# Tipología y ciclo de vida de los datos: PRA2

Autores: Erick Franz García Miranda y Rafael Eduardo Garcia Agramonte

Junio 2021

---

## Descripción del dataset

---

El dataset corresponde al generado en la PRA1 utilizando tecnicas de webscraping con Python al sitio web: https://www.volcanodiscovery.com/es/earthquakes/ que muestra información de eventos sismicos al rededor del mundo. Se realizó una extensión del código de webscraping, donde se agregaron 2 columnas: "Latitud" y "Longitud" del sismo.

El dataset contiene información de terremotos de 15 paises de centro y sur america como son: Argentina, Bolivia, Brasil, Chile, Colombia, Costa Rica, Ecuador, Guatemala, México, Panamá, Paraguay, Perú, Puerto Rico, República Dominicana, Venezuela); en el perido de 01 de enero del 2010 hasta el 05 de abril del 2021. El dataset tiene formato .csv y tiene **12507 registros.**

Cada registro representa un evento sismico y contiene los siguientes campos:

- **Fecha y Hora:** Fecha en formato día-mes-año y hora específicos en que ocurrió el sismo.
- **Magnitud:** Indica la de magnitud del sismo en escala sismológica de richter.
- **Profundidad:** Distancia de donde se origina el fenómeno con respecto al centro de la tierra.
- **Ubicación:** Lugar en donde se produjo el sismo.
- **Año:** Año en que ocurrió el sismo.
- **País:** Nombre de la nación afectada por el sismo.
- **Latitud:** Coordenada Y de la ubicación del sismo.
- **Longitud:** Coordenada X de la ubicación del sismo.

La importancia del dataset se encuentra en lo impredecible y potencialmente debastador que es un evento sismico. Un registro completo de estos eventos hace que el potencial de analisis del dataset sea elevado.

En este proyecto se utilizará el dataset para responder la pregunta general de **"¿Existe algún patrón o patrones en la actividad sismica de america latina (centro america y sur america)?"**
Este pregunta la podemos dividir en 3 preguntas especificas:

- **¿Cuál es el país con actividad sismica más intensa?**
- **¿Qué variables son más relevantes o significativas en la ocurrencia de un sismo moderado?**
- **¿Cual es la relación entre las variables de un sismo?**

---

## Integración y selección

---

Primero, importaremos el dataset.

```
# Importamos el data set
data_terremotos<-read.csv("../data/terremotos_centro_sur_america.csv",
encoding="UTF-8", header=T,sep=",")
paises_de_america<-read.csv("../data/Paises.csv", encoding="ANSI",
header=T,sep=",")
head(data_terremotos)
```

```
##               Fecha.y.Hora Magnitud Profundidad
## 1 30 Dec 2010 16:35:33 GMT      4.1      109 km
## 2 21 Dec 2010 20:04:47 GMT      4.7      187 km
## 3 19 Dec 2010 07:43:52 GMT      4.5      108 km
## 4 18 Dec 2010 13:07:52 GMT      4.3      124 km
## 5 17 Dec 2010 22:14:26 GMT      4.8      568 km
## 6 17 Dec 2010 20:34:24 GMT      4.8       72 km
##
Ubicacion
## 1                       24 km al norte de Provincia de Nazca, Ica, Perú
## 2         16 km al sureste de Vilavila, Provincia de Lampa, Puno, Perú
## 3        46 km al suroeste de Yauri, Provincia de Espinar, Cusco, Perú
## 4 7.6 km al noroeste de Chuñune, Provincia de Caylloma, Arequipa, Perú
## 5                 85 km al suroeste de Tarauaca, Estado de Acre, Brasil
## 6              30 km al noreste de Ilo, Departamento de Moquegua, Perú
##   Longitud Latitud  Año Pais
## 1  -74.980 -14.620 2010 peru
## 2  -70.587 -15.310 2010 peru
## 3  -71.732 -15.069 2010 peru
## 4  -71.553 -15.284 2010 peru
## 5  -71.460  -8.501 2010 peru
## 6  -71.103 -17.481 2010 peru
```

```
# Cantidad de registros
dim(data_terremotos)[1]
```

```
## [1] 12726
```

***Fecha y hora:***

Convertiremos la fecha y hora del tipo de cadena de texto al de fecha y hora:

```
# Cargamos la librería Lubridate
library(lubridate)

# Convertimos a tipo fecha y hora
data_terremotos$Fecha.y.Hora <-
dmy_hms(substr(data_terremotos$Fecha.y.Hora, start = 1 , stop = 20) )
head(data_terremotos$Fecha.y.Hora)
```

```
## [1] "2010-12-30 16:35:33 UTC" "2010-12-21 20:04:47 UTC"
## [3] "2010-12-19 07:43:52 UTC" "2010-12-18 13:07:52 UTC"
## [5] "2010-12-17 22:14:26 UTC" "2010-12-17 20:34:24 UTC"
```

Creamos algunas columnas adicionales que serán útiles al momento de realizar los analisis.

```
# Número de mes
data_terremotos$Mes <- month(data_terremotos$Fecha.y.Hora)
# Hora
data_terremotos$Hora <- hour(data_terremotos$Fecha.y.Hora)
# Día de la semana
data_terremotos$Dia.Semana <- wday(data_terremotos$Fecha.y.Hora, label =
TRUE, abbr = FALSE)
```

### *Magnitud y Profundidad*

Tranformaremos los campos de Magnitud y Profundidad de tipo cadena de texto a valores numéricos:

```
# Convertimos a tipo numérico la Magnitud
data_terremotos$Magnitud <- as.numeric(data_terremotos$Magnitud)

# Quitamos los caracteres y convertimos a tipo numérico la Profundidad
data_terremotos$Profundidad <- gsub('Â', '', data_terremotos$Profundidad)
data_terremotos$Profundidad <- gsub('km', '',
data_terremotos$Profundidad)
data_terremotos$Profundidad <- gsub('.{1}$', '',
data_terremotos$Profundidad)
data_terremotos$Profundidad <- as.numeric(data_terremotos$Profundidad)
```

### *Zona de referencia*

Vamos a crear un campo llamado “Zona de referencia” que contenga la ubicación resumida del sismo, a partir del campo ‘Ubicacion’ que tiene una descripción más larga.

```
library(stringr)

# Funcion que busca una palabra a partir de cierta posicion en R.
# Autor: Hadley Wickham
# Fuente: https://github.com/tidyverse/stringr/blob/master/R/word.r
```

```r
word <- function(string, start = 1L, end = start, sep = fixed(" ")) {
  args <- vctrs::vec_recycle_common(string = string, start = start, end =
end)
  string <- args$string
  start <- args$start
  end <- args$end

  breaks <- str_locate_all(string, sep)
  words <- lapply(breaks, invert_match)

  # Convert negative values into actual positions
  len <- vapply(words, nrow, integer(1))

  neg_start <- !is.na(start) & start < 0L
  start[neg_start] <- start[neg_start] + len[neg_start] + 1L

  neg_end <- !is.na(end) & end < 0L
  end[neg_end] <- end[neg_end] + len[neg_end] + 1L

  # Replace indexes past end with NA
  start[start > len] <- NA
  end[end > len] <- NA

  # To return all words when trying to extract more words than available
  start[start < 1L] <- 1

  # Extract locations
  starts <- mapply(function(word, loc) word[loc, "start"], words, start)
  ends <-   mapply(function(word, loc) word[loc, "end"], words, end)

  str_sub(string, starts, ends)
}


# Verifica si el texto contiene alguno de los puntos cardinales.

# Función para punto cardinal

 existe_punto_cardinal <- function(x) {
  library(dplyr)
  case_when(
    tolower(word(word(x, start = 2, sep = fixed("km al ")), 1)) ==
"norte" ~ "norte",
    tolower(word(word(x, start = 2, sep = fixed("km al ")), 1)) ==
"noreste" ~  "noreste",
    tolower(word(word(x, start = 2, sep = fixed("km al ")), 1)) == "este"
~  "este",
    tolower(word(word(x, start = 2, sep = fixed("km al ")), 1)) ==
"sureste" ~  "sureste",
```

```
    tolower(word(word(x, start = 2, sep = fixed("km al ")), 1)) == "sur"
~  "sur",
    tolower(word(word(x, start = 2, sep = fixed("km al ")), 1)) ==
"suroeste" ~  "suroeste",
    tolower(word(word(x, start = 2, sep = fixed("km al ")), 1)) ==
"oeste" ~  "oeste",
    tolower(word(word(x, start = 2, sep = fixed("km al ")), 1)) ==
"noroeste" ~  "noroeste",
#    TRUE ~ x
    TRUE ~ data_terremotos$Ubicacion
  )
 }

 # Devuelve contenido de la Ubicacion del sismo  sin el texto que
antecede a los puntos cardinales. Sino, retornara la misma columna.

 zona_de_sismo2 <- function(x) {
  y <- word(data_terremotos$Ubicacion, start = 2, sep = fixed(paste(x,'
de ',sep="")))
  case_when(
    x == "norte" ~ y,
    x == "noreste" ~ y,
    x == "este" ~  y,
    x == "sureste" ~  y,
    x == "sur" ~  y,
    x == "suroeste" ~  y,
    x == "oeste" ~  y,
    x == "noroeste" ~  y,
    TRUE ~ data_terremotos$Ubicacion
  )
 }

# Variable que combina las dos funciones anteriores para limpiar la
columna Ubicacion.

data_terremotos_puntoCardinal <-
str_trim(existe_punto_cardinal(data_terremotos$Ubicacion))

# Analiza si la columna Ubicacion contiene alguno de los paises de
America (variable declara al inicio de este fichero). Retorna la limpieza
final de esta columna en una nueva columna que se agrega al dataframe en
cuestion.

validar_pais3 <- function(z) {
  r <- data.frame(z)
  s <- data.frame(paises_de_america)
  h <- ""
  delimitador <- ".*"
```

```
  for (a in 1:nrow(z)) {
    for (g in 1:nrow(s)) {
      if (grepl(paises_de_america$Paises[g], z$Zona.de.referencia[a])) {
        r$Zona.de.referencia[a] <- str_trim(
          paste((gsub(
            pattern = (paste((
              str_trim(s$Paises[g])
            ), delimitador, sep = "")), "", x = z$Zona.de.referencia[a]
          )), (str_trim(s$Paises[g])), sep = ""))
        break
      } else {
        r$Zona.de.referencia[a] <- str_trim(z$Zona.de.referencia[a])
      }
    }
  }
  return(r)
}

# Agrega la columna Zona de referencia con ayuda de la funcion zona de
sismo2.

data_terremotos$Zona.de.referencia <-
zona_de_sismo2(data_terremotos_puntoCardinal)

data_terremotos <- validar_pais3(data_terremotos)

head(data_terremotos)

##          Fecha.y.Hora Magnitud Profundidad
## 1 2010-12-30 16:35:33      4.1         109
## 2 2010-12-21 20:04:47      4.7         187
## 3 2010-12-19 07:43:52      4.5         108
## 4 2010-12-18 13:07:52      4.3         124
## 5 2010-12-17 22:14:26      4.8         568
## 6 2010-12-17 20:34:24      4.8          72
##
Ubicacion
## 1                       24 km al norte de Provincia de Nazca, Ica, Perú
## 2         16 km al sureste de Vilavila, Provincia de Lampa, Puno, Perú
## 3        46 km al suroeste de Yauri, Provincia de Espinar, Cusco, Perú
## 4 7.6 km al noroeste de Chuñune, Provincia de Caylloma, Arequipa, Perú
## 5                85 km al suroeste de Tarauaca, Estado de Acre, Brasil
## 6              30 km al noreste de Ilo, Departamento de Moquegua, Perú
##   Longitud Latitud  Año Pais Mes Hora Dia.Semana
## 1  -74.980 -14.620 2010 peru  12   16   Thursday
## 2  -70.587 -15.310 2010 peru  12   20    Tuesday
## 3  -71.732 -15.069 2010 peru  12    7     Sunday
## 4  -71.553 -15.284 2010 peru  12   13   Saturday
## 5  -71.460  -8.501 2010 peru  12   22     Friday
## 6  -71.103 -17.481 2010 peru  12   20     Friday
```

```
##                                Zona.de.referencia
## 1                   Provincia de Nazca, Ica, Perú
## 2       Vilavila, Provincia de Lampa, Puno, Perú
## 3       Yauri, Provincia de Espinar, Cusco, Perú
## 4 Chuñune, Provincia de Caylloma, Arequipa, Perú
## 5               Tarauaca, Estado de Acre, Brasil
## 6            Ilo, Departamento de Moquegua, Perú
```

**Seleccionaremos el dataset completo, ya que nuestro interés es analizar los terremotos de latino america entre los años 2010 hasta 2021 (05 abril). Por lo tanto es necesario considerar todos los regitros.**

---

# Limpieza

---

## Elementos vacios

Vamos a revisar los valores vacios del dataset:

```
# Estadísticas de valores vacíos
colSums(is.na(data_terremotos))
```

```
##       Fecha.y.Hora           Magnitud        Profundidad
Ubicacion
##                  0                   0                   0
0
##           Longitud            Latitud                Año
Pais
##                  0                   0                   0
0
##                Mes                Hora         Dia.Semana
Zona.de.referencia
##                  0                   0                   0
0
```

```
colSums(data_terremotos[c("Magnitud","Profundidad","Ubicacion","Longitud"
,"Latitud")]=="")
```

```
##    Magnitud Profundidad   Ubicacion    Longitud     Latitud
##           0           0           0           0           0
```

Podemos observar que no existen valores nulos, por lo que no será necesario realizar un tratamiento.

Luego, validamos que no existan otros valores que puedan hacer la función de vacios para las variables numéricas, como: cero, valores de tipo '9999' y negativos.

***Cero***

```
# Validamos los valores cero
colSums(data_terremotos[c("Magnitud","Profundidad","Longitud","Latitud")]
==0)

##    Magnitud Profundidad    Longitud     Latitud
##           0           0           0           0
```

Observamos que no existen valores iguales a cero.

***Valores negativos y valores de tipo '9999'*** Para ello calcularemos el rango de cada variable. Si alguna variable tiene estos valores en sus límites inferiores o superiores, analizaremos si son valores que reemplazan al vacio.

```
# Rango de la Magnitud
range(data_terremotos$Magnitud)

## [1] 4.0 4.9

# Rango de la Profundidad
range(data_terremotos$Profundidad)

## [1]   0.4 750.0

# Rango de la Longitud
range(data_terremotos$Longitud)

## [1] -117.70  -27.96

# Rango de la Latitud
range(data_terremotos$Latitud)

## [1] -51.30000  33.06617
```

Observamos que no existen valores del tipo '9999'. Las variables Longitud y Latitud admiten valores negativos, por lo que no las consideraremos como valores que reemplazan al vacio.

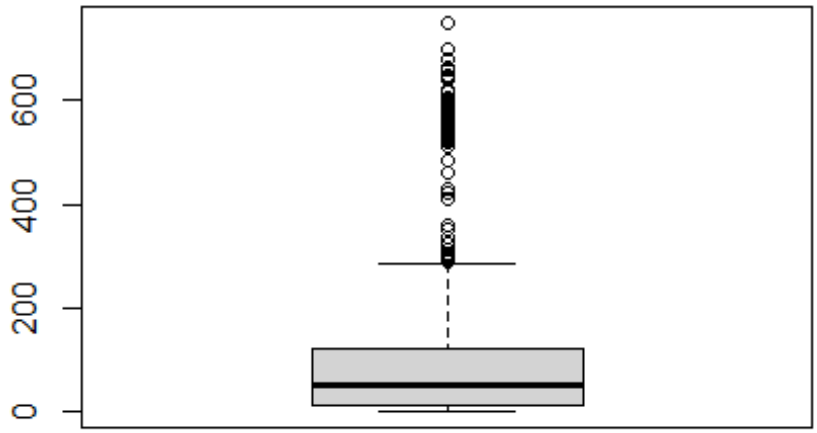
**Valores negativos**

## Valores extremos

Primero, realizaremos un analisis de los valores extremos para la Profundidad y la Magnitud de manera individual a través de boxplots.

Analizaremos la Profundidad.

```
# Calcular el boxplot de la Profundidad
Profundidad.bp <- boxplot(data_terremotos$Profundidad)
```

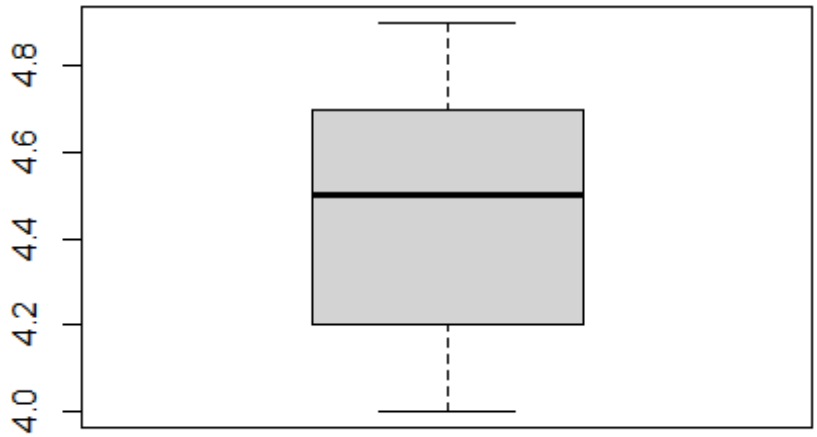



```
# Calculamos el rango de los outliers
range(Profundidad.bp$out)

## [1] 288 750

# Calcular la cantidad de los outliers
length(Profundidad.bp$out)

## [1] 161
```

Podemos observar que existen 159 valores outliers de la variable Profundidad que **tienen valores entre 288 a 750 km.**`

Ahora, analizaremos la Magnitud.

```
# Calcular el boxplot de la Magnitud
Magnitud.bp <- boxplot(data_terremotos$Magnitud)
```



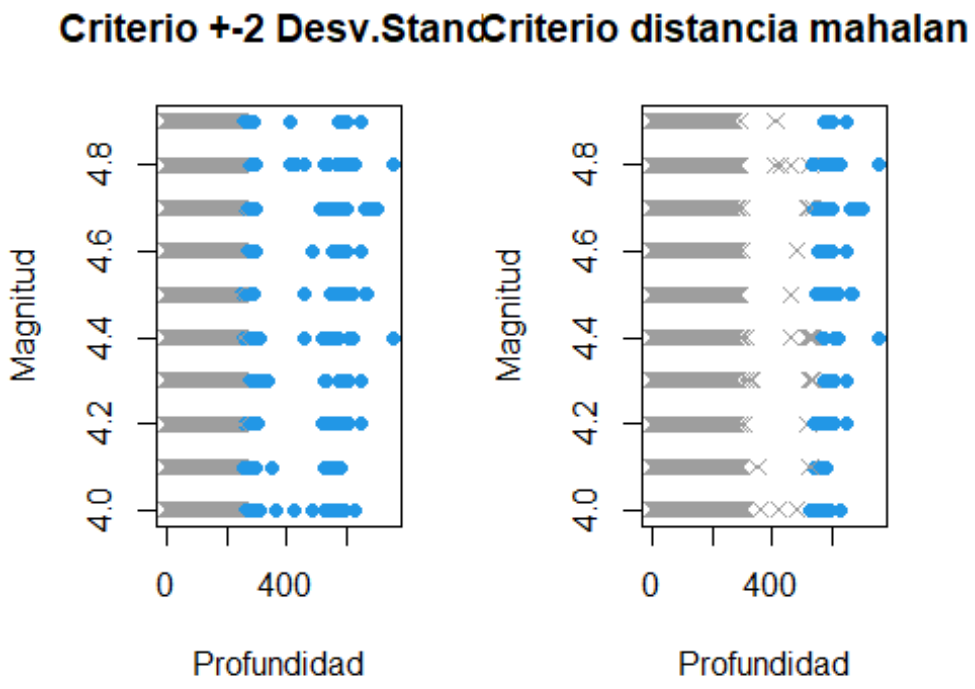

```
# Calculamos el rango de los outliers
range(Magnitud.bp$out)

## [1]  Inf -Inf

# Calcular la cantidad de los outliers
length(Magnitud.bp$out)

## [1] 0
```

Observamos que o existen outliers de la variable Magnitud.

También, analizaremos las variables Profundidad y Magnitud utilizando la **distancia de Mahalanobis** para encontrar 100 outliers.

```
#Criterio +/-2SD
ap <- data_terremotos[,c("Profundidad","Magnitud")]
Profundidad.outlier <- abs(scale(data_terremotos$Profundidad)) > 2
Magnitud.outlier <- abs(scale(data_terremotos$Magnitud)) > 2
pch <- (Profundidad.outlier | Magnitud.outlier) * 12
pch <- pch + 4
col <- (Profundidad.outlier | Magnitud.outlier) * 12
col <- pch + 4

par(mfrow=c(1,2))
plot(ap, pch=pch,col=col,main="Criterio +-2 Desv.Stand.")

#Criterio distancia Mahalanobis (los dos outliers más extremos)
```

```r
n.outliers <- 100
m.dist.order <- order(mahalanobis(ap, colMeans(ap), cov(ap)),
decreasing=TRUE)
is.outlier <- rep(FALSE, nrow(ap))
is.outlier[m.dist.order[1:n.outliers]] <- TRUE
pch <- is.outlier * 12
pch <- pch + 4
col <- is.outlier * 12
col <- pch + 4
plot(ap, pch=pch,col=col,main="Criterio distancia mahalanobis")
```




Podemos observar que los outliers, según el criterio de mahalanobis, son puntos con altos valores de Profundidad; al igual que con el criterio de -+2 desviaciones estandar. Esto se debe a la manera en que están distribuidos los datos, donde se observa que la Profundidad es la variable con mayor dispersión.

Hallaremos el rango de los 100 puntos extremos calculados:

```r
# Rango de los outliers
range(ap[is.outlier,]$Profundidad)
```

```
## [1] 523 750
```

Finalmente, concluimos que existen 159 outliers con **Profundidades mayores a 288 km (criterio +-2Desv.).** De los cuales los 100 outliers más extremos tienen **Profundidades mayores a 523 km (Criterio mahalanobis).**

Al ser datos reales de sismos, no haremos ningún tratamiento adicional a estos valores.

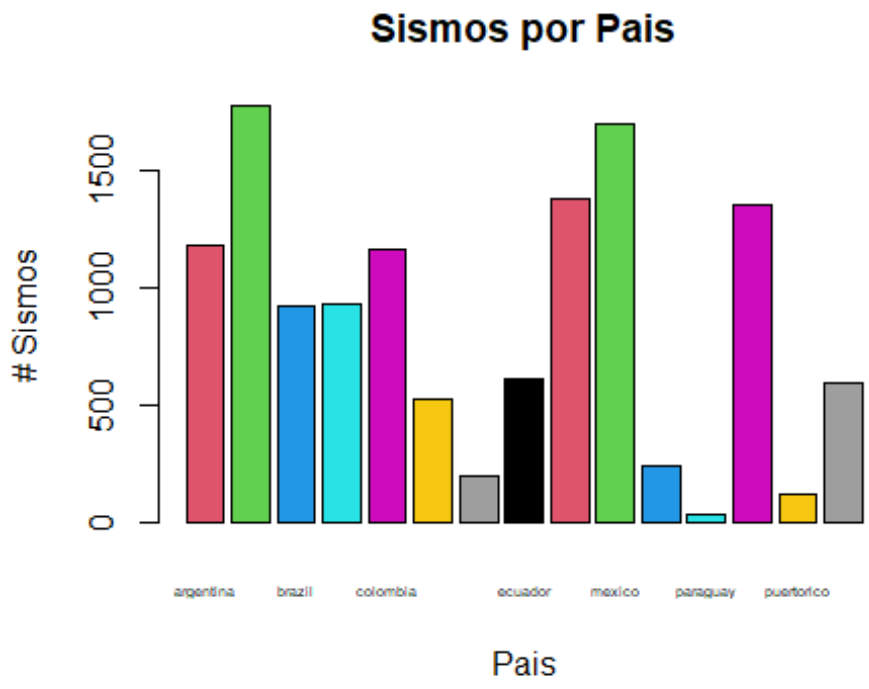
---

## Datos preprocesados

---

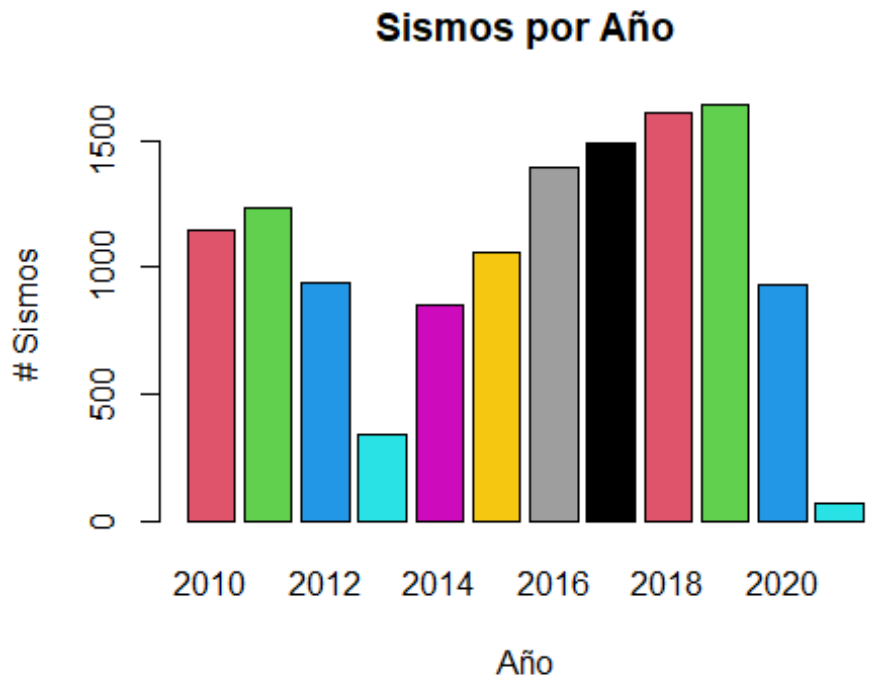
Presentamos como resumen las distribuciones de las variables discretas y continuas. Primero, presentamos las distribuciones de las variables discretas **(Pais, Año, Mes, Hora y Dia.Semana)** a través de gráficos de barra:

```
# Grafico de barras por pais
barplot(table(data_terremotos$Pais),
        legend = FALSE,
        col = 2:20,
        main = "Sismos por Pais",
        xlab="Pais",
        ylab="# Sismos",
        cex.names=0.45)
```

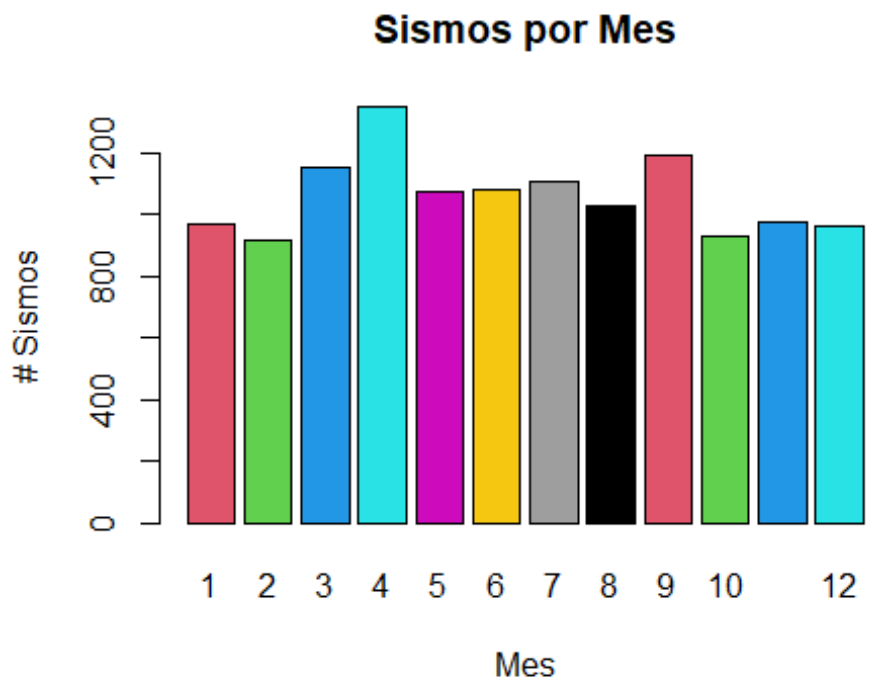



```
# Grafico de barras por Año
barplot(table(data_terremotos$Año),
        legend = FALSE,
        col = 2:20,
        main = "Sismos por Año",
```

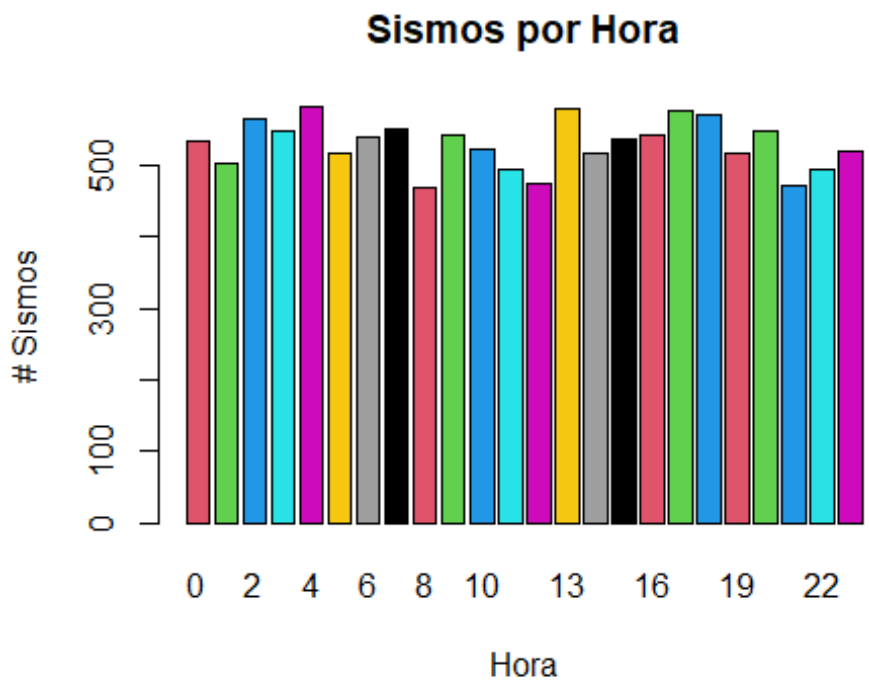
```
        xlab="Año",
        ylab="# Sismos")

# Grafico de barras por Año
barplot(table(data_terremotos$Año),
        legend = FALSE,
        col = 2:20,
        main = "Sismos por Año",
        xlab="Año",
        ylab="# Sismos")
```



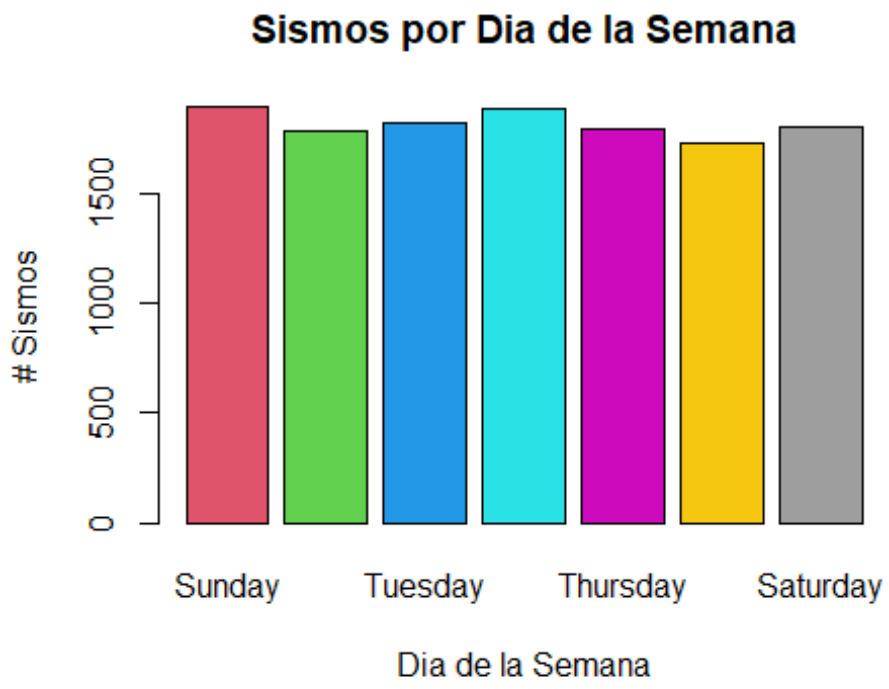

```
# Grafico de barras por Mes
barplot(table(data_terremotos$Mes),
        legend = FALSE,
        col = 2:20,
        main = "Sismos por Mes",
        xlab="Mes",
        ylab="# Sismos")
```




```r
# Grafico de barras por Hora
barplot(table(data_terremotos$Hora),
        legend = FALSE,
        col = 2:20,
        main = "Sismos por Hora",
        xlab="Hora",
        ylab="# Sismos")
```

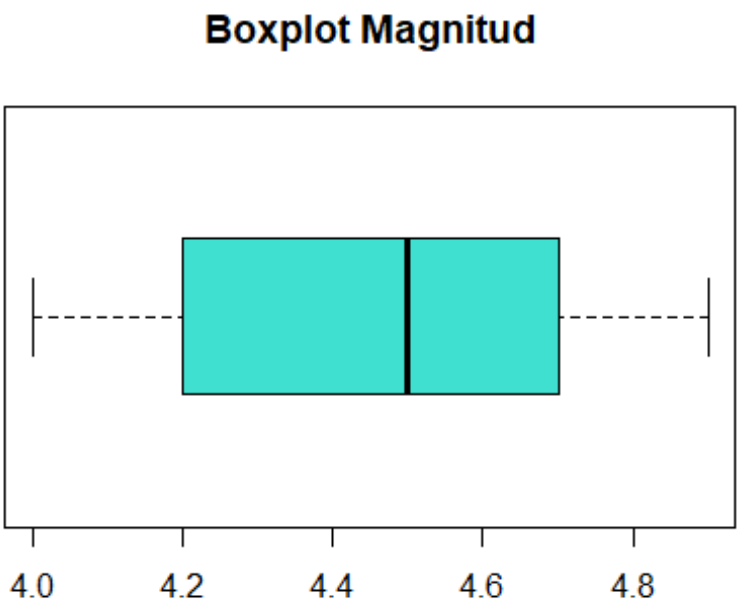



```r
# Grafico de barras por Dia.Semana
barplot(table(data_terremotos$Dia.Semana),
        legend = FALSE,
        col = 2:20,
        main = "Sismos por Dia de la Semana",
        xlab="Dia de la Semana",
        ylab="# Sismos")
```

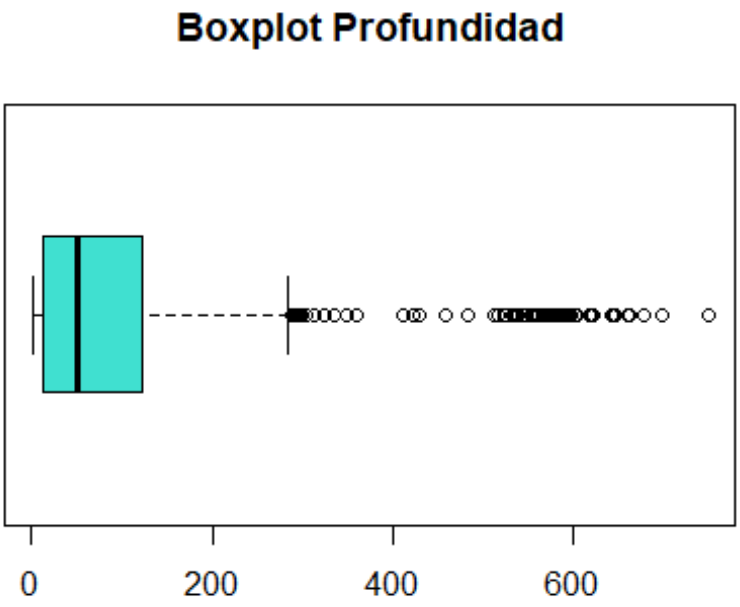



Luego, presentamos las distribuciones de las variables continuas **(Pais, Año, Mes, Hora y Dia.Semana)** a través de diagramas de cajas:

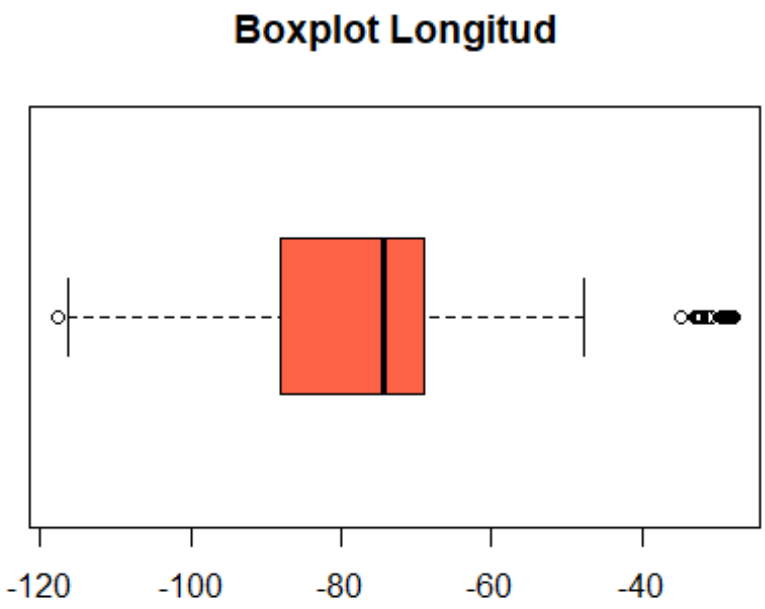
```
# Diagrama de cajas de Magnitud
boxplot(data_terremotos$Magnitud, horizontal = TRUE,
        col=c("turquoise"),
        main="Boxplot Magnitud")
```



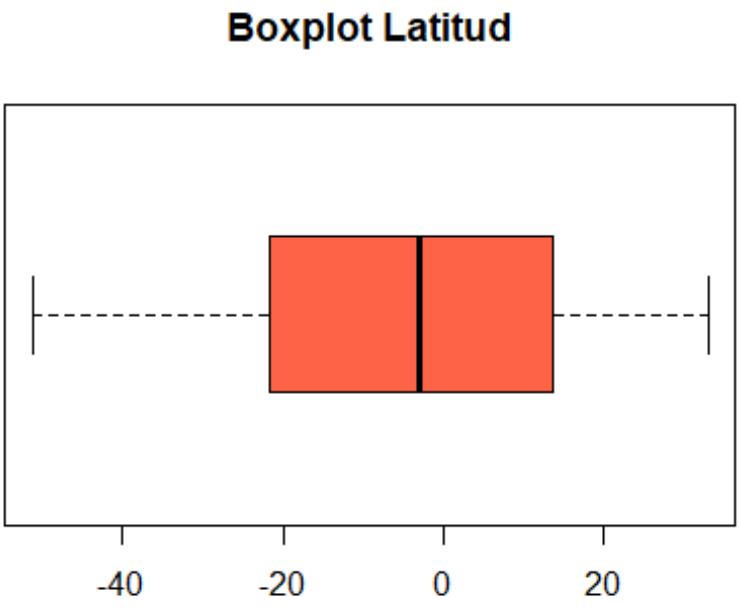

```r
# Diagrama de cajas de Profundidad
boxplot(data_terremotos$Profundidad, horizontal = TRUE,
        col=c("turquoise"),
        main="Boxplot Profundidad")
```




```r
# Diagrama de cajas de Longitud
boxplot(data_terremotos$Longitud, horizontal = TRUE,
        col=c("tomato"),
        main="Boxplot Longitud")
```




```r
# Diagrama de cajas de Latitud
boxplot(data_terremotos$Latitud, horizontal = TRUE,
        col=c("tomato"),
        main="Boxplot Latitud")
```




---

## Analisis de datos

---

### Planificación

Con el objetivo de responder las 3 preguntas especificas planteadas en el *apartado 1. Descripción*, se realizarán tres analisis estadisticos:

- **Contraste de hipotesis**, para determinar qué paises se dan los sismos de mayor Magnitud/Profundidad y con qué significancia.
- Un modelo de **regresion logísica**, donde se modelará la probabilidad de que ocurra un sismo de acuerdo a las varibles del dataset. Nos ayudará a determinar que variables son significativas para la ocurrencia de un sismo.
- Analisis de **correlacion**, para determinar si existe una correlación entre: 1) magnitud y profundidad, 2) longitud y magnitud, 3) latitud y magnitud, 4) longitud y profundidad y 5) latitud y profundidad.

Por lo tanto, para el contraste de hipotesis los grupos a comparar serán los paises. Y para la regresión y la correlación el analisis se realizará al conjunto de datos en su totalidad.

## Normalidad y Homocedasticidad

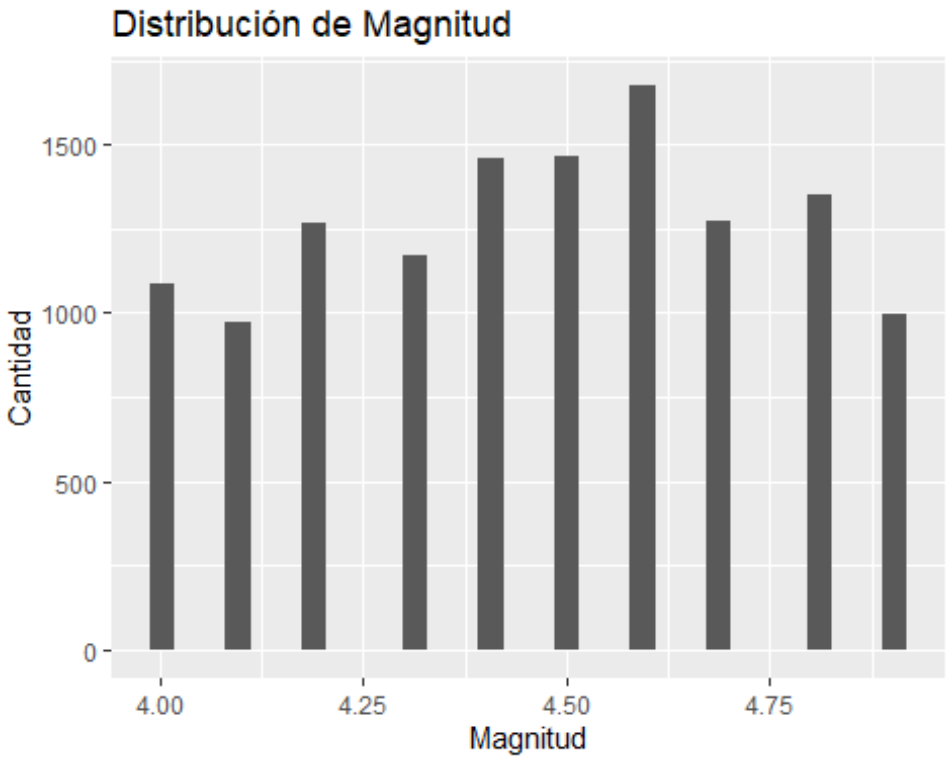
Antes de realizar las pruebas estadisticas, vamos a comprobar la Normalidad y Homocedasticidad de los datos.

### Normalidad

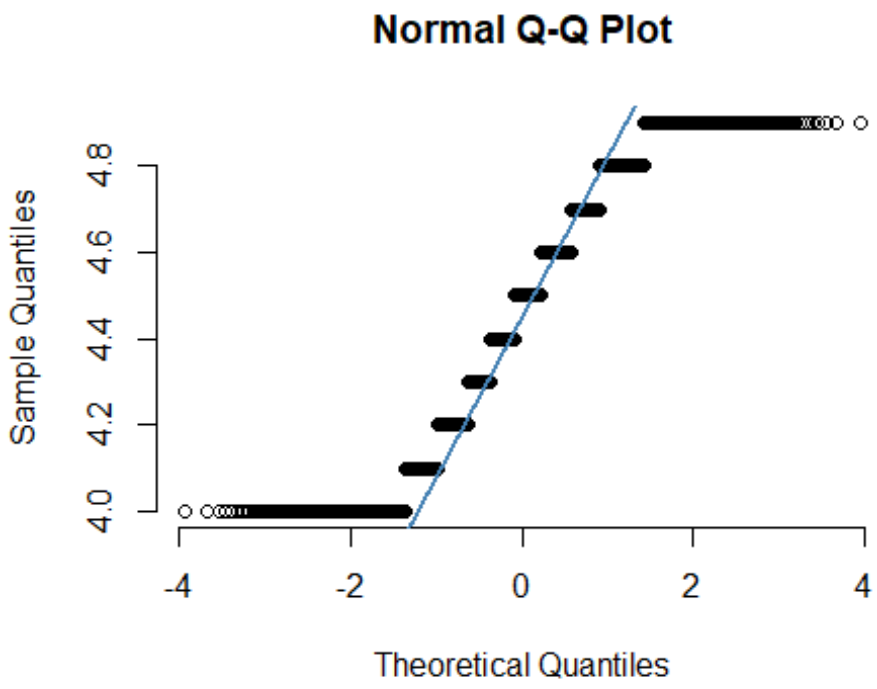
Primero, comprobaremos la ***normalidad*** de los datos. Para ello graficaremos la distribución, construiremos gráficas **QQplots** y aplicaremos el test de **Kolmogorov-Smirnov** para las variables **Magnitud, Profunidad, Longitud y Latitud.** No optamos usar el test Shapiro-wilk porque el tamaño de los registros superan los 5000.

***Magnitud***

```
# Importamos la librería ggplot2
library(ggplot2)

#Distribución de la Magnitud
ggplot(data = data_terremotos, aes (x=data_terremotos$Magnitud)) +
geom_histogram() +
ggtitle("Distribución de Magnitud") +
xlab("Magnitud") +
ylab("Cantidad")
```




```
#Gráfica QQplots
qqnorm(data_terremotos$Magnitud, pch = 1, frame = FALSE)
qqline(data_terremotos$Magnitud, col = "steelblue", lwd = 2)
```



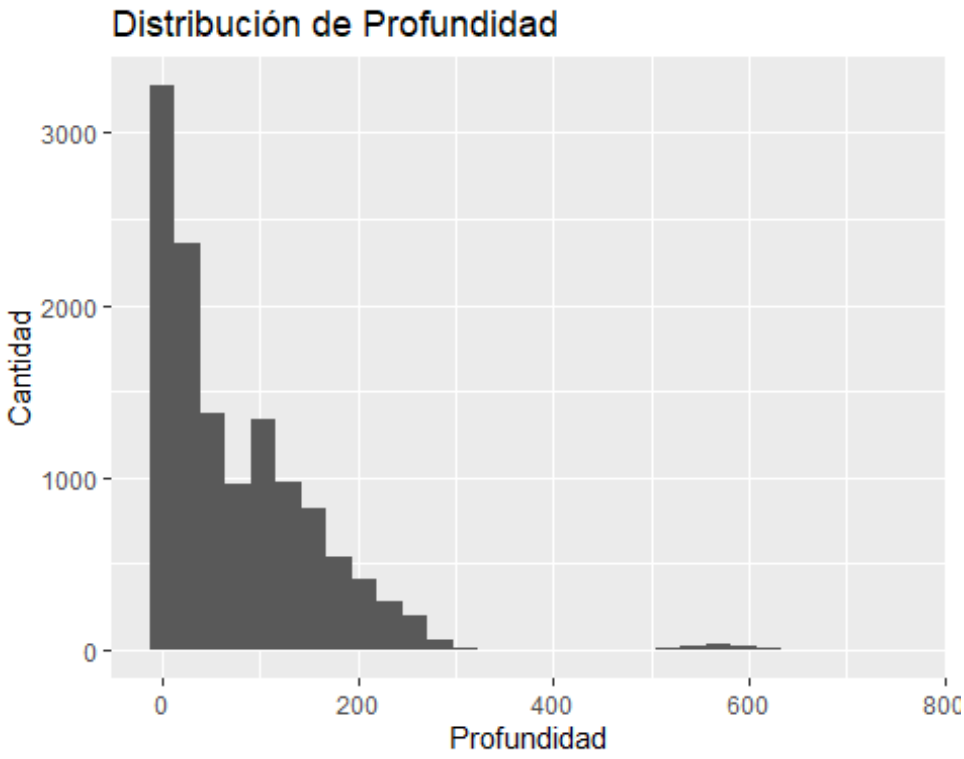

```
# Test Kolmogorov-Smirnov aplicado a la Magnitud
ks.test(data_terremotos$Magnitud, pnorm, mean(data_terremotos$Magnitud),
sd(data_terremotos$Magnitud))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_terremotos$Magnitud
## D = 0.10931, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

De los gráficos podemos observar que la magnitud tiene datos centrales muy cercanos a la distribución normal, sin embargo por sus sus valores extremos la Magnitud no tiene la distribución normal.

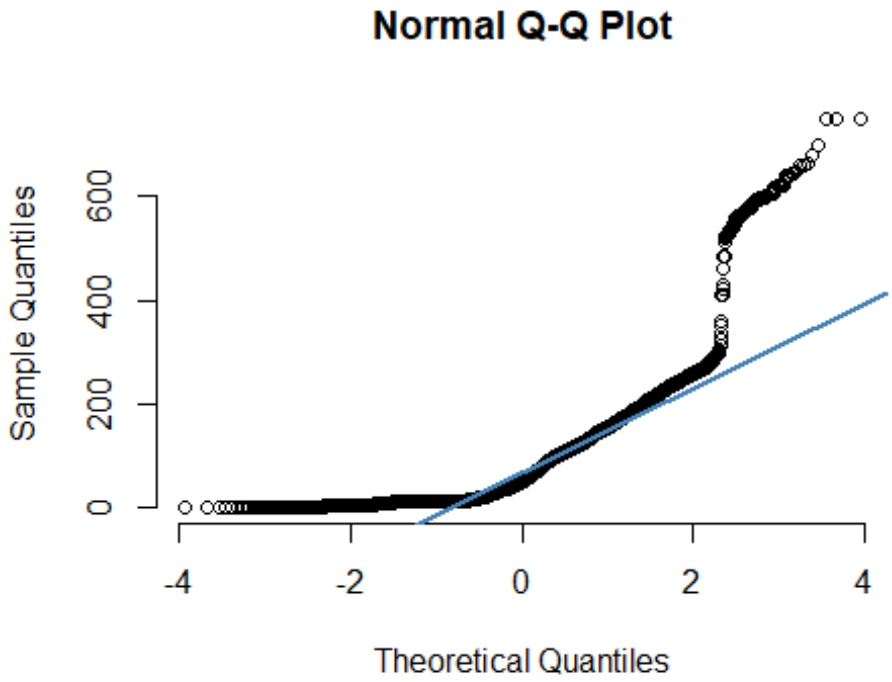
Del test observamos que el p-value es menor a 0.05 por lo que se rechaza la hipotesis nula de normalidad. Se acepta la hipotesis alternativa, que la **Magnitud** no sigue una distribución Normal.

***Profundidad***

```
#Distribución de la Profundidad
ggplot(data = data_terremotos, aes (x=data_terremotos$Profundidad)) +
geom_histogram() +
ggtitle("Distribución de Profundidad") +
xlab("Profundidad") +
ylab("Cantidad")
```




```r
#Gráfica QQplots
qqnorm(data_terremotos$Profundidad, pch = 1, frame = FALSE)
qqline(data_terremotos$Profundidad, col = "steelblue", lwd = 2)
```



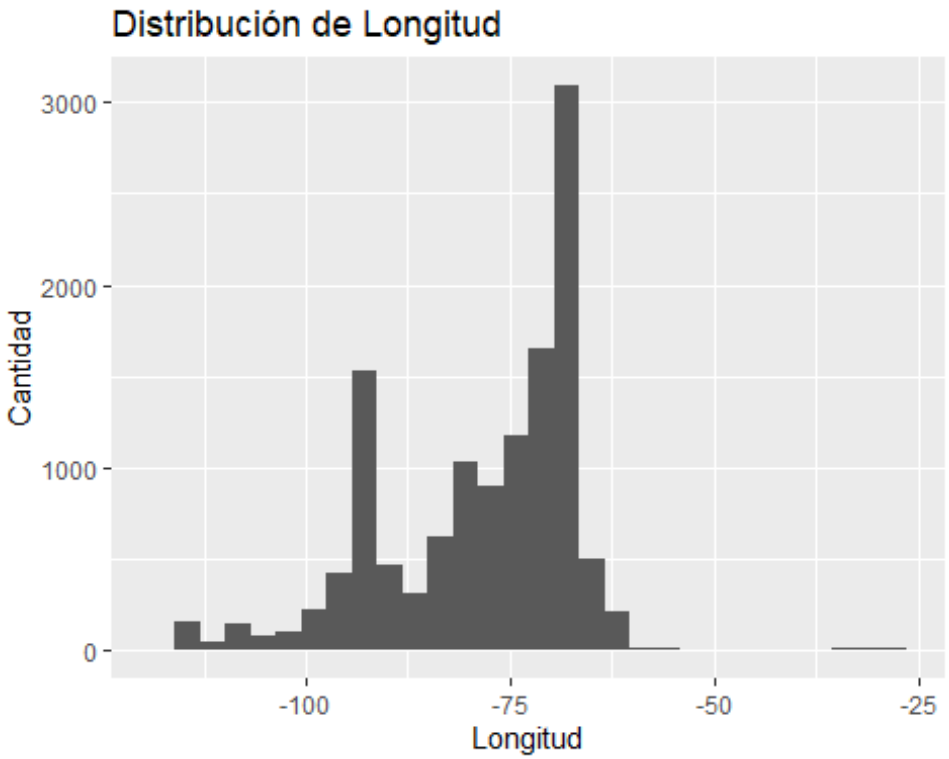
```
# Test Kolmogorov-Smirnov aplicado a la Profundidad
ks.test(data_terremotos$Profundidad, pnorm,
mean(data_terremotos$Profundidad), sd(data_terremotos$Profundidad))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_terremotos$Profundidad
## D = 0.17688, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

De los gráficos podemos observar que la profundidad es unimodal pero con cola a la derecha. Por lo que visualmente la profundidad no sigue una distribución normal.

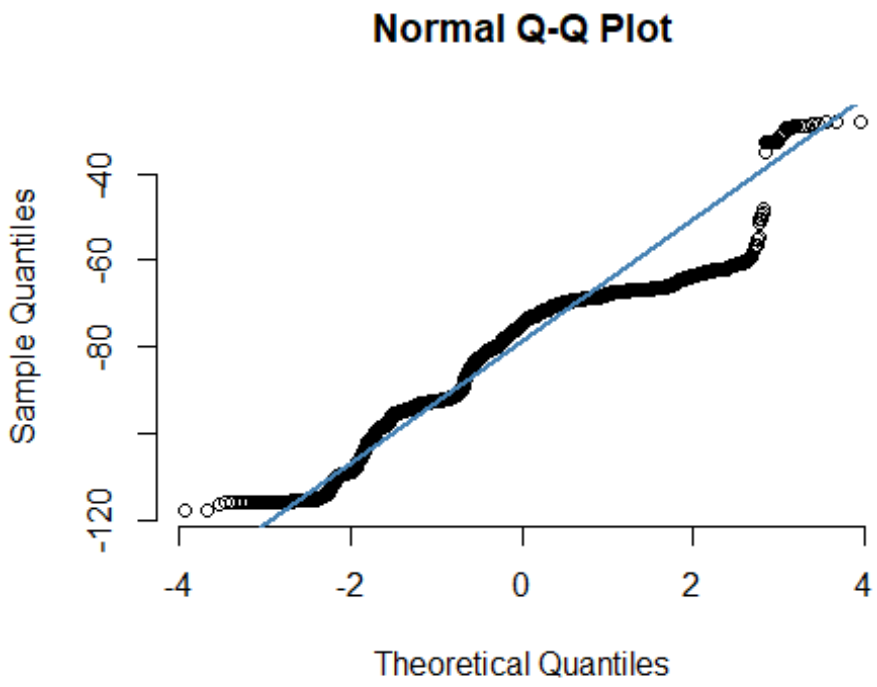
Del test observamos que el p-value es menor a 0.05 por lo que se rechaza la hipotesis nula de normalidad. Se acepta la hipotesis alternativa, que la **Profundidad** no sigue una distribución Normal.

***Longitud***

```
#Distribución de la Longitud
ggplot(data = data_terremotos, aes (x=data_terremotos$Longitud)) +
geom_histogram() +
ggtitle("Distribución de Longitud") +
xlab("Longitud") +
ylab("Cantidad")
```



```
#Gráfica QQplots
qqnorm(data_terremotos$Longitud, pch = 1, frame = FALSE)
qqline(data_terremotos$Longitud, col = "steelblue", lwd = 2)
```




```
# Test Kolmogorov-Smirnov aplicado a la Longitud
ks.test(data_terremotos$Longitud, pnorm, mean(data_terremotos$Longitud),
sd(data_terremotos$Longitud))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_terremotos$Longitud
## D = 0.14398, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

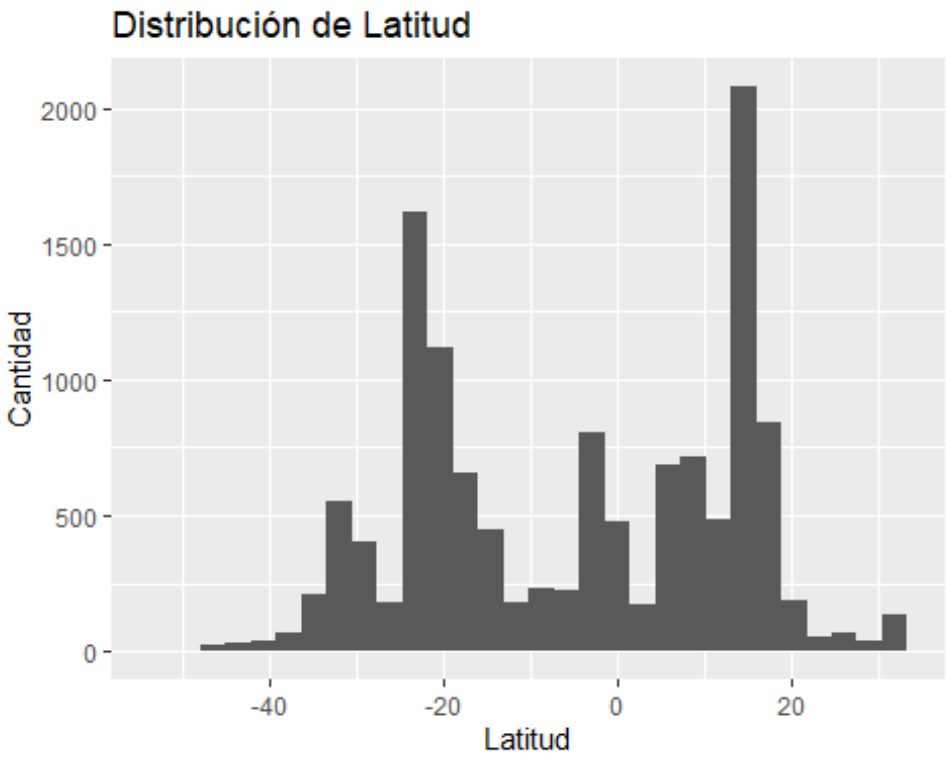
De los gráficos podemos observar que la longitud es bimodal y sus valores más altos hacen que la variable se aleje de la distribución normal.

Del test observamos que el p-value es menor a 0.05 por lo que se rechaza la hipotesis nula de normalidad. Se acepta la hipotesis alternativa, que la **Longitud** no sigue una distribución Normal.

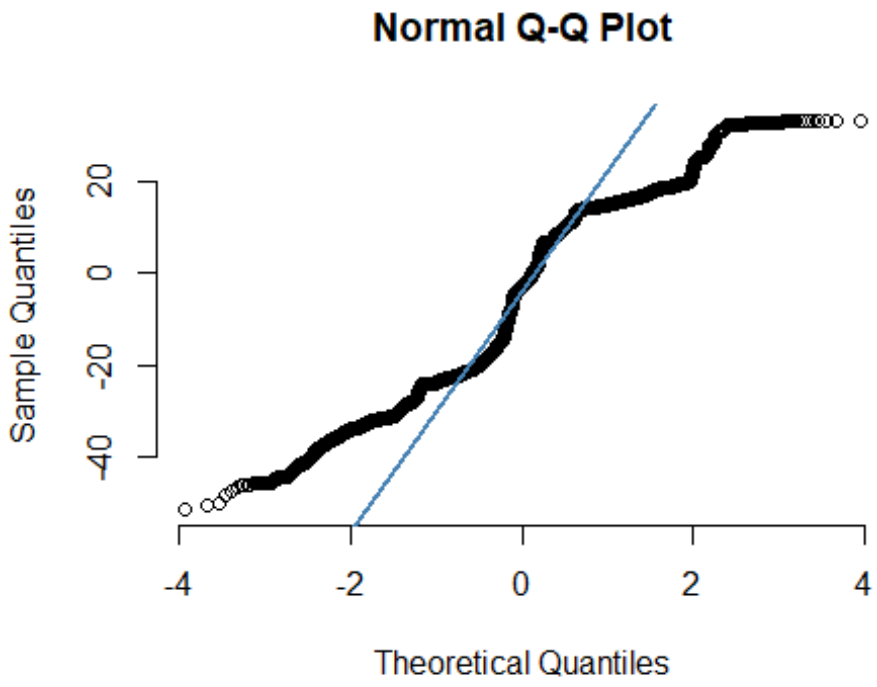
***Latitud***

```
#Distribución de la Latitud
ggplot(data = data_terremotos, aes (x=data_terremotos$Latitud)) +
geom_histogram() +
ggtitle("Distribución de Latitud") +
```

```
xlab("Latitud") +
ylab("Cantidad")
```




```
#Gráfica QQplots
qqnorm(data_terremotos$Latitud, pch = 1, frame = FALSE)
qqline(data_terremotos$Latitud, col = "steelblue", lwd = 2)
```




```
# Test Kolmogorov-Smirnov aplicado a la Latitud
ks.test(data_terremotos$Latitud, pnorm, mean(data_terremotos$Latitud),
sd(data_terremotos$Latitud))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_terremotos$Latitud
## D = 0.13574, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

De los gráficos podemos observar que la latitud es tiene muchos picos y es bastante dispersa, por lo que visualmente podemos ver que no se aproxima a una distribución normal.

Del test observamos que el p-value es menor a 0.05 por lo que se rechaza la hipotesis nula de normalidad. Se acepta la hipotesis alternativa, que la **Latitud** no sigue una distribución Normal.

Por lo tanto, concluimos que ***la Magnitud, la Profundidad, la longitud y la Latitud no siguen una distribución normal.***

### Homocedasticidad

Ahora, comprobaremos la ***homocedasticidad*** (homogeneidad de la varianza). Para ello utilizaremos el test **Fligner-Killeen,** utilizado para datos que no cumplen con la

condición de normalidad, para las variables **Magnitud, Profunidad, Longitud y Latitud** según los grupos de **Paises.**

***Magnitud***

```
# Test Fligner-Killeen aplicado a la Magnitud para los grupos de paises.
fligner.test(Magnitud ~ Pais, data = data_terremotos)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Magnitud by Pais
## Fligner-Killeen:med chi-squared = 1281.3, df = 14, p-value < 2.2e-16
```

El p-value es menor a 0.05, por lo que se rechaza la hipotesis nula. Se acepta la hipotesis alternativa, que **la Magnitud presenta varianzas estadisticamente diferentes para los distintos grupos de Paises.**

***Profundidad***

```
# Test Fligner-Killeen aplicado a la Profundidad para los grupos de
paises.
fligner.test(Profundidad ~ Pais, data = data_terremotos)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Profundidad by Pais
## Fligner-Killeen:med chi-squared = 1646.4, df = 14, p-value < 2.2e-16
```

El p-value es menor a 0.05, por lo que se rechaza la hipotesis nula. Se acepta la hipotesis alternativa, que **la Profundidad presenta varianzas estadisticamente diferentes para los distintos grupos de Paises.**

***Longitud***

```
# Test Fligner-Killeen aplicado a la Longitud para los grupos de paises.
fligner.test(Longitud ~ Pais, data = data_terremotos)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Longitud by Pais
## Fligner-Killeen:med chi-squared = 2786.2, df = 14, p-value < 2.2e-16
```

El p-value es menor a 0.05, por lo que se rechaza la hipotesis nula. Se acepta la hipotesis alternativa, que **la Longitud presenta varianzas estadisticamente diferentes para los distintos grupos de Paises.**

***Latitud***

```
# Test Fligner-Killeen aplicado a la Latitud para los grupos de paises.
fligner.test(Latitud ~ Pais, data = data_terremotos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Latitud by Pais
## Fligner-Killeen:med chi-squared = 4295.8, df = 14, p-value < 2.2e-16
```

El p-value es menor a 0.05, por lo que se rechaza la hipotesis nula. Se acepta la hipotesis alternativa, que **la Latitud presenta varianzas estadisticamente diferentes para los distintos grupos de Paises.**

Por lo tanto, concluimos que ***la Magnitud, la Profundidad, la Longitud y la Latitud presentan varianzas diferentes para los Paises.***

## Pruebas estadísticas

En este apartado realizaremos los 3 análisis descritos en la sección *4.1 Planificación.*

### Contraste de Hipotesis

Realizaremos una analisis de contraste de hipotesis para saber qué pais tiene una magnitud y profundidad estadisticamente mayor a los demás.

Mostramos las medias de Magnitud y Profundidad de los paises:

```
library(plyr)
# Contruimos una tabla con la media y varianza de la Magnitud de los
Paises
# Calculamos la media
t_mean <- aggregate(data_terremotos$Magnitud, list(data_terremotos$Pais)
, mean)
colnames(t_mean) <- c("Pais","Media")
# Calculamos la varianza
t_var <- aggregate(data_terremotos$Magnitud, list(data_terremotos$Pais) ,
var)
colnames(t_var) <- c("Pais","Varianza")
# Calculamos la cantidad de sismos
t_count <- count(data_terremotos,"Pais")
colnames(t_count) <- c("Pais","Cantidad")
# Contruimos la tabla y la ordenamos
table_pais_mag <- merge(t_mean,t_var, by = "Pais")
table_pais_mag <- merge(table_pais_mag,t_count, by = "Pais")
table_pais_mag<- table_pais_mag[order(table_pais_mag$Media,decreasing =
TRUE),]
table_pais_mag
```

```
##                  Pais    Media   Varianza Cantidad
## 3              brazil 4.763459 0.01543824      925
## 4               chile 4.744635 0.01272092      932
```

```
## 10             mexico 4.521824 0.03517384     1700
## 12           paraguay 4.517241 0.04576355       29
## 13               peru 4.506282 0.05452411     1353
## 11             panama 4.423651 0.07539661      241
## 5            colombia 4.416309 0.07208774     1165
## 6           costarica 4.403263 0.06677779      521
## 8             ecuador 4.402303 0.07551363      608
## 1           argentina 4.388354 0.07155346     1185
## 2             bolivia 4.370883 0.06819556     1779
## 15          venezuela 4.370169 0.07143461      590
## 7  dominican-republic 4.341237 0.06678810      194
## 14         puertorico 4.321186 0.07142764      118
## 9           guatemala 4.304906 0.05970154     1386
```

```
# Contruimos una tabla con la media y varianza de la Profundidad de los
Paises
# Calculamos la media
t_mean <- aggregate(data_terremotos$Profundidad,
list(data_terremotos$Pais) , mean)
colnames(t_mean) <- c("Pais","Media")
# Calculamos la varianza
t_var <- aggregate(data_terremotos$Profundidad,
list(data_terremotos$Pais) , var)
colnames(t_var) <- c("Pais","Varianza")
# Calculamos la cantidad de sismos
t_count <- count(data_terremotos,"Pais")
colnames(t_count) <- c("Pais","Cantidad")
# Contruimos la tabla y la ordenamos
table_pais_pro <- merge(t_mean,t_var, by = "Pais")
table_pais_pro <- merge(table_pais_pro,t_count, by = "Pais")
table_pais_pro <- table_pais_pro[order(table_pais_pro$Media,decreasing =
TRUE),]
table_pais_pro
```

```
##                  Pais     Media   Varianza Cantidad
## 12           paraguay 497.51724 41545.1872       29
## 1           argentina 159.44920 12402.4957     1185
## 2             bolivia 132.47499  7247.4546     1779
## 3              brazil  91.35914  8165.0509      925
## 15          venezuela  82.94932  4837.1711      590
## 13               peru  76.13651  6725.8860     1353
## 4               chile  73.87157  4593.6480      932
## 9           guatemala  56.74784  2501.0205     1386
## 5            colombia  53.35536  3597.0331     1165
## 10             mexico  42.06835  2338.6816     1700
## 8             ecuador  41.10674  2147.2249      608
## 14         puertorico  37.31356  1386.1484      118
## 7  dominican-republic  32.32629  1711.9406      194
## 6           costarica  30.33858  1194.1111      521
## 11             panama  20.75726   212.0074      241
```

En el caso de Paraguay, decidimos no considerarla por la baja cantidad de sismos (30) y alta varianza. Podemos observar que los 3 paises con mayor magnitud son: Brazil, Chile y Mexico. Pero todos los paises tienen valores aparentemente parecidos. También, observamos que los 3 paises con mayor profundidad de sismo son: Argentina, Bolivia y Brazil.

A continuación, realizaremos el test Kruskal-Wallis (test no paramétrico) para comprobar que las medias de la Magnitud y la Profundidad sean iguales en todos los paises. Comenzaremos con la Magnitud:

```
# Test Kruskal para la Magnitud según los Paises. Sin considerar a
Paraguay
kruskal.test(Magnitud ~ Pais, data =
data_terremotos[data_terremotos$Pais!='paraguay',])

##
##  Kruskal-Wallis rank sum test
##
## data:  Magnitud by Pais
## Kruskal-Wallis chi-squared = 3331.7, df = 13, p-value < 2.2e-16
```

Dado que el p-value es menor a la 0.05 por lo que rechazamos la hipotesis nula y aceptamos la hipotesis alternativa, que es que la Magnitud muestra diferencias significativas para los diferentes paises. Ahora realizaremos el test a la Profundidad:

```
# Test Kruskal para la Progundidad los Paises. Sin considerar a Paraguay
kruskal.test(Profundidad ~ Pais, data =
data_terremotos[data_terremotos$Pais!='paraguay',])

##
##  Kruskal-Wallis rank sum test
##
## data:  Profundidad by Pais
## Kruskal-Wallis chi-squared = 2994.8, df = 13, p-value < 2.2e-16
```

Dado que el p-value es menor a la 0.05 por lo que rechazamos la hipotesis nula y aceptamos la hipotesis alternativa, que es que la Profundida muestra diferencias significativas para los diferentes paises.

Por lo tanto, concluimos que **los paises no tienen la misma Magnitud y la Profundida estadisticamente.** Una vez comprobada las diferencias de las medidas, comprobaremos si los paises en los primeros lugares son estadisticamente superiores a los demás.

***Primero, analizaremos la Magnitud.*** Estamos frente a un constraste de dos muestras independientes con varianzas desconocidas diferentes; además por el teorema de límite central (muestras mayores a 30) se puede asumir normalidad. Por lo tanto podemos utilizar el t-test para realizar este contraste de hipotesis.

Validaremos si la Magnitud de Brazil (1°) es superior a Chile (2°). Las hipotesis son: H0: La Magnitud media de Brazil y Chile son iguales. H1: La Magnitud media de Brazil es mayor a la de Chile.

Aplicando el t-test:

```
t.test(data_terremotos$Magnitud[data_terremotos$Pais=='brazil'],data_terr
emotos$Magnitud[data_terremotos$Pais=='chile'],alternative="greater",
var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data:  data_terremotos$Magnitud[data_terremotos$Pais == "brazil"] and
data_terremotos$Magnitud[data_terremotos$Pais == "chile"]
## t = 3.4176, df = 1835.1, p-value = 0.0003228
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.009759691         Inf
## sample estimates:
## mean of x mean of y
##  4.763459  4.744635
```

El p-value es menor al una significancia de 0.01. Por lo que se rechaza la hipotesis nula y se acepta la alternativa. Es decir aceptamos que Brazil tiene una Magnitud media mayor que Chile, con un nivel de confianza de 99%.

Validaremos si la Magnitud de Brazil (1°) es superior a México (3°). Las hipotesis son: H0: La Magnitud media de Brazil y México son iguales. H1: La Magnitud media de Brazil es mayor a la de México

Aplicando el t-test:

```
t.test(data_terremotos$Magnitud[data_terremotos$Pais=='brazil'],data_terr
emotos$Magnitud[data_terremotos$Pais=='mexico'],alternative="greater",
var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data:  data_terremotos$Magnitud[data_terremotos$Pais == "brazil"] and
data_terremotos$Magnitud[data_terremotos$Pais == "mexico"]
## t = 39.522, df = 2524.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2315757       Inf
## sample estimates:
## mean of x mean of y
##  4.763459  4.521824
```

El p-value es menor al una significancia de 0.01. Por lo que se rechaza la hipotesis nula y se acepta la alternativa. Es decir aceptamos que Brazil tiene una Magnitud media mayor que México, con un nivel de confianza de 99%.

**Con esto validamos que Brazil es el pais con mayor magnitud sismica de latino america**

***Segundo, analizaremos la Profundidad*** Estamos frente a un constraste de dos muestras independientes con varianzas desconocidas diferentes; además por el teorema de límite central (muestras mayores a 30) se puede asumir normalidad. Por lo tanto podemos utilizar el t-test para realizar este contraste de hipotesis.

Validaremos si la Profundidad de Argentina (1°) es superior a Bolivia (2°). Las hipotesis son: H0: La Profundidad media de Argentina y Bolivia son iguales. H1: La Profundidad media de Argentina es mayor a la de Bolivia

Aplicando el t-test:

```
t.test(data_terremotos$Profundidad[data_terremotos$Pais=='argentina'],dat
a_terremotos$Profundidad[data_terremotos$Pais=='bolivia'],alternative="gr
eater", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_terremotos$Profundidad[data_terremotos$Pais ==
"argentina"] and data_terremotos$Profundidad[data_terremotos$Pais ==
"bolivia"]
## t = 7.074, df = 2075.7, p-value = 1.025e-12
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  20.69933      Inf
## sample estimates:
## mean of x mean of y
##  159.4492  132.4750
```

El p-value es menor al una significancia de 0.01. Por lo que se rechaza la hipotesis nula y se acepta la alternativa. Es decir aceptamos que Argentina tiene una Profundidad media mayor que Bolivia, con un nivel de confianza de 99%.

Validaremos si la Profundidad de Argentina (1°) es superior a Brazil (3°). Las hipotesis son: H0: La Magnitud media de Argentina y Brazil son iguales. H1: La Magnitud media de Argentina es mayor a la de Brazil

Aplicando el t-test:

```
t.test(data_terremotos$Profundidad[data_terremotos$Pais=='argentina'],dat
a_terremotos$Profundidad[data_terremotos$Pais=='brazil'],alternative="gre
ater", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_terremotos$Profundidad[data_terremotos$Pais ==
"argentina"] and data_terremotos$Profundidad[data_terremotos$Pais ==
"brazil"]
## t = 15.502, df = 2104.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  60.862    Inf
## sample estimates:
## mean of x mean of y
## 159.44920  91.35914
```

El p-value es menor al una significancia de 0.01. Por lo que se rechaza la hipotesis nula y se acepta la alternativa. Es decir aceptamos que Argentina tiene una Profundidad media mayor que Brazil, con un nivel de confianza de 99%.

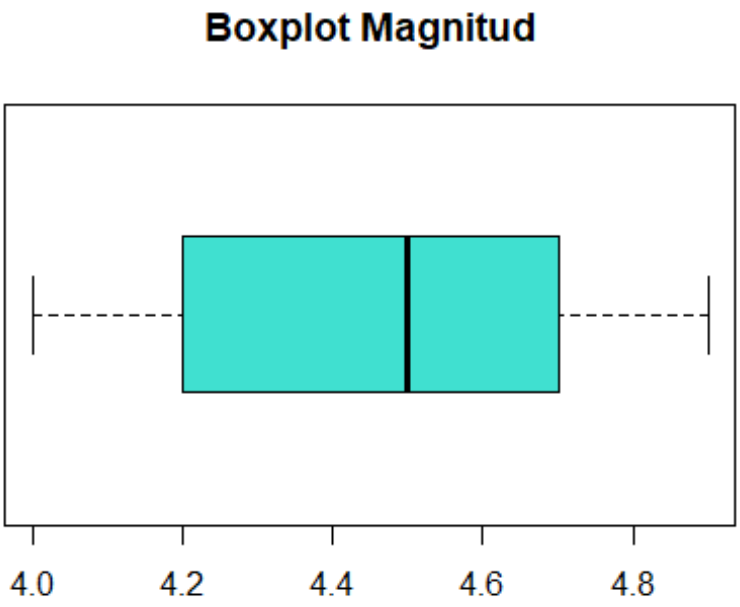
**Con esto validamos que Argentina es el pais con mayor profundidad sismica de latino america**

### Regresión Logística

Vamos a utilizar la regresión logística para determinar qué variables son relevantes en la ocurrencia de un sismo moderado.

Se construirá una variable dicotómica para utilizar como variable dependiente. Vamos a utilizar la magnitud y determinar el umbral para la variable dicotómica. Primero, observaremos la distribución de la magnitud.

```
# Diagrama de cajas de Magnitud
boxplot(data_terremotos$Magnitud, horizontal = TRUE,
        col=c("turquoise"),
        main="Boxplot Magnitud")
```




En base a lo observado en el gráfico previo utilizaremos un valor cercano a la mediana como umbral. En este caso utilizaremos 4.5. Si la magnitud del sismo es igual o superior a este valor lo consideraremos Moderado/Fuerte. En caso de ser inferior consideraremos al sismo leve.

```
# Se crea la variable dicotómica en base a la Magnitud.
data_terremotos$Nivel.Sismo <- case_when(data_terremotos$Magnitud >= 4.5
~ 1,
                                         data_terremotos$Magnitud < 4.5 ~
0)

table(data_terremotos$Nivel.Sismo)

##
##    0    1
## 5963 6763

attach(data_terremotos)
```

Ahora, vamos a construir el modelo de regresión logística. Para ello vamos a incluiyendo variable por variable para determinar qué tan significantes para el modelo.

**Paso 1: País.**

Vamos a empezar agregando el país al modelo de regresión logística.

```
# Construimos el modelo logistico
model.logist1 = glm(formula= Nivel.Sismo ~
Pais,family=binomial(link=logit))
summary(model.logist1)

##
## Call:
## glm(formula = Nivel.Sismo ~ Pais, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2335  -1.0721   0.1037   1.0178   1.6420
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.25281    0.05856  -4.317 1.58e-05 ***
## Paisbolivia              -0.19878    0.07612  -2.611  0.00902 **
## Paisbrazil                4.29257    0.25890  16.580  < 2e-16 ***
## Paischile                 5.47533    0.45199  12.114  < 2e-16 ***
## Paiscolombia              0.13767    0.08291   1.660  0.09684 .
## Paiscostarica             0.03313    0.10583   0.313  0.75428
## Paisdominican-republic -0.29669    0.16014  -1.853  0.06392 .
## Paisecuador               0.03485    0.10043   0.347  0.72859
## Paisguatemala            -0.79450    0.08475  -9.375  < 2e-16 ***
## Paismexico                0.68778    0.07678   8.957  < 2e-16 ***
## Paispanama                0.16147    0.14164   1.140  0.25430
## Paisparaguay              0.74529    0.38716   1.925  0.05423 .
## Paisperu                  0.64045    0.08061   7.945 1.95e-15 ***
## Paispuertorico           -0.49163    0.20554  -2.392  0.01676 *
## Paisvenezuela            -0.23098    0.10302  -2.242  0.02496 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17592  on 12725  degrees of freedom
## Residual deviance: 14640  on 12711  degrees of freedom
## AIC: 14670
##
## Number of Fisher Scoring iterations: 7
```

Se observa que la mayoría de los países son significantes para el modelo (p-value < 0.05). Por lo tanto, **la variable país será relevante para el modelo.** El índice AIC es 14670.

**Paso 2: Longitud.**

Vamos a agregar la variable longitud al modelo de regresión logística.

```
# Construimos el modelo logistico
model.logist2 = glm(formula= Nivel.Sismo ~ Pais + Longitud,
```

```
family=binomial(link=logit))
summary(model.logist2)

##
## Call:
## glm(formula = Nivel.Sismo ~ Pais + Longitud, family = binomial(link =
logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2658  -1.0603   0.1007   1.0609   1.6513
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              0.741713   0.337688   2.196  0.02806 *
## Paisbolivia             -0.179518   0.076399  -2.350  0.01879 *
## Paisbrazil               4.315658   0.259085  16.657  < 2e-16 ***
## Paischile                5.574350   0.453516  12.291  < 2e-16 ***
## Paiscolombia             0.291253   0.097541   2.986  0.00283 **
## Paiscostarica            0.291128   0.136550   2.132  0.03300 *
## Paisdominican-republic -0.241903   0.161195  -1.501  0.13344
## Paisecuador              0.231460   0.120031   1.928  0.05381 .
## Paisguatemala           -0.431192   0.148122  -2.911  0.00360 **
## Paismexico               1.155020   0.174462   6.620 3.58e-11 ***
## Paispanama               0.372207   0.158224   2.352  0.01865 *
## Paisparaguay             0.673594   0.387954   1.736  0.08252 .
## Paisperu                 0.762412   0.090400   8.434  < 2e-16 ***
## Paispuertorico          -0.506446   0.205608  -2.463  0.01377 *
## Paisvenezuela           -0.203890   0.103452  -1.971  0.04874 *
## Longitud                 0.014787   0.004945   2.990  0.00279 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17592  on 12725  degrees of freedom
## Residual deviance: 14631  on 12710  degrees of freedom
## AIC: 14663
##
## Number of Fisher Scoring iterations: 7
```

Se observa que la longitud es significante para el modelo (p-value < 0.05). Por lo tanto, **la variable longitud será relevante para el modelo.** Se observa que la longitud no mejora significativamente el índice AIC. (Se reduce solo a 14663). Para determinar qué variable de estas dos aporta más al modelo se utilizará el test de anova.

```
# Aplicamos el test basado en la devianza
anova(model.logist2,test="Chisq")

## Analysis of Deviance Table
##
```

```
## Model: binomial, link: logit
##
## Response: Nivel.Sismo
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     12725      17592
## Pais     14  2951.94     12711      14640 < 2.2e-16 ***
## Longitud  1     8.93     12710      14631  0.002803 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se observa del Test Anova que el país contribuye más al modelo que la longitud. **Por lo tanto, no vamos a considerar la longitud en nuestro modelo.**

**Paso 3: Latitud.**

Vamos a agregar la variable latitud al modelo de regresión logística.

```
# Construimos el modelo logistico
model.logist3 = glm(formula= Nivel.Sismo ~ Pais + Latitud,
family=binomial(link=logit))
summary(model.logist3)
```

```
##
## Call:
## glm(formula = Nivel.Sismo ~ Pais + Latitud, family = binomial(link =
logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2644  -1.0529   0.1024   1.0634   1.6607
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.593658   0.149855  -3.962 7.45e-05 ***
## Paisbolivia             -0.119993   0.082551  -1.454  0.14607
## Paisbrazil               4.355840   0.260466  16.723  < 2e-16 ***
## Paischile                5.453994   0.452051  12.065  < 2e-16 ***
## Paiscolombia             0.516684   0.174325   2.964  0.00304 **
## Paiscostarica            0.494516   0.214602   2.304  0.02120 *
## Paisdominican-republic   0.280434   0.283145   0.990  0.32196
## Paisecuador              0.351580   0.162835   2.159  0.03084 *
## Paisguatemala           -0.272513   0.227569  -1.197  0.23111
## Paismexico               1.260421   0.244244   5.161 2.46e-07 ***
## Paispanama               0.606849   0.229221   2.647  0.00811 **
## Paisparaguay             0.769604   0.387358   1.987  0.04694 *
## Paisperu                 0.848375   0.116701   7.270 3.60e-13 ***
## Paispuertorico           0.085313   0.311024   0.274  0.78386
```

```
## Paisvenezuela           0.216322   0.208235   1.039  0.29888
## Latitud                -0.012628   0.005108  -2.472  0.01343 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17592  on 12725  degrees of freedom
## Residual deviance: 14634  on 12710  degrees of freedom
## AIC: 14666
##
## Number of Fisher Scoring iterations: 7
```

Se observa que la latitud es significante para el modelo (p-value < 0.05). Por lo tanto, **la variable latitud será relevante para el modelo.** Sin embargo, se observa que la latitud no mejora significativamente el índice AIC. (Se reduce solo a 14666). Para determinar qué variable de estas dos aporta más al modelo se utilizará el test de anova.

```
# Aplicamos el test basado en la devianza
anova(model.logist3,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Nivel.Sismo
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    12725      17592
## Pais    14  2951.94     12711      14640  < 2e-16 ***
## Latitud  1     6.09     12710      14634  0.01356 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se observa del Test Anova que el país contribuye más al modelo que la latitud. **Por lo tanto, no vamos a considerar la latitud en nuestro modelo.**

**Paso 4: Año.**

Vamos a agregar la variable año al modelo de regresión logística.

```
# Construimos el modelo logistico
model.logist4 = glm(formula= Nivel.Sismo ~ Pais + Año,
family=binomial(link=logit))
summary(model.logist4)
```

```
##
## Call:
## glm(formula = Nivel.Sismo ~ Pais + Año, family = binomial(link =
logit))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.248  -1.055   0.103   1.037   1.669
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -23.178783  12.731479  -1.821  0.06867 .
## Paisbolivia             -0.196730   0.076142  -2.584  0.00977 **
## Paisbrazil               4.294708   0.258912  16.588  < 2e-16 ***
## Paischile                5.461861   0.452046  12.083  < 2e-16 ***
## Paiscolombia             0.135317   0.082935   1.632  0.10277
## Paiscostarica            0.038330   0.105887   0.362  0.71736
## Paisdominican-republic  -0.263797   0.161198  -1.636  0.10174
## Paisecuador              0.028967   0.100501   0.288  0.77317
## Paisguatemala           -0.789985   0.084792  -9.317  < 2e-16 ***
## Paismexico               0.699488   0.077081   9.075  < 2e-16 ***
## Paispanama               0.163206   0.141664   1.152  0.24929
## Paisparaguay             0.750643   0.387218   1.939  0.05256 .
## Paisperu                 0.646123   0.080693   8.007 1.17e-15 ***
## Paispuertorico          -0.491080   0.205584  -2.389  0.01691 *
## Paisvenezuela           -0.238107   0.103117  -2.309  0.02094 *
## Año                      0.011374   0.006316   1.801  0.07174 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17592  on 12725  degrees of freedom
## Residual deviance: 14636  on 12710  degrees of freedom
## AIC: 14668
##
## Number of Fisher Scoring iterations: 7
```

Se observa que el año no es significante para el modelo (p-value > 0.05). Por lo tanto, **esta variable no será tomanda en cuenta para el modelo.**

**Paso 5: Mes.**

Vamos a agregar la variable mes como factor (no hay ordinalidad) al modelo de regresión logística.

```
# Construimos el modelo logistico
model.logist5 = glm(formula= Nivel.Sismo ~ Pais + factor(Mes),
family=binomial(link=logit))
summary(model.logist5)
```

```
##
## Call:
## glm(formula = Nivel.Sismo ~ Pais + factor(Mes), family = binomial(link
= logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2493  -1.0445   0.1053   1.0685   1.7177
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.202756   0.090989  -2.228   0.0259 *
## Paisbolivia             -0.201294   0.076254  -2.640   0.0083 **
## Paisbrazil               4.293455   0.258963  16.579  < 2e-16 ***
## Paischile                5.476721   0.452019  12.116  < 2e-16 ***
## Paiscolombia             0.137912   0.083046   1.661   0.0968 .
## Paiscostarica            0.039531   0.106090   0.373   0.7094
## Paisdominican-republic -0.316113   0.162856  -1.941   0.0523 .
## Paisecuador              0.035224   0.100614   0.350   0.7263
## Paisguatemala           -0.794368   0.084887  -9.358  < 2e-16 ***
## Paismexico               0.682861   0.077069   8.860  < 2e-16 ***
## Paispanama               0.165020   0.141940   1.163   0.2450
## Paisparaguay             0.764473   0.387775   1.971   0.0487 *
## Paisperu                 0.636284   0.080764   7.878 3.32e-15 ***
## Paispuertorico          -0.486753   0.206063  -2.362   0.0182 *
## Paisvenezuela           -0.230938   0.103189  -2.238   0.0252 *
## factor(Mes)2            -0.145516   0.102303  -1.422   0.1549
## factor(Mes)3            -0.111652   0.097054  -1.150   0.2500
## factor(Mes)4            -0.082316   0.094679  -0.869   0.3846
## factor(Mes)5             0.009505   0.099123   0.096   0.9236
## factor(Mes)6            -0.218346   0.099133  -2.203   0.0276 *
## factor(Mes)7            -0.081207   0.097190  -0.836   0.4034
## factor(Mes)8            -0.089737   0.099807  -0.899   0.3686
## factor(Mes)9             0.052165   0.096827   0.539   0.5901
## factor(Mes)10            0.113139   0.102189   1.107   0.2682
## factor(Mes)11           -0.088082   0.100595  -0.876   0.3812
## factor(Mes)12            0.083187   0.101309   0.821   0.4116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17592  on 12725  degrees of freedom
## Residual deviance: 14618  on 12700  degrees of freedom
## AIC: 14670
##
## Number of Fisher Scoring iterations: 7
```

Se observa que la mayoría de los factores del mes no son significantes para el modelo (p-value > 0.05). Por lo tanto, **la variable Mes no se tomará en cuenta para el modelo.**

**Paso 6: Día de semana.**

Vamos a agregar la variable Dia.Semana al modelo de regresión logística.

```
# Construimos el modelo logistico
model.logist6 = glm(formula= Nivel.Sismo ~ Pais + Dia.Semana,
family=binomial(link=logit))
summary(model.logist6)
```

```
##
## Call:
## glm(formula = Nivel.Sismo ~ Pais + Dia.Semana, family = binomial(link
= logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2427  -1.0376   0.1031   1.0331   1.6810
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.252155   0.058588  -4.304 1.68e-05 ***
## Paisbolivia             -0.201117   0.076168  -2.640  0.00828 **
## Paisbrazil               4.293277   0.258915  16.582  < 2e-16 ***
## Paischile                5.473758   0.451997  12.110  < 2e-16 ***
## Paiscolombia             0.138678   0.082951   1.672  0.09456 .
## Paiscostarica            0.030669   0.105892   0.290  0.77211
## Paisdominican-republic -0.302798   0.160370  -1.888  0.05901 .
## Paisecuador              0.036836   0.100478   0.367  0.71391
## Paisguatemala           -0.795721   0.084777  -9.386  < 2e-16 ***
## Paismexico               0.689139   0.076825   8.970  < 2e-16 ***
## Paispanama               0.161727   0.141719   1.141  0.25379
## Paisparaguay             0.757042   0.387463   1.954  0.05072 .
## Paisperu                 0.638535   0.080643   7.918 2.41e-15 ***
## Paispuertorico          -0.491322   0.205638  -2.389  0.01688 *
## Paisvenezuela           -0.232491   0.103062  -2.256  0.02408 *
## Dia.Semana.L            -0.010393   0.052017  -0.200  0.84164
## Dia.Semana.Q            -0.065572   0.051698  -1.268  0.20467
## Dia.Semana.C            -0.114584   0.052315  -2.190  0.02850 *
## Dia.Semana^4            -0.002115   0.052397  -0.040  0.96780
## Dia.Semana^5             0.020787   0.052613   0.395  0.69278
## Dia.Semana^6             0.021851   0.051977   0.420  0.67419
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17592  on 12725  degrees of freedom
```

```
## Residual deviance: 14633  on 12705  degrees of freedom
## AIC: 14675
##
## Number of Fisher Scoring iterations: 7
```

Se observa que la mayoría de los factores del día de semana no son significantes para el modelo (p-value > 0.05). Por lo tanto, **la variable Dia.Semana no se tomará en cuenta para el modelo.**

**Paso 7: Hora.**

Vamos a agregar la variable Hora al modelo de regresión logística.

```
# Construimos el modelo logistico
model.logist7 = glm(formula= Nivel.Sismo ~ Pais + factor(Hora),
family=binomial(link=logit))
summary(model.logist7)

##
## Call:
## glm(formula = Nivel.Sismo ~ Pais + factor(Hora), family =
binomial(link = logit))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.253  -1.045   0.103   1.048   1.696
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.314147   0.109931  -2.858  0.00427 **
## Paisbolivia             -0.196610   0.076250  -2.578  0.00992 **
## Paisbrazil               4.295644   0.258961  16.588  < 2e-16 ***
## Paischile                5.479683   0.452015  12.123  < 2e-16 ***
## Paiscolombia             0.139465   0.083080   1.679  0.09322 .
## Paiscostarica            0.038515   0.106008   0.363  0.71636
## Paisdominican-republic -0.296499   0.160458  -1.848  0.06463 .
## Paisecuador              0.042316   0.100625   0.421  0.67410
## Paisguatemala           -0.790891   0.084889  -9.317  < 2e-16 ***
## Paismexico               0.692282   0.076968   8.994  < 2e-16 ***
## Paispanama               0.166497   0.141930   1.173  0.24076
## Paisparaguay             0.770725   0.387736   1.988  0.04684 *
## Paisperu                 0.647418   0.080791   8.013 1.12e-15 ***
## Paispuertorico          -0.492798   0.205781  -2.395  0.01663 *
## Paisvenezuela           -0.228136   0.103249  -2.210  0.02714 *
## factor(Hora)1            0.120278   0.138318   0.870  0.38453
## factor(Hora)2            0.119562   0.134149   0.891  0.37278
## factor(Hora)3            0.014703   0.136116   0.108  0.91398
## factor(Hora)4            0.044915   0.132235   0.340  0.73411
## factor(Hora)5            0.036079   0.135928   0.265  0.79068
## factor(Hora)6           -0.039059   0.135214  -0.289  0.77268
## factor(Hora)7           -0.015617   0.134656  -0.116  0.90767
```

```
## factor(Hora)8           0.097779   0.139545   0.701  0.48349
## factor(Hora)9          -0.039968   0.133280  -0.300  0.76427
## factor(Hora)10         -0.004898   0.135256  -0.036  0.97111
## factor(Hora)11         -0.020488   0.139637  -0.147  0.88335
## factor(Hora)12          0.073775   0.139341   0.529  0.59649
## factor(Hora)13          0.070679   0.132861   0.532  0.59474
## factor(Hora)14          0.089068   0.136317   0.653  0.51351
## factor(Hora)15          0.020324   0.135690   0.150  0.88094
## factor(Hora)16          0.102653   0.134190   0.765  0.44428
## factor(Hora)17          0.258982   0.132884   1.949  0.05130 .
## factor(Hora)18         -0.009435   0.132949  -0.071  0.94342
## factor(Hora)19          0.106241   0.134969   0.787  0.43119
## factor(Hora)20          0.168528   0.135323   1.245  0.21299
## factor(Hora)21          0.025261   0.139634   0.181  0.85644
## factor(Hora)22          0.234132   0.137854   1.698  0.08943 .
## factor(Hora)23         -0.061908   0.136530  -0.453  0.65023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17592  on 12725  degrees of freedom
## Residual deviance: 14622  on 12688  degrees of freedom
## AIC: 14698
##
## Number of Fisher Scoring iterations: 7
```

Se observa que la mayoría de los factores de la hora no son significantes para el modelo (p-value > 0.05). Por lo tanto, **la variable Hora no se tomará en cuenta para el modelo.**

**Se concluye que el país es la única variable relevante para la ocurrencia de un sismo moderado.**

### Correlación

Vamos a evaluar la correlación entre las variables numéricas: Magnitud, profundidad, longitud y latitud. No se tomará en cuenta la correlación entre la longitud y latitud que son variables geográficas que se complementan.

A continuación se evalúa las correlaciones entre variables con el método Spearman (no paramétrico por la ausencia de normalidad en las variables).

#### *Magnitud y profundidad*

```
cor.test(data_terremotos$Magnitud, data_terremotos$Profundidad,
method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  data_terremotos$Magnitud and data_terremotos$Profundidad
```

```
## S = 3.5155e+11, p-value = 0.008159
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.02344946
```

*Magnitud y Longitud*

```
cor.test(data_terremotos$Magnitud, data_terremotos$Longitud,
method="spearman")

##
##  Spearman's rank correlation rho
##
## data:  data_terremotos$Magnitud and data_terremotos$Longitud
## S = 3.2817e+11, p-value = 4.736e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.04462992

length(data_terremotos$Magnitud) #12507

## [1] 12726

length(data_terremotos$Lo) #12507

## [1] 12726
```

*Magnitud y Latitud*

```
cor.test(data_terremotos$Magnitud, data_terremotos$Latitud,
method="spearman")

##
##  Spearman's rank correlation rho
##
## data:  data_terremotos$Magnitud and data_terremotos$Latitud
## S = 4.0807e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1879715
```

*Profundidad y Longitud*

```
cor.test(data_terremotos$Profundidad, data_terremotos$Longitud,
method="spearman")

##
##  Spearman's rank correlation rho
##
## data:  data_terremotos$Profundidad and data_terremotos$Longitud
## S = 1.8847e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##      rho
## 0.451308
```

*Profundidad y Latitud*

```
cor.test(data_terremotos$Profundidad, data_terremotos$Latitud,
method="spearman")

##
##  Spearman's rank correlation rho
##
## data:  data_terremotos$Profundidad and data_terremotos$Latitud
## S = 4.6578e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.3559914
```

---

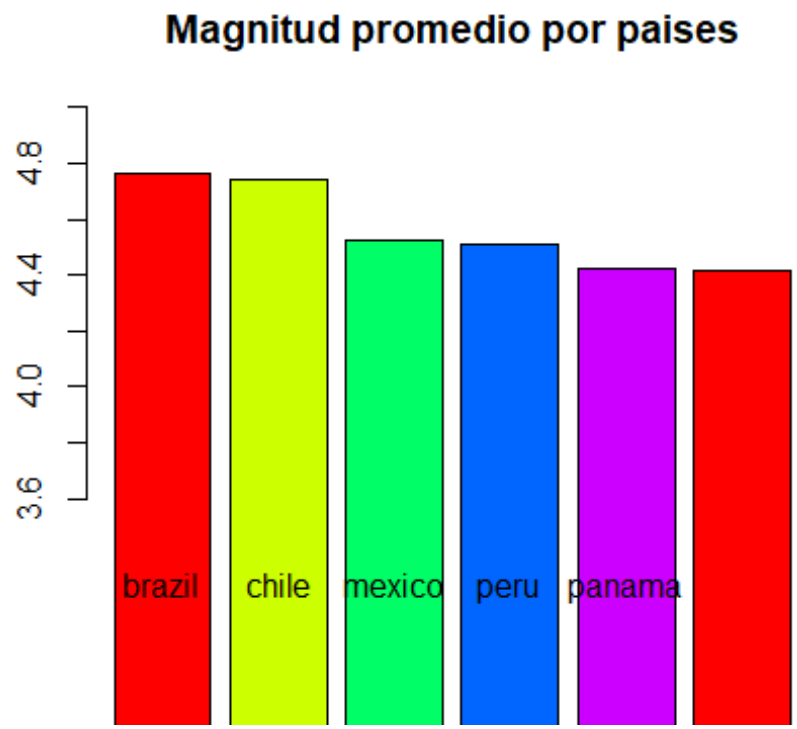
# Resultados

---

## Contraste de Hipotesis

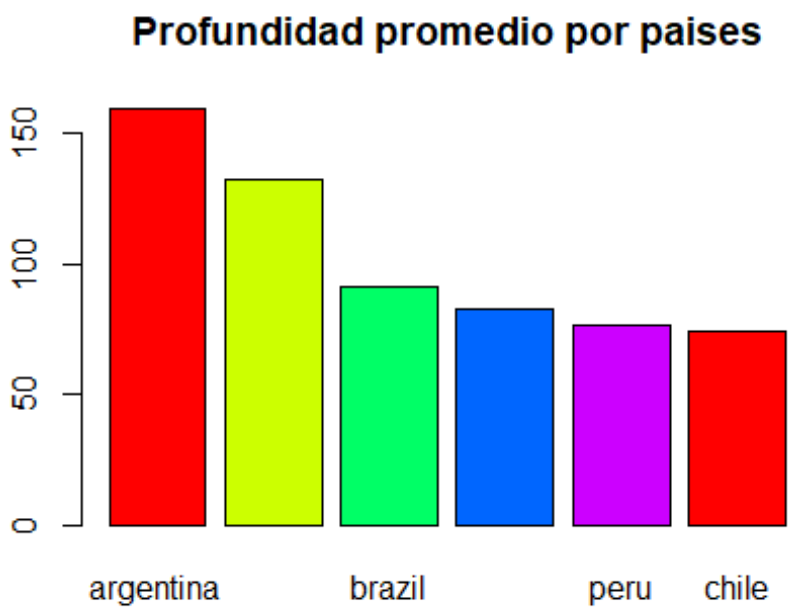
Del analisis del contraste de hipotesis pudimos validar el los paises con mayor Magnitud y Profundidad. En el caso de la Magnitud, **Brazil** es el pais con mayor Magnitud medida (4.75) en los últimos 11 años.

```
# Cargamos ggplot2
library(ggplot2)

# Graficamos las Magnitudes medias de los paises
# Calculamos el top 5 de paises
table_pais_mag.top <-
head(table_pais_mag[table_pais_mag$Pais!='paraguay',])
# Graficamos el barplot
barplot(height = table_pais_mag.top$Media, names.arg =
table_pais_mag.top$Pais,  main = "Magnitud promedio por paises", col =
rainbow(5), ylim=c(3.5,5))
```




En el caso de la Profundidad, **Argentina** es el pais con mayor Profundidad medida (159.1 km) en los últimos 11 años.

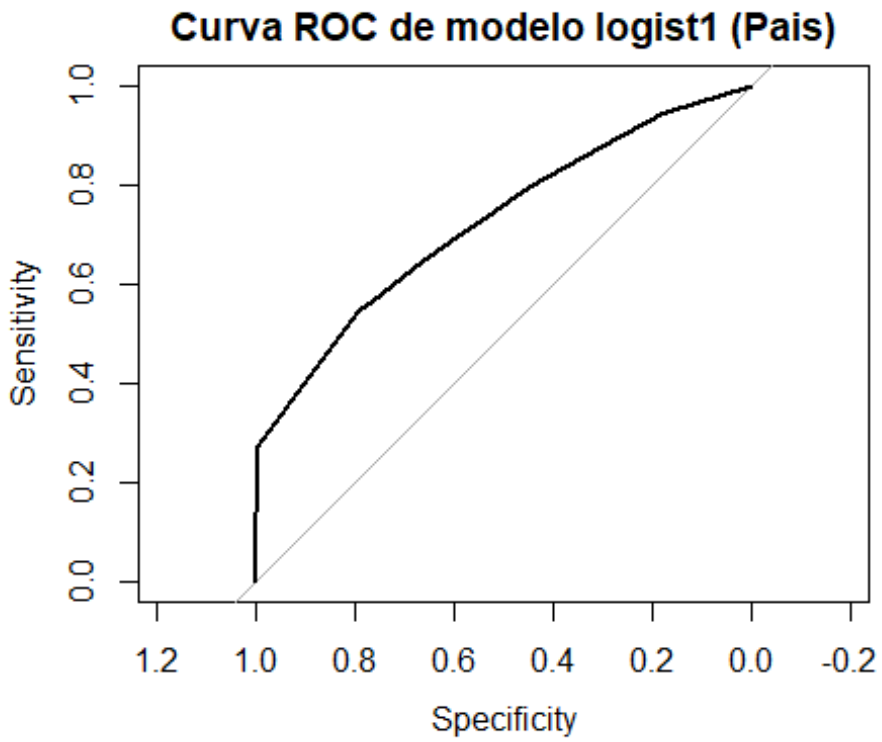
```
# Graficamos las Progundidades medias de los paises
# Calculamos el top 5 de paises
table_pais_pro.top <-
head(table_pais_pro[table_pais_pro$Pais!='paraguay',])
# Graficamos el barplot
barplot(height = table_pais_pro.top$Media, names.arg =
table_pais_pro.top$Pais,  main = "Profundidad promedio por paises", col =
rainbow(5))
```




## Regresión Logística

Vamos a graficar la Curva Roc del mejor modelo de regresión logística (variable independiente: país).

```
# Cargar librería pROC
library(pROC)

# Graficamos la curva ROC.
prob=predict(model.logist1, data_terremotos, type="response")
r=roc(Nivel.Sismo,prob, data=data_terremotos)
plot (r,main="Curva ROC de modelo logist1 (Pais)")
```




Vamos a calcular el área debajo de la curva.

```
#Area debajo de la curva
auc(r)

## Area under the curve: 0.7249
```

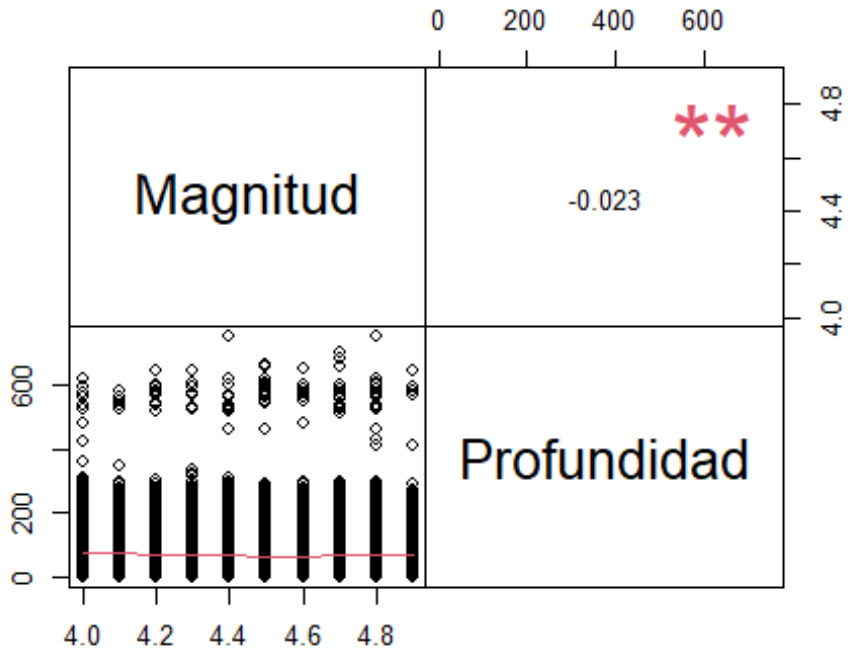
El área debajo de la curva es 0.72, por lo cual concluimos que el modelo es bueno.

## Correlación

Vamos a graficar la correlación de las variables utilizando la función "chart.Correlation".

### Magnitud y profundidad

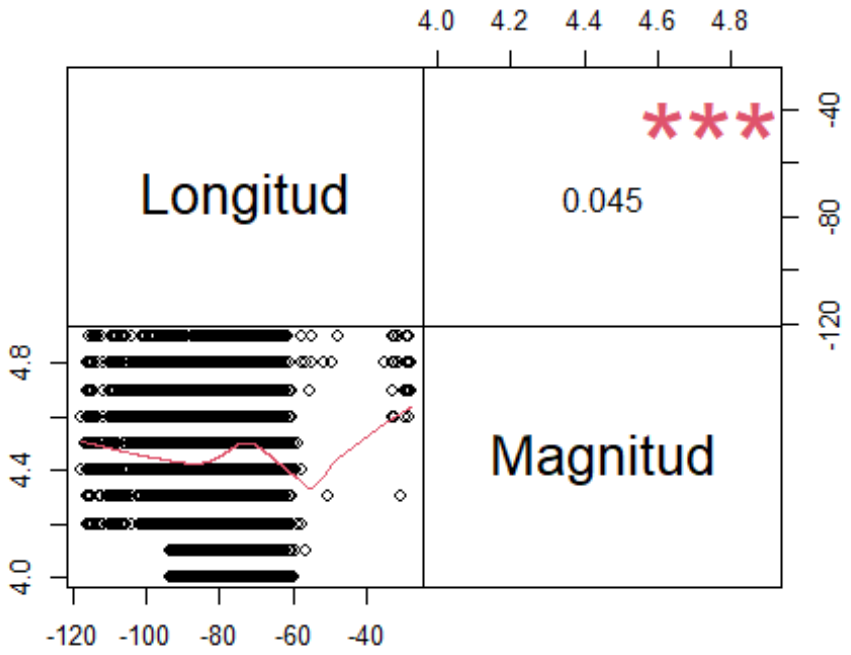
```
# Se cargan las librerías que ayudan a colocar los gráficos.
library(readxl)
library(GGally)
library(Hmisc)
library(corrplot)
library(PerformanceAnalytics)
chart.Correlation(data_terremotos[c("Magnitud", "Profundidad")], method =
"spearman", histogram = F, pch = 19)
```




```
?chart.Correlation
```

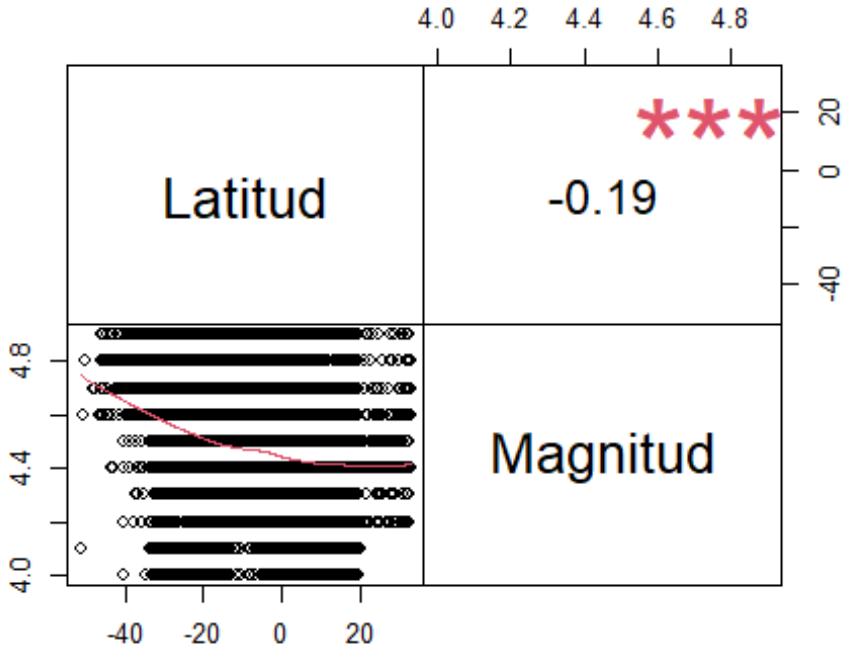
### Longitud y magnitud

```
chart.Correlation(data_terremotos[c("Longitud", "Magnitud")], method =
"spearman", histogram = F, pch = 19)
```



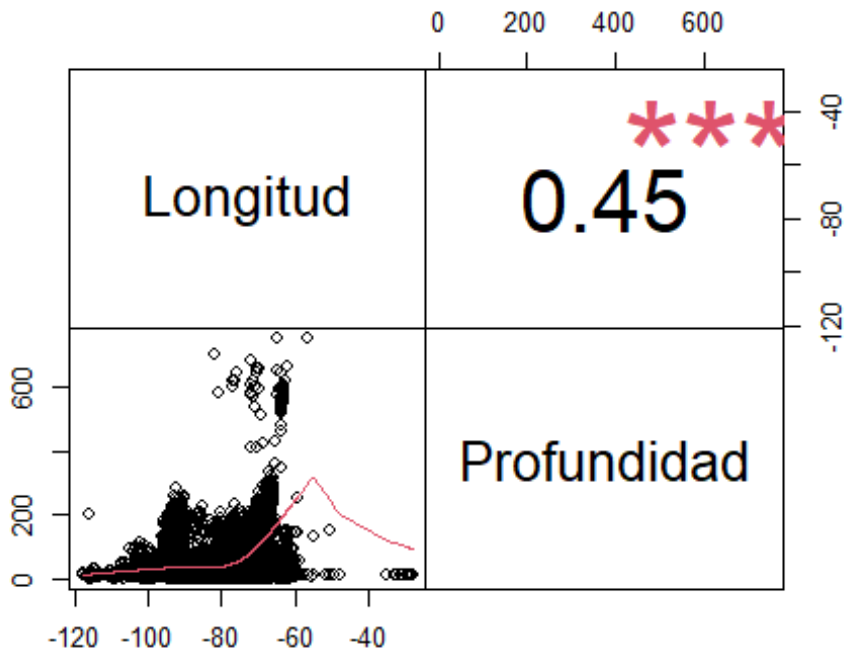

### Latitud y magnitud

```
chart.Correlation(data_terremotos[c("Latitud", "Magnitud")], method =
"spearman", histogram = F, pch = 19)
```



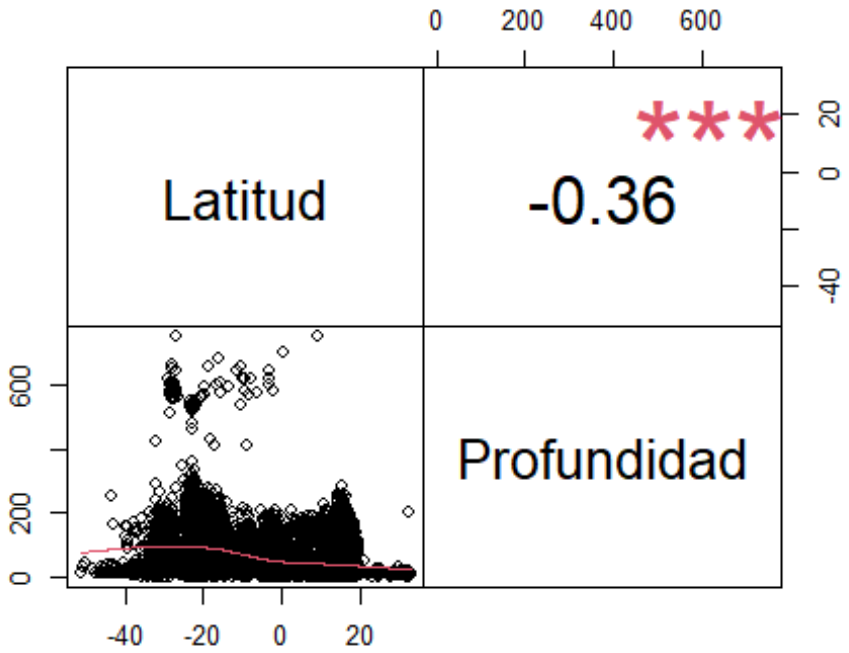
### Longitud y profundidad

```
chart.Correlation(data_terremotos[c("Longitud", "Profundidad")], method =
"spearman", histogram = F, pch = 19)
```




### Latitud y profundidad

```
chart.Correlation(data_terremotos[c("Latitud", "Profundidad")], method =
"spearman", histogram = F, pch = 19)
```




## Resolución del problema

Respondiendo a la pregunta general:

**"¿Existe algún patrón o patrones en la actividad sismica de america latina (centro america y sur america)?"**

Si existen patrones de la actividad sismica de america latina que nos permitió sacar algunas conclusiones. Y respondiendo a las 3 preguntas especificas, tenemos:

- **¿Cuál es el país con actividad sismica más intensa?**

Los paises con actividades sismicas más intensas son Brazil y Argentina.

Brazil con una Magnitud media de 4.75, mayor a los demás paises con un nivel de confianza de 99%. Y Argentina con una Profundidad media 159.1 km, mayor a los demás paises con un nivel de confianza de 99%.

- **¿Qué variables son más relevantes o significativas en la ocurrencia de un sismo moderado?**

A llevar a cabo la regresión logística concluimos que el país es la única variable relevante para la ocurrencia de un sismo moderado. Además, el área debajo de la curva ROC de este modelo fue 0.72.

- **¿Cual es la relación entre las variables de un sismo?**

En base a las pruebas de Correlación entre las variables: Magnitud, profundidad, longitud y latitud podemos llegar a las siguientes conclusiones:

- Los países que se acercaban más a la derecha del meridiano de Greenwich (latitud positiva) considerablemente presentaban una mayor profundidad en el sismo. (Rho: 0.45)
- Mientras más al Sur del Ecuador se encontraban los países donde ocurrían los sismos, tendían moderadamente a tener terremotos de mayor profundidad. (Rho: -0.35)
- Existe muy poca relación entre sismos de zonas geográficas con latitudes más a la izquierda del meridiano de Greenwich (número más negativos) y una alta magnitud del sismo. (Rho: -0.18)
- No hay relación significativa entre magnitud y profundidad de un sismo. (Rho: -0.02)
- No hay relación significativa entre magnitud y longitud de un sismo. (Rho: 0.04)

---

## Código y datos preprocesados

---

El código en formato .rmd se puede descargar en Github desde la siguiente dirección:

https://github.com/ergm569/terremotos_limpieza_analisis/tree/main/code/terremotos_limpieza_analisis.Rmd

Loos datos preprocesados se exportan mediante el siguiente comando:

```
# Exportación de los datos limpios en .csv
write.csv(data_terremotos, "terremotos_centro_sur_america_clean.csv")
```

Y también pueden ser descargados en Github desde la siguiente dirección:

https://github.com/ergm569/terremotos_limpieza_analisis/tree/main/data/terremotos_centro_sur_america_clean.csv

| Contribuciones | Firma |
| --- | --- |
| **Investigación previa** | EG, R |
| **Redacción de respuestas** | EG, R |
| **Desarrollo código** | EG, R |