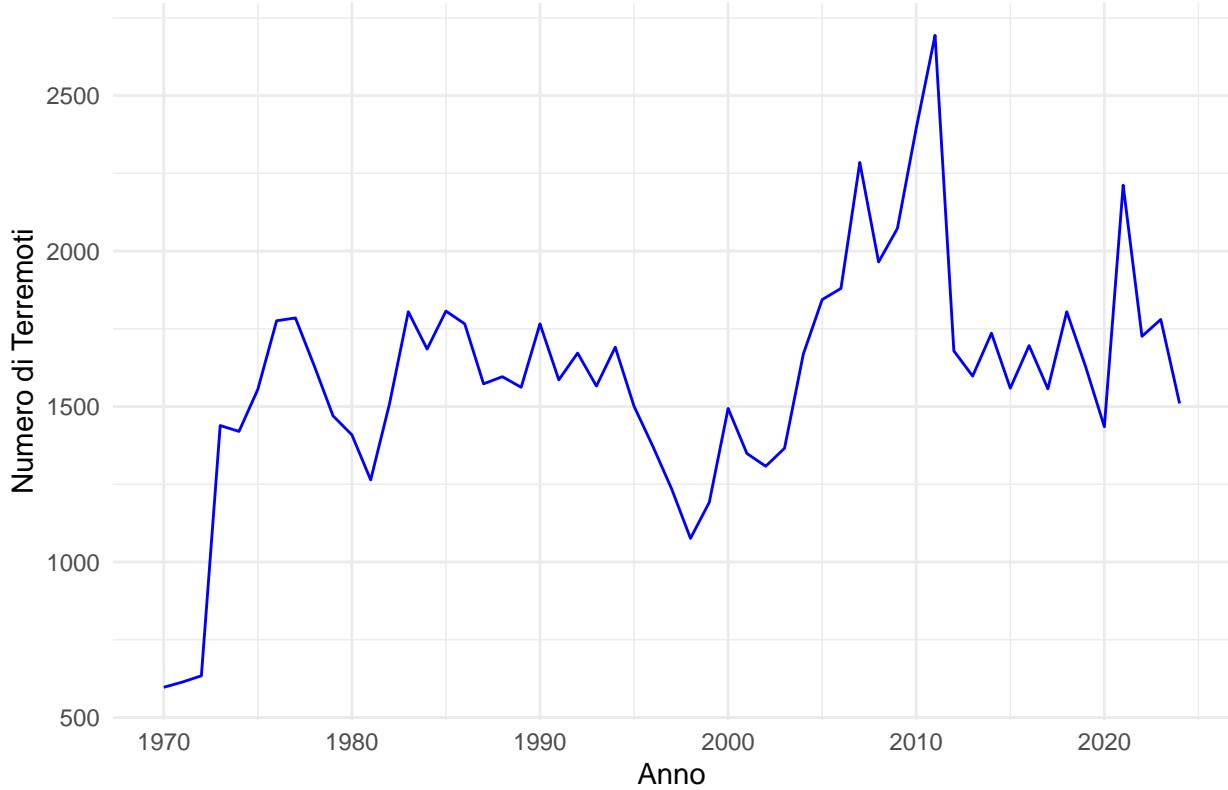


Infine, abbiamo deciso di analizzare l'evoluzione del numero di terremoti avvenuti a livello globale nel periodo di riferimento (1970-2025).

```
gg=  terremoti %>%
  mutate(Anno=year(as.Date(time)))%>%
  group_by(Anno) %>%
  summarise(Conteggio= n(), magnitudomedia=mean(mag))

ggplot(gg, aes(x = Anno, y = Conteggio)) +
  geom_line(color = "blue") +
  theme_minimal()+
  labs(title = "Serie Temporale dei Terremoti (Raggruppati per Anno)",
       x = "Anno", y = "Numero di Terremoti")
```

## Serie Temporale dei Terremoti (Raggruppati per Anno)



La serie storica annuale evidenzia come il numero di terremoti sia contenuto nell'intervallo tra 1500-2000 terremoti per la maggior parte degli anni studiati. Tuttavia, nel periodo compreso tra il 2005 e il 2010 c'è stato un incremento significativo del numero di terremoti, raggiungendo il picco, superiore a 2500 terremoti, nel 2010.

Nel grafico sottostante, viene riproposta la serie storica dei valori, con una differenza rispetto alla versione precedente: i valori di magnitudine sono stati suddivisi in intervalli di 0.5. Per ciascun intervallo, è stato conteggiato il numero di terremoti per ogni anno. Al fine di migliorare la visibilità dei gruppi con magnitudine inferiore, è stata applicata una scala logaritmica sull'asse delle ordinate.

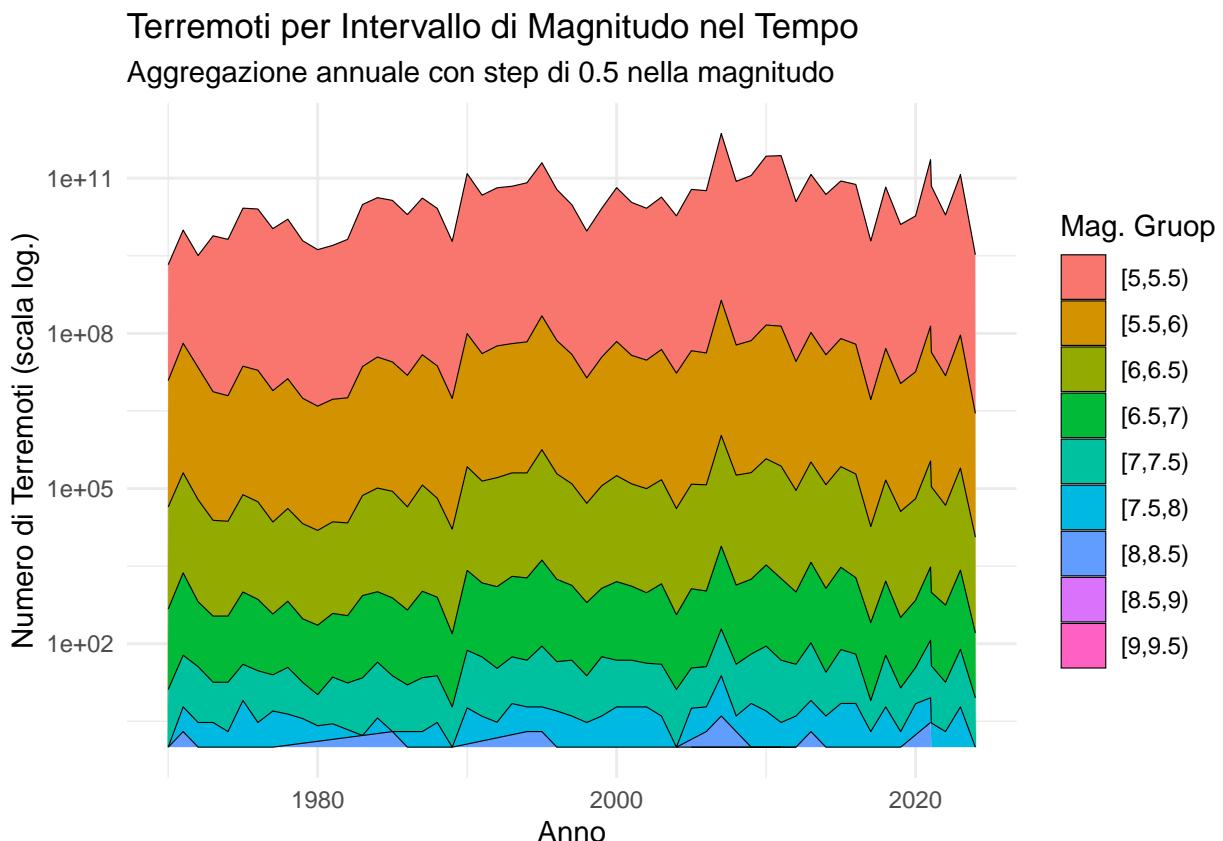
```
df_year <- terremoti %>%
  mutate(time = ymd_hms(time, tz = "UTC"),
        year = floor_date(time, unit = "year"),
        mag_bin = cut(mag,
                      breaks = seq(floor(min(mag, na.rm = TRUE)),
                                    ceiling(max(mag, na.rm = TRUE)),
                                    by = 0.5),
                      include.lowest = TRUE,
                      right = FALSE)) %>%
  group_by(year, mag_bin) %>%
  summarise(count = n(), .groups = "drop")

# Creazione del grafico ad area
ggplot(df_year, aes(x = year, y = count, fill = mag_bin)) +
  geom_area(color = "black", size = 0.2) +
  scale_y_log10() +
  labs(title = "Terremoti per Intervallo di Magnitudo nel Tempo",
       subtitle = "Aggregazione annuale con step di 0.5 nella magnitudo",
```

```

x = "Anno",
y = "Numero di Terremoti (scala log.)",
fill = "Mag. Gruop") +
theme_minimal()

```



Avendo informazioni relative alla latitudine e alla longitudine c'è sembrato interessante anche studiare questo fenomeno in base al continente dove è avvenuto, con l'idea di cercare di capire in quali continenti il fenomeno risulta particolarmente concentrato. A tal fine mediante il codice che riportiamo di seguito abbiamo prima collegato il continente e lo stato alle osservazioni del database oggetto di studio, e successivamente lo abbiamo filtrato andando a considerare soltanto una parte delle osservazioni contenute in esso. Questo perché a causa di una diversa approssimazione di latitudine e longitudine per alcune osservazioni il collegamento con lo stato e il continente non risultava particolarmente preciso.

```

library(sp)
library(rworldmap)

coordinates_continents = function(points)
{
  countriesSP <- getMap(resolution='high')
  pointsSP = SpatialPoints(points, proj4string=CRS(proj4string(countriesSP)))
  indices = over(pointsSP, countriesSP)
  indices$REGION
}

coordinates_country = function(points)
{

```

```

countriesSP <- getMap(resolution='high')
pointsSP = SpatialPoints(points, proj4string=CRS(proj4string(countriesSP)))
indices = over(pointsSP, countriesSP)
indices$ADMIN
}

terremoti2 = terremoti

terremoti2$continent = coordinates_continents(data.frame(
  as.integer(terremoti$longitude),
  as.integer(terremoti$latitude)))

terremoti2$continent = recode(
  terremoti2$continent,
  "South America and the Caribbean"="South America")

terremoti2$country = coordinates_country(data.frame(
  as.integer(terremoti$longitude),
  as.integer(terremoti$latitude)))

terremoti2 = filter(terremoti2,
  type=="earthquake" &
    continent!="<NA>" &
    continent!="Antarctica")

```

Innanzitutto, riportiamo il pie chart relativo al continente nel quale è avvenuto l'evento, da cui si evince come la maggior parte dei terremoti ha avuto luogo in Asia (circa il 44%) e in Sudamerica (circa il 32%).

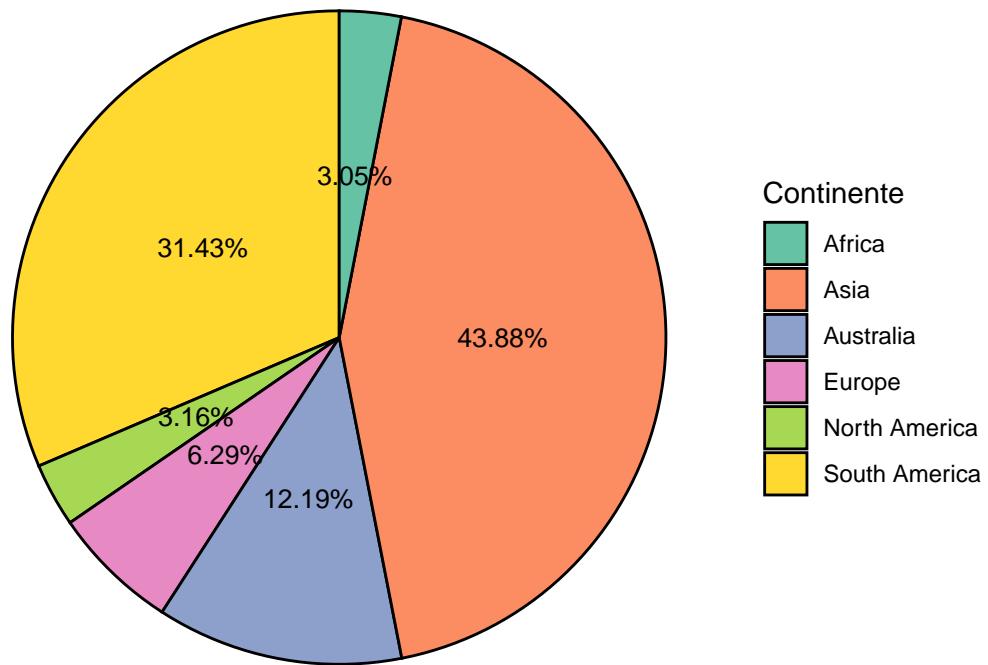
```

data_piechart = terremoti2 %>% count(continent) %>%
  arrange(desc(continent)) %>%
  mutate(perc = round(n/sum(n)*100, 2),
    ypos_l = cumsum(perc)-0.5*perc,
    perc_l = paste0(round(n/sum(n)*100, 2), "%"))

ggplot(data_piechart, aes(x="", y=perc, fill=continent)) +
  geom_bar(stat="identity", width=1, color="black") +
  geom_text(aes(y=ypos_l, label=perc_l), color="black", size=3.5) +
  coord_polar("y", start=0, direction=-1) +
  theme_void() +
  scale_fill_brewer(palette="Set2") +
  labs(title="PIE CHART di continente",
    fill="Continente")

```

## PIE CHART di continente



```
rm(data_piechart)
```

Successivamente, al fine di vedere il comportamento del fenomeno nei diversi continenti, abbiamo deciso di concentrare l'attenzione sulla variabile relativa alla magnitudo. A tal fine, di seguito riportiamo alcune statistiche di sintesi della magnitudo per continente.

```
cbind(min = tapply(terremoti2$mag, droplevels(terremoti2$continent), min),
      max = tapply(terremoti2$mag, droplevels(terremoti2$continent), max),
      mean = tapply(terremoti2$mag, droplevels(terremoti2$continent), mean),
      median = tapply(terremoti2$mag, droplevels(terremoti2$continent), median),
      sd = tapply(terremoti2$mag, droplevels(terremoti2$continent), sd))

##          min max     mean median       sd
## Africa      5 7.3 5.306256   5.20 0.3731212
## Asia        5 7.9 5.341656   5.20 0.4131597
## Australia    5 8.1 5.354538   5.20 0.4171046
## Europe      5 7.8 5.368365   5.20 0.4263263
## North America 5 7.9 5.384919   5.24 0.4437618
## South America 5 8.8 5.398311   5.22 0.4642237
```

Inoltre, riportiamo di seguito prima il ridgeline graph e poi il cleveland plot. Avendo notato che il valore massimo della magnitudo è stato rilevato in Sudamerica, abbiamo deciso di approfondire il comportamento del fenomeno in questa macroregione. Specifichiamo che nei grafici relativi al solo Sudamerica sono stati considerati soltanto gli stati per cui il numero di osservazioni a disposizione fosse almeno pari a 30.

```
library(ggridges)
ridge_1 = ggplot(terremoti2, aes(x=mag, y=continent, fill=continent)) +
  geom_density_ridges() +
```

```

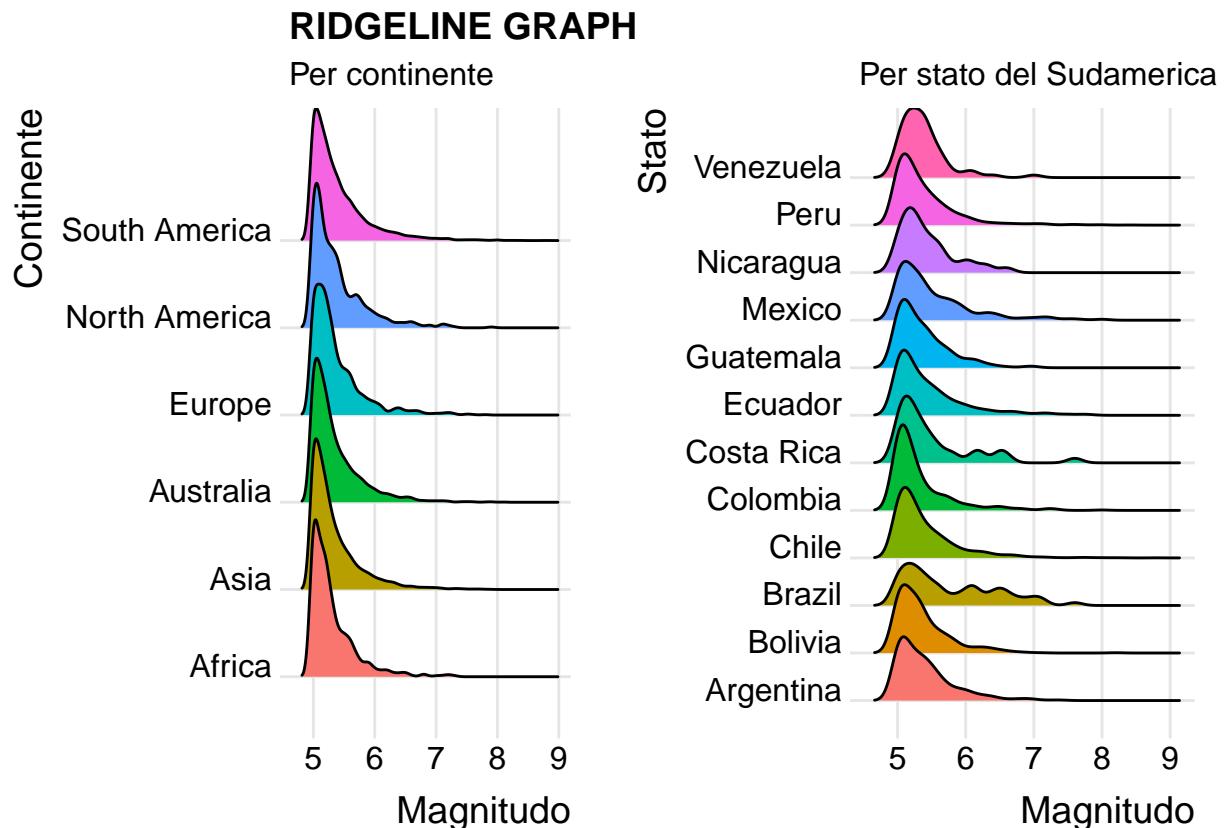
theme_ridges() +
theme(legend.position="none") +
labs(title="RIDGEGRAPH",
    subtitle="Per continente",
    x="Magnitudo",
    y="Continente")

data_ridge_2a = terremoti2 %>% filter(continent=="South America")
tab_ridge_2 = data.frame(table(droplevels(data_ridge_2a$country))) %>%
  filter(Freq>=30) %>%
  rename(country=Var1)
data_ridge_2b = inner_join(data_ridge_2a, tab_ridge_2, by="country")

ridge_2 = ggplot(data_ridge_2b, aes(x=mag, y=country, fill=country)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position="none") +
  labs(title="",
    subtitle="Per stato del Sudamerica",
    x="Magnitudo",
    y="Stato")

grid.arrange(ridge_1, ridge_2, nrow=1)

```



```
rm(ridge_1, ridge_2, data_ridge_2a, data_ridge_2b, tab_ridge_2)
```

Dal ridgeline graph per continente si evince che le varie distribuzioni presentano un andamento molto simile tra loro e simile a quello relativo alla distribuzione della magnitudo sull'intero database, mostrato nell'istogramma di pagina 5. Concentrando poi l'attenzione sul Sudamerica, la distribuzione della magnitudo risulta sempre simile tra i vari stati con una forte assimmetria positiva; soltanto il Brasile presenta una distribuzione visibilmente differente.

```
data_cleveland = terremoti2 %>% group_by(continent) %>%
  summarise(max=max(mag))

cleveland_1 = ggplot(data_cleveland, aes(x=max, y=reorder(continent,max))) +
  geom_point(color="darkblue", size = 2.5) +
  geom_segment(aes(x=min(terremoti2$mag), xend=max,
                    y=reorder(continent,max),
                    yend=reorder(continent,max)),
                color="lightblue") +
  theme_minimal() +
  theme(panel.grid.major=element_blank(),
        panel.grid.minor=element_blank()) +
  labs(title="CLEVELAND PLOT di magnitudo massima",
       subtitle="Per continente",
       x="Magnitudo",
       y="")

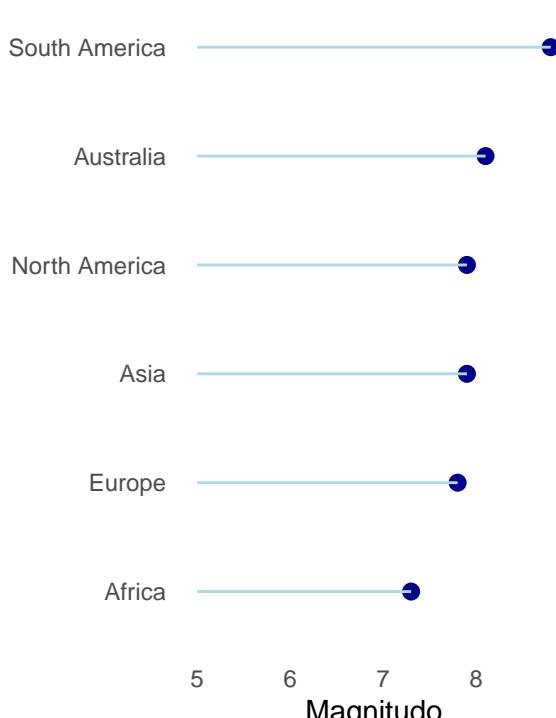
data_cleveland_2a = terremoti2 %>% filter(continent=="South America")
tab_cleveland_2 = data.frame(table(droplevels(data_cleveland_2a$country))) %>%
  filter(Freq>=30) %>% rename(country=Var1)
data_cleveland_2b = inner_join(data_cleveland_2a, tab_cleveland_2,
                                by="country") %>%
  group_by(country) %>% summarise(max=max(mag))

cleveland_2 = ggplot(data_cleveland_2b, aes(x=max, y=reorder(country,max))) +
  geom_point(color="darkblue", size = 2.5) +
  geom_segment(aes(x=min(terremoti2$mag), xend=max,
                    y=reorder(country,max),
                    yend=reorder(country,max)),
                color="lightblue") +
  theme_minimal() +
  theme(panel.grid.major=element_blank(),
        panel.grid.minor=element_blank()) +
  labs(title="",
       subtitle="Per stato del Sudamerica",
       x="Magnitudo",
       y="")

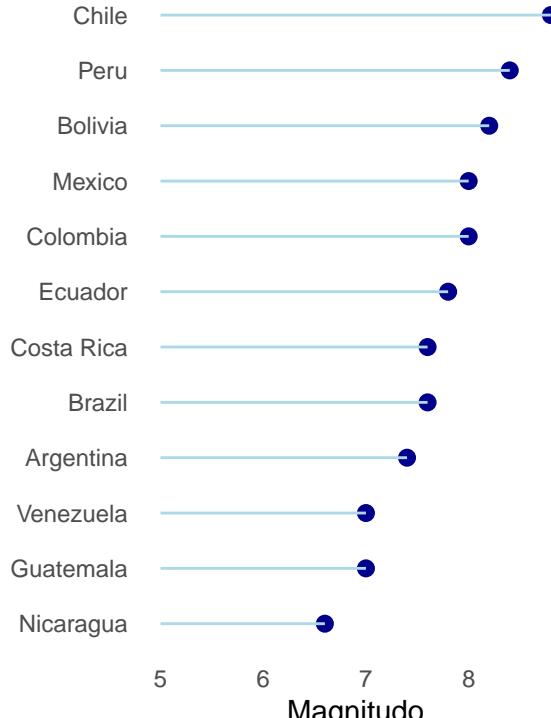
grid.arrange(cleveland_1, cleveland_2, nrow=1)
```

## CLEVELAND PLOT di magnitudo massima

Per continente



Per stato del Sudamerica



```
rm(data_cleveland, cleveland_1, cleveland_2, data_cleveland_2a,
  data_cleveland_2b, tab_cleveland_2)
```

Infine, il Cleveland plot soprastante riporta per i vari continenti e per i vari stati sudamericani la magnitudo massima rilevata. Da essi si evince come il valore massimo della magnitudo nel periodo 1975-2025 è stato rilevato in Sudamerica e in particolare in Cile.

Come accennato in precedenza, esiste una correlazione tra le faglie tettoniche e la probabilità di verificarsi di un terremoto. A tal proposito, qui sotto è riportata una mappa del mondo in cui sono evidenziate in rosso le faglie attive. I pallini rappresentano un sotto-campione contenente il 15% di tutti i terremoti presenti nel dataset di origine tettonica. Per ogni terremoto, la dimensione del pallino è proporzionale alla magnitudo dell'evento sismico.

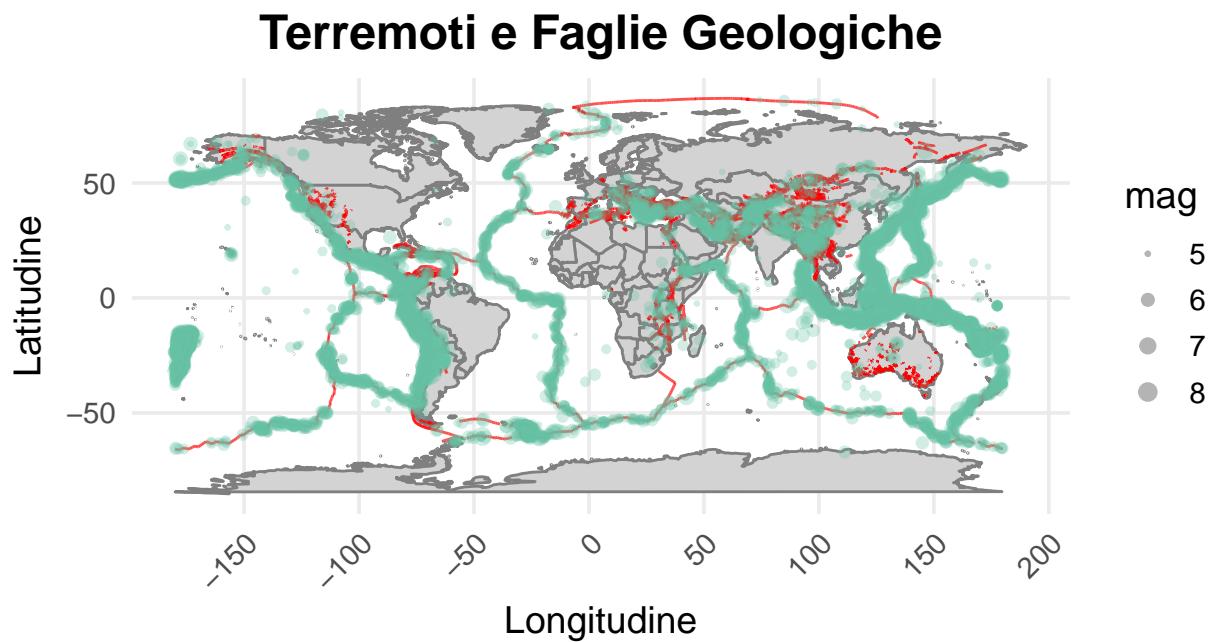
```
terremoti_sampled <- terremoti %>%
  group_by(type) %>%
  filter(!runif(n()) < 0.85) %>%
  filter(type == "earthquake")

ggplot() +
  borders("world", colour = "gray50", fill = "lightgray") +
  geom_sf(data = faglie, color = "red", size = 1, alpha = 0.7) +
  geom_point(data = terremoti_sampled, aes(x = longitude, y = latitude,
                                             color = type,
                                             size = mag), alpha = 0.3) +
  scale_color_brewer(palette = "Set2") +
  scale_size_continuous(range = c(0.5, 3)) +
  labs(title = "Terremoti e Faglie Geologiche",
```

```

x = "Longitudine",
y = "Latitudine") +
theme_minimal(base_size = 14) +
theme(
  axis.text.x = element_text(angle = 45, hjust = 1),
  plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
  legend.position = "right"
) +
guides(color = "none")

```



Come si può osservare dal grafico, i terremoti tendono a concentrarsi principalmente lungo le faglie tettoniche. Tuttavia, nel dataset sono presenti anche terremoti di origine diversa, come quelli causati dall'uomo o da vulcani. Per questo motivo, nel grafico sottostante sono rappresentati tutti i terremoti di origine non tettonica. Nel grafico i terremoti sono stati suddivisi per colore in funzione della loro origine.

```

terremoti_sampled_2 <- terremoti %>%
  filter(type != "earthquake")

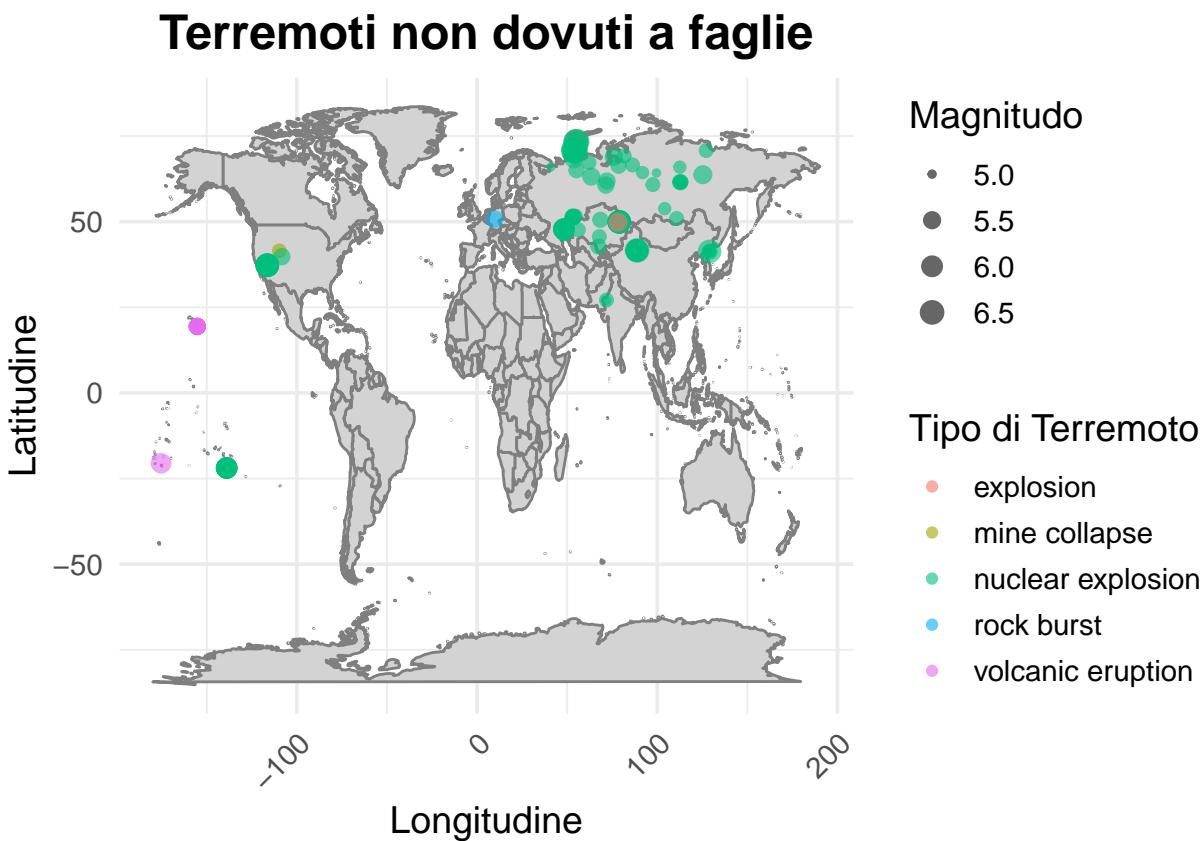
ggplot() +
  borders("world", colour = "gray50", fill = "lightgray") +
  geom_point(data = terremoti_sampled_2, aes(x = longitude,
                                              y = latitude,
                                              color = type,
                                              size = mag),
              alpha = 0.6) +
  scale_size_continuous(range = c(1, 4)) +
  labs(title = "Terremoti non dovuti a faglie",

```

```

x = "Longitudine", y = "Latitudine",
color = "Tipo di Terremoto", size = "Magnitudo") +
theme_minimal(base_size = 14) +
theme(
  axis.text.x = element_text(angle = 45, hjust = 1),
  plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
  legend.position = "right"
)

```



Come osservato in precedenza, i terremoti sono per la maggior parte di origine tettonica. Per questo motivo, è stato creato un barplot che mostra la distribuzione percentuale tra i terremoti di origine tettonica e tutti gli altri eventi sismici, dal quale emerge che solo l'1% di tutti i terremoti verificatisi nel periodo studiato è stato causato da fattori legati all'attività umana o vulcanica, mentre la restante parte è dovuta a fenomeni di origine esclusivamente naturale.

```

frequenza_terremoti <- terremoti %>%
  count(type) %>%
  mutate(Percentuale = n / sum(n) * 100)

frequenza_terremoti <- frequenza_terremoti %>%
  mutate(Percentuale = n / sum(n) * 100) %>%
  mutate(type = ifelse(Percentuale < 1, "Altri", as.character(type))) %>%
  group_by(type) %>%
  summarise(n = sum(n), Percentuale = sum(Percentuale)) %>%
  ungroup()

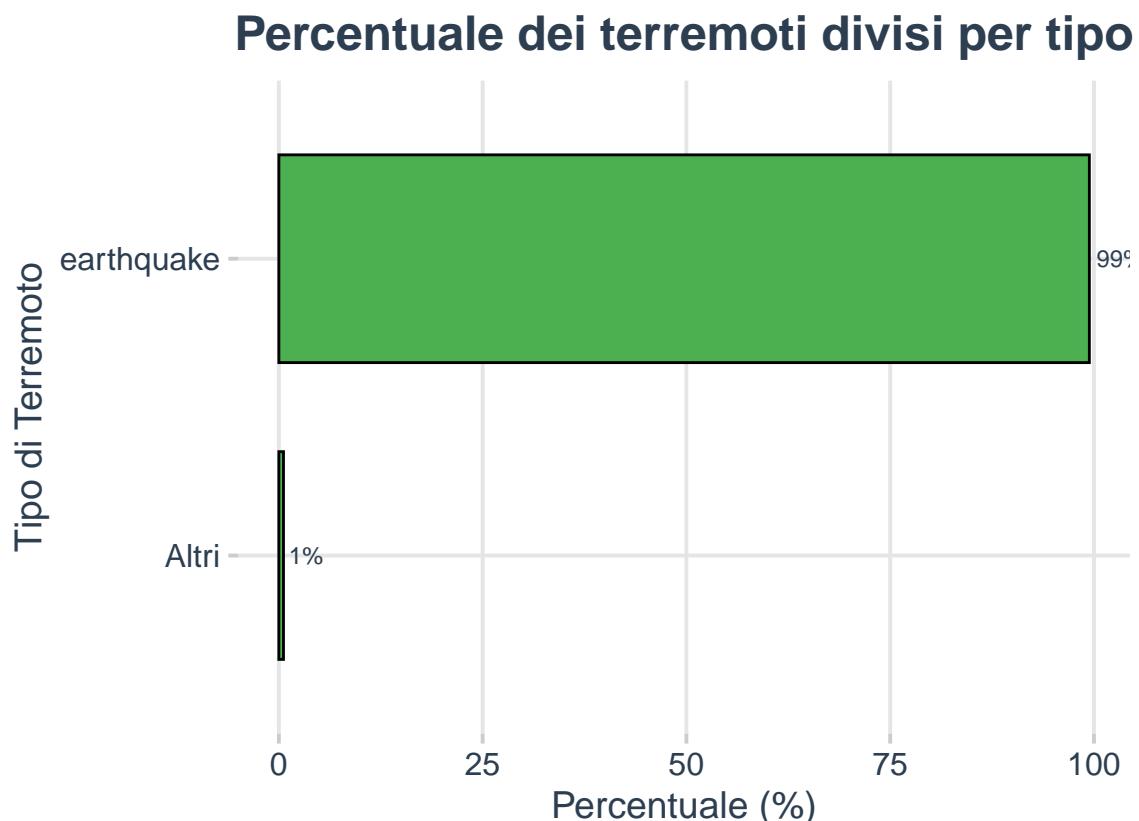
ggplot(data = frequenza_terremoti,

```

```

aes(x = reorder(type, Percentuale), y = Percentuale)) +
geom_bar(stat = "identity", fill = "#4CAF50", color = "black", width = 0.7) +
coord_flip() +
labs(title = "Percentuale dei terremoti divisi per tipo",
x = "Tipo di Terremoto",
y = "Percentuale (%)") +
theme_minimal(base_size = 15) +
theme(
  plot.title = element_text(hjust = 0.5, size = 18,
                            face = "bold",
                            color = "#2C3E50"),
  axis.text.x = element_text(size = 12, color = "#2C3E50"),
  axis.text.y = element_text(size = 12, color = "#2C3E50"),
  axis.title = element_text(size = 14, color = "#2C3E50"),
  panel.grid.major = element_line(color = "gray90"),
  panel.grid.minor = element_blank(),
  axis.ticks = element_line(color = "gray80"),
  plot.margin = margin(10, 20, 10, 30),
) +
geom_text(aes(label = paste0(round(Percentuale, 0), "%")),
          hjust = -.15, color = "#2C3E50", size = 3)

```



In particolare, osservando il grafico sottostante, si può notare che la maggior parte dei terremoti non dovuti a faglie sono statati causati dall'uomo con delle esplosioni nucleari.

```

frequenza_terremoti_minori <- terremoti %>%
  filter(type != "earthquake") %>%

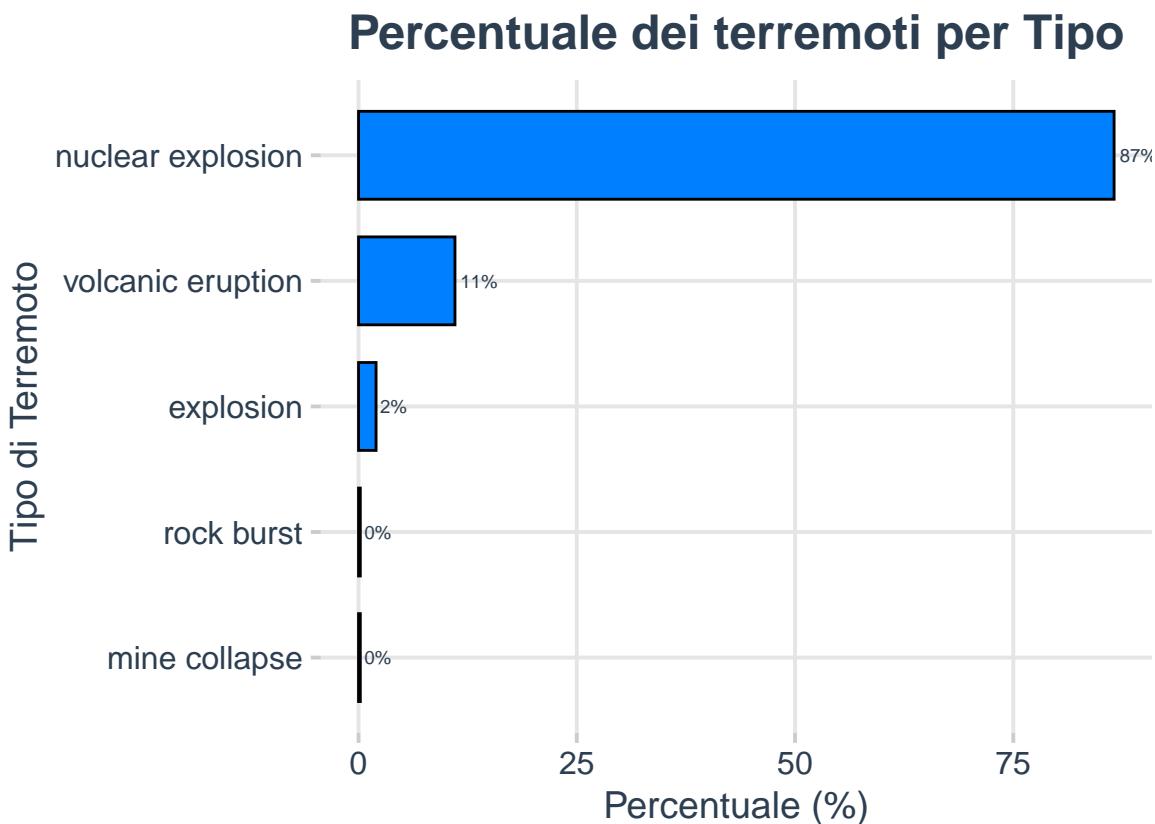
```

```

count(type) %>%
mutate(Percentuale = n / sum(n) * 100)

ggplot(data = frequenza_terremoti_minori, aes(x = reorder(type, Percentuale),
                                               y = Percentuale)) +
  geom_bar(stat = "identity", fill = "#007fff", color = "black", width = 0.7) +
  coord_flip() +
  labs(title = "Percentuale dei terremoti per Tipo",
       x = "Tipo di Terremoto",
       y = "Percentuale (%)") +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold",
                               color = "#2C3E50"),
    axis.text.x = element_text(size = 12, color = "#2C3E50"),
    axis.text.y = element_text(size = 12, color = "#2C3E50"),
    axis.title = element_text(size = 14, color = "#2C3E50"),
    panel.grid.major = element_line(color = "gray90"),
    panel.grid.minor = element_blank(),
    axis.ticks = element_line(color = "gray80"),
    plot.margin = margin(10, 20, 10, 20)
  ) +
  geom_text(aes(label = paste0(round(Percentuale, 0), "%")), hjust = -0.15,
            color = "#2C3E50", size = 2.6)

```



Nel dataset analizzato, oltre alle colonne precedenti, è presente anche la variabile **place**, che contiene informazioni sulla localizzazione del terremoto, incluso lo stato in cui si è verificato. Per estrapolare lo stato di appartenenza di ciascun terremoto, è stato adottato un approccio differente rispetto al precedente. La difficoltà principale risiedeva nel fatto che gli Stati Uniti d'America non sono indicati come un'unica nazione (USA), ma come singoli stati ("Alabama", "Alaska", "Arizona", ecc.). Di conseguenza, è stato necessario raggrupparli sotto un unico nome. Inoltre, a fianco del nome della nazione, a volte comparivano termini come "region", "earthquake", "baja", che hanno richiesto un'eliminazione manuale. Questo approccio ha escluso alcuni terremoti nei quali la variabile è risultata vuota o non rientrava nel pattern utilizzato per gli altri.

Tuttavia, come sarà spiegato successivamente, questo nuovo approccio ha permesso di includere alcuni terremoti che in precedenza erano stati esclusi, ma ha probabilmente ne ha escluso altri. Qui di seguito è riportato il codice utilizzato per il filtraggio e il raggruppamento degli eventi sismici.

```
usa_state = c("Alabama", "Alaska", "Arizona", "Arkansas",
             "California", "Colorado", "Connecticut",
             "Delaware", "Florida", "Georgia", "Hawaii",
             "Idaho", "Illinois", "Indiana", "Iowa",
             "Kansas", "Kentucky", "Louisiana", "Maine",
             "Maryland", "Massachusetts", "Michigan",
             "Minnesota", "Mississippi", "Missouri", "Montana",
             "Nebraska", "Nevada", "New Hampshire",
             "New Jersey", "New Mexico", "New York", "North Carolina",
             "North Dakota", "Ohio", "Oklahoma",
             "Oregon", "Pennsylvania", "Rhode Island", "South Carolina",
             "South Dakota", "Tennessee", "Texas", "Utah", "Vermont",
             "Virginia", "Washington", "West Virginia", "Wisconsin", "Wyoming")

usa_state = tolower(usa_state)

# Ho filtrato anche le parole: region, earthquake, sequence, california-baja
earthquakes <- terremoti %>%
  filter(type == "earthquake") %>%
  separate(place, into = c("location", "state"), sep = ", ", extra = "drop") %>%
  mutate(state = trimws(state)) %>%
  filter(!is.na(state)) %>%
  mutate(state = tolower(state)) %>%
  mutate(state = str_replace_all(state,
                                 "region|earthquake|sequence|california-baja",
                                 "")) %>%
  mutate(state = str_trim(state)) %>%
  mutate(state = case_when(
    state %in% usa_state ~ "usa",
    TRUE ~ state # mantiene il valore originale per le altre nazioni
  )) %>%
  count(state) %>%
  mutate(Percentuale = n / sum(n) * 100) %>%
  arrange(-n)

state_map <- map_data("state")
world_map <- map_data("world")

centroids_state <- state_map %>%
  group_by(region) %>%
```

```

mutate(region = tolower(region)) %>%
summarise(
  latitude = mean(lat),
  longitude = mean(long)
)

centroids_world <- world_map %>%
  group_by(region) %>%
  mutate(region = tolower(region)) %>%
  summarise(
    latitude = mean(lat),
    longitude = mean(long)
  )

# Ora uniamo i due dataset
merged_data <- merge(earthquakes, centroids_world, by.x = "state",
                      by.y = "region", all.x = TRUE)

# Rimuovi i valori NA nelle coordinate
merged_data_clean <- merged_data %>%
  filter(!is.na(longitude) & !is.na(latitude)) %>%
  arrange(-n)

# ----- Vediamo quali sono gli stati con più terremoti -----
limite = 5

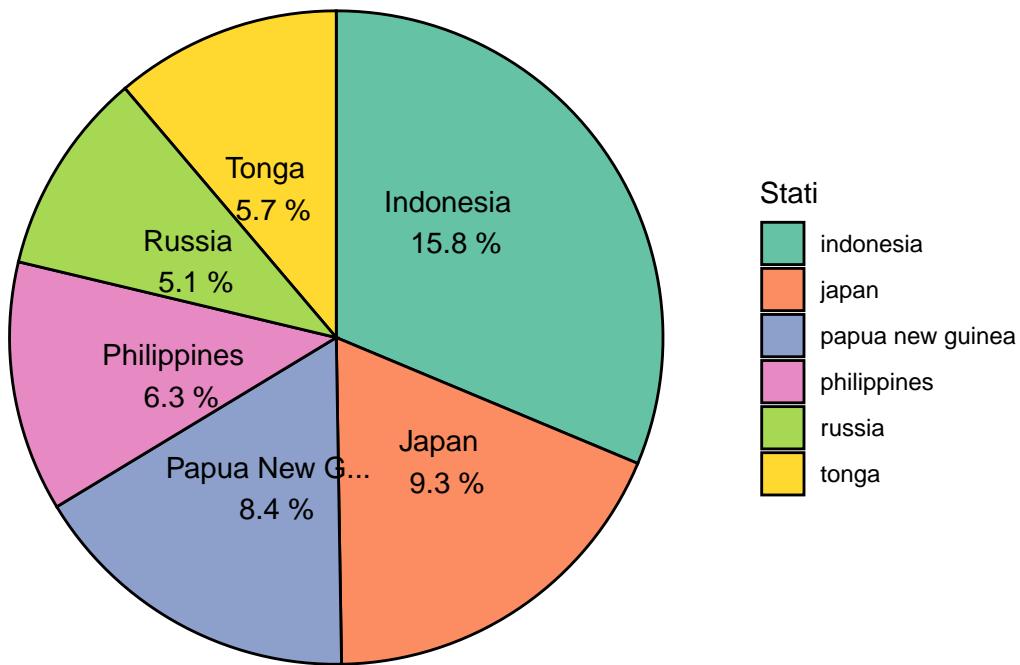
# Funzione per filtrare gli stati a più alta percentuale
filtra_percentuale <- function(dataset, soglia = 3) {
  dataset %>%
    filter(Percentuale >= soglia)
}

stati_importanti <- filtra_percentuale(merged_data_clean, limite)

ggplot(stati_importanti, aes(x="", y=Percentuale, fill=state)) +
  geom_bar(stat="identity", width=1, color="black") +
  geom_text(aes(label = paste(str_trunc(str_to_title(state), width = 14,
                                    ellipsis = "..."), "\n",
                                    round(Percentuale, 1), "%")),
            position = position_stack(vjust = 0.6), size = 3.8,
            color = "black") +
  coord_polar("y", start=0, direction=-1) +
  theme_void() +
  scale_fill_brewer(palette="Set2") +
  labs(title="PIE CHART degli Stati",
       fill="Stati")

```

## PIE CHART degli Stati



Come si può osservare dal dataset, è stato possibile suddividere i terremoti in base alla nazione di appartenenza utilizzando il nuovo metodo. Grazie a questo approccio, è stato possibile creare un grafico a torta che mostra la distribuzione percentuale delle nazioni con la maggiore frequenza di terremoti nella serie storica considerata.

Per un ulteriore analisi, è stato realizzato un altro grafico in cui gli stati sono stati colorati in modo diverso in base alla percentuale globale di terremoti che si sono verificati al loro interno. Le nazioni per cui non è stato possibile reperire dati, o che probabilmente sono stati esclusi durante il processo di filtraggio, sono stati rappresentati in grigio.

```
world <- ne_countries(scale = "medium", returnclass = "sf")

asia <- world %>% filter(continent == "Asia")

# Converte i dati dei punti in un oggetto sf
merged_data_sf <- st_as_sf(merged_data_clean,
                           coords = c("longitude", "latitude"),
                           crs = st_crs(asia))

# Filtro i punti che sono dentro il continente africano
merged_data_filtered <- st_intersection(merged_data_sf, asia)

# Estraggono le coordinate dalla geometria
merged_data_filtered <- merged_data_filtered %>%
  mutate(longitude = st_coordinates(geometry)[, 1],
        latitude = st_coordinates(geometry)[, 2])

st_crs(faglie) <- 4326 # Imposta il CRS di 'faglie'
```

```

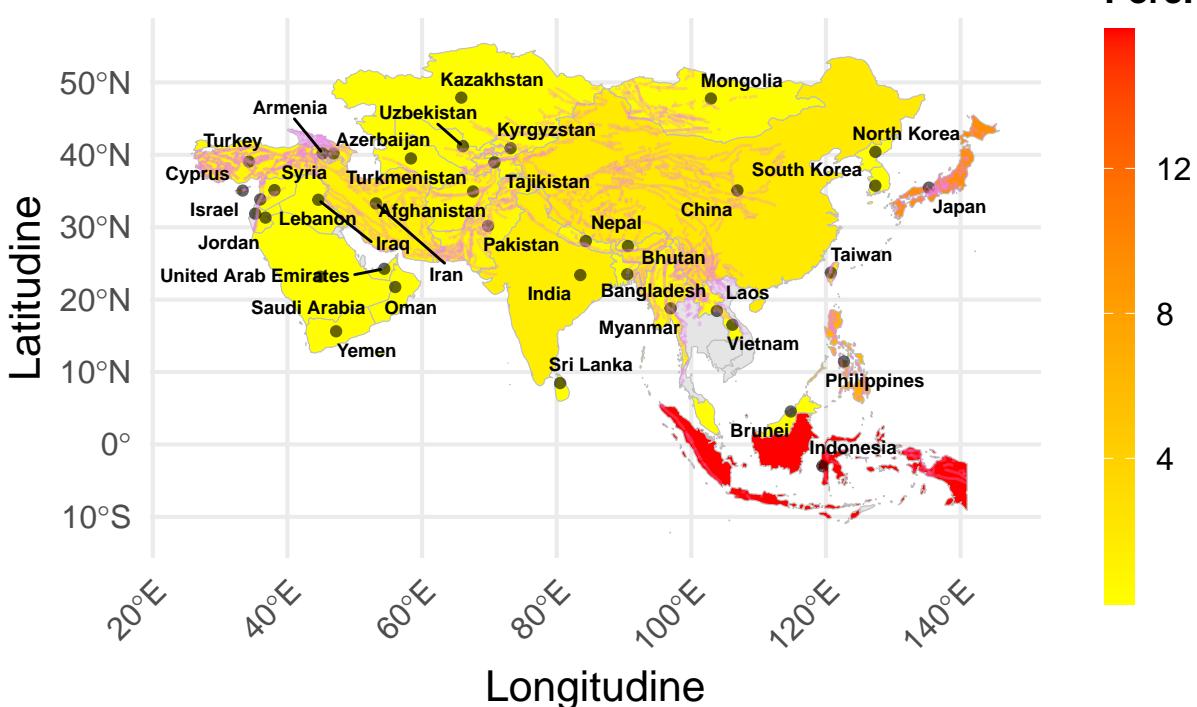
st_crs(asia) <- 4326      # Imposta il CRS di 'asia'
faglie_in_asia <- st_intersection(faglie, asia)

asia_merged <- st_join(asia, merged_data_sf %>% select(state, Percentuale))

ggplot(data = asia_merged) +
  geom_sf(aes(fill = Percentuale), color = "gray") +
  scale_fill_gradient(low = "yellow", high = "red", na.value = "gray90",
                      name = "Percentuale") +
  geom_point(data = merged_data_filtered, aes(x = longitude,
                                              y = latitude), alpha = 0.6) +
  geom_sf(data = faglie_in_asia, color = "violet", size = 1, alpha = 0.3) +
  geom_text_repel(data = merged_data_filtered,
                  aes(x = longitude, y = latitude, label = str_to_title(state)),
                  size = 2.5, fontface = "bold") +
  labs(title = "Asia e faglie geologiche", x = "Longitudine",
       y = "Latitudine",
       fill = "Percentuale",
       caption = 'In grigio sono indicati gli stati su cui non abbiamo dati.') +
  guides(fill = guide_colorbar(title = "Perc.",
                               title.position = "top",
                               barwidth = 0.8, barheight = 15)) +
  theme_minimal(base_size = 16) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 12),
    axis.text.y = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, size = 20, face = "bold"),
    legend.position = "right",
    legend.title = element_text(size = 14, face = "bold"),
    legend.text = element_text(size = 12),
  )
)

```

# Asia e faglie geologiche



In grigio sono indicati gli stati su cui non abbiamo dati.

Dal nuovo dataset è stato possibile riprodurre nuovamente il grafico Cleveland, che mostra gli stati con i terremoti più forti nell'intera serie storica considerata. Come si può osservare, ora è visibile un terremoto di magnitudo 9.1 in Giappone, che era stato escluso nel grafico precedente, nella categoria Asia, per lo stesso tipo di grafico. Questo evidenzia come il nuovo approccio abbia consentito di includere eventi sismici che in precedenza erano stati omessi.

```
max_magitudes <- terremoti %>%
  filter(type == "earthquake") %>%
  separate(place, into = c("location", "state"), sep = ",",
         extra = "drop") %>%
  mutate(state = trimws(state)) %>%
  filter(!is.na(state)) %>%
  mutate(state = tolower(state)) %>%
  mutate(state = str_replace_all(
    state,
    "region|earthquake|sequence|california-baja", ""))
  mutate(state = str_trim(state)) %>%
  mutate(state = case_when(
    state %in% usa_state ~ "usa",
    TRUE ~ state
  )) %>%
  group_by(state) %>%
  summarise(Max_Mag = max(mag, na.rm = TRUE))

limite = 7.9

# Funzione per filtrare gli stati a più alta mag
```

```

filtra_percentuale <- function(dataset, soglia = 3) {
  dataset %>%
    filter(Max_Mag >= soglia)
}

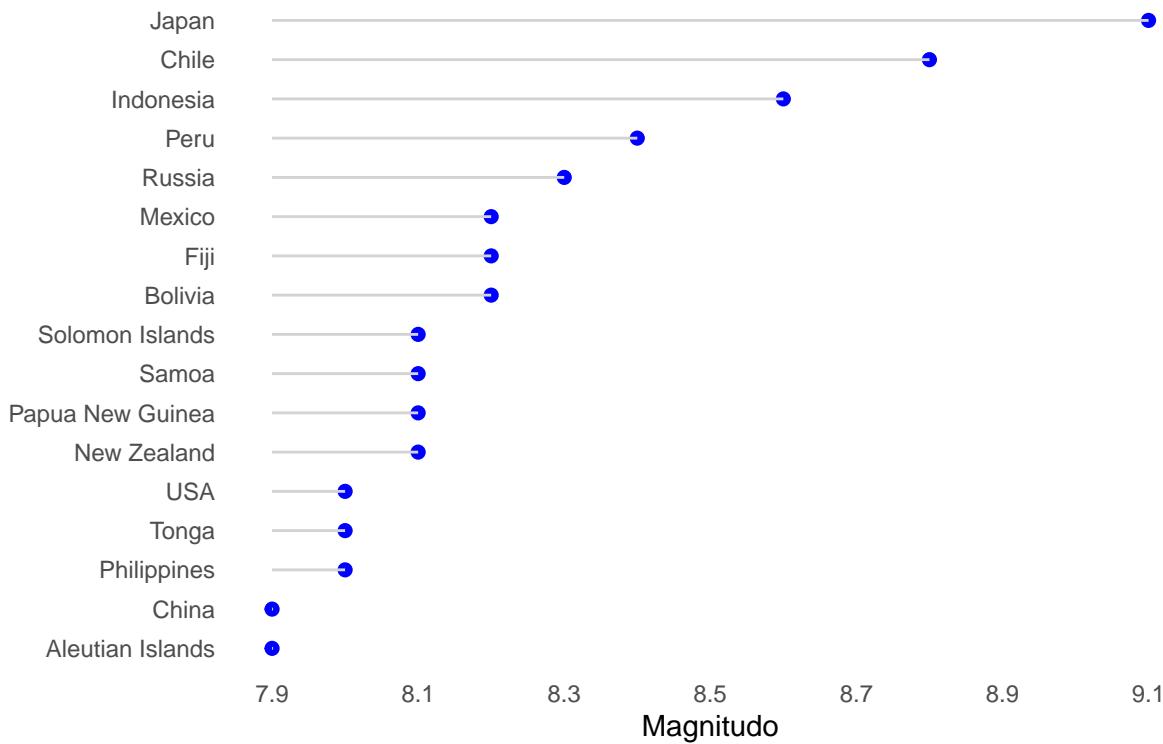
max_magnitudes <- filtra_percentuale(max_magnitudes, limite)

ggplot(max_magnitudes,
       aes(x=Max_Mag,
           y=reorder(state, Max_Mag))) +
  geom_point(color="blue",
             size = 2) +
  geom_segment(aes(x = limite,
                   xend = Max_Mag,
                   y = reorder(state, Max_Mag),
                   yend = reorder(state, Max_Mag)),
               color = "lightgrey") +
  scale_x_continuous(breaks = seq(limite, 10, by = 0.2)) +
  scale_y_discrete(labels = function(x) ifelse(tolower(x) == "usa",
                                                "USA", str_to_title(x))) +
  labs (x = "Magnitudo",
        y = "",
        title = "Terremoti più forti per ogni stato",
        subtitle = "Dal 1970 al 2014") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())

```

## Terremoti più forti per ogni stato

Dal 1970 al 2014



Infine, per visualizzare il contributo dei vari tipi di terremoti rispetto al totale nella serie storica, è stato creato il grafico di cui sotto. Tuttavia, poiché la maggior parte dei terremoti è di origine tettonica (earthquake), è stato necessario escludere questa tipologia per evidenziare le altre. La serie storica è stata suddivisa in gruppi con intervalli di 5 anni e, per ogni gruppo, è stato misurato il numero di terremoti per ciascuna tipologia. Il grafico mostra, quindi, come è variata nel tempo la percentuale delle diverse tipologie di terremoti rispetto al totale.

```
df_filtered <- terremoti %>%
  filter(type != "earthquake") %>%
  mutate(time = ymd_hms(time, tz = "UTC"),
         year = year(time)) %>%
  mutate(year_group = cut(year,
                          breaks = seq(min(year), max(year) + 5, by = 5),
                          include.lowest = TRUE,
                          right = FALSE)) %>%
  group_by(type, year_group) %>%
  summarise(count = n(), .groups = "drop")

ggplot(df_filtered, aes(x = year_group, fill = type, weight = count)) +
  geom_bar(position = "fill") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.margin = margin(0, 50, 0, 0)) +
  scale_fill_brewer(palette = "Set1") +
  scale_y_continuous(labels = scales::percent) +
  labs(y = 'Percentuale', x = "Gruppi di Anni",
       fill = "Tipologia",
```

```
title = "Percentuale per tipologie di terremoti negli anni")
```

