Actividad 2 EDA

Luis Daniel Dimas Ramirez

19/12/2022

Registro de 1,000 datos de eventos sismicos desde 1964 en Fiji.

Los datos se obtuvieron del package datasets de R Studio.

1: Descripción de las variables

a) Describe cada una de las variables consideradas.

Para este analisis, dejaremos de lado las variables de latitud, longitud y profundidad. Adicionalmente agregare la variable de Clasificación la cual clasificara en categorias dependiendo la magnitud del sismo.

Variables	Descripción
Magnitud	Magnitud del sismo (Escala Richter)
Estaciones	Cantidad de estaciones que reportaron el sismo
Clasificación	Clasificacion del sismo segun su magnitud

b) En una tabla indica el tipo de variable

Variables	Descripción
Magnitud	Cuantitativa Continua Razón
Estaciones	Cuantitativa Discreta Razón
Clasificación	Cualitativa Nominal

2: Análisis descriptivo de la variable cualitativa

a) Realiza una tabla de frecuencias que resuma la informacion contenida en esta variable. Las columnas que debe incluir son: frecuencias absolutas, frecuencias relativas y sus respectivas acumuladas; ademas de los totales. Interpreta los resultados en el contexto de tu problema.

Como parte del tratamiento de datos; para mi variable cualitativa creare los filtros con base en la clasificación que realiza el SSM.

Primero revisare el rango de los datos de la magnitud.

```
range(x$mag)
```

```
## [1] 4.0 6.4
```

```
x$risk <- NA
x$risk[x$mag <= 4.5] <- "Light"
x$risk[x$mag >4.5 & x$mag <= 5.5] <- "Moderate"
x$risk[x$mag > 5.5] <- "Strong"

x <- x[,c(4:6)]
names(x)
```

```
## [1] "mag"      "stations" "risk"
```

Realizando la tabla de frecuencias…

```
risk <- x$risk
(tabla_risk <- table(risk))
```

```
## risk
##    Light Moderate   Strong
##      484      492       24
```

```
(tablar_risk <- round(prop.table(tabla_risk), digits = 4))
```

```
## risk
##    Light Moderate   Strong
##    0.484    0.492    0.024
```

```
levels(as.factor(risk))
```

```
## [1] "Light"    "Moderate" "Strong"
```

```
(f.relativa_risk <- c(
                    sum(risk == "Light")/nrow(x),
                    sum(risk == "Moderate")/nrow(x),
                    sum(risk == "Strong")/nrow(x)))
```

```
## [1] 0.484 0.492 0.024
```

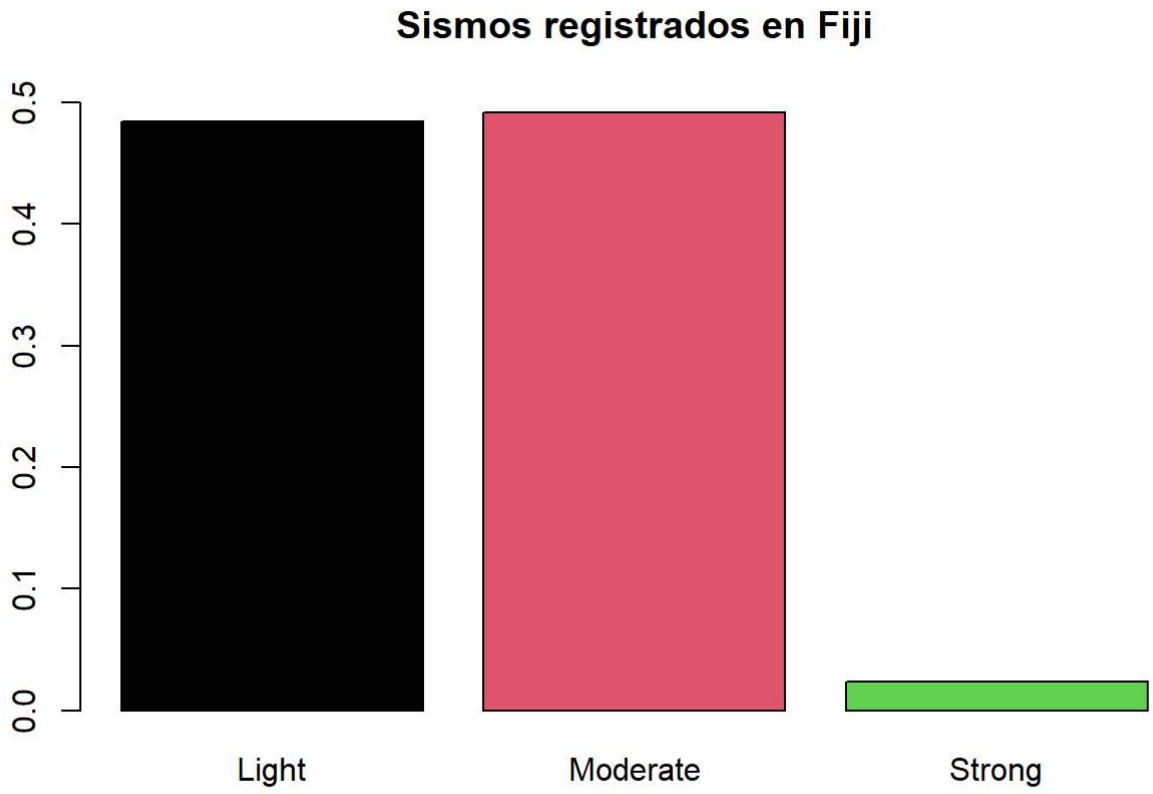
```
tabla.frec_risk <- matrix(cbind(tabla_risk[1],tabla_risk[2],tabla_risk[3], sum(tabla_risk),
    tabla_risk[1], sum(tabla_risk[1:2]),  sum(tabla_risk[1:3]), sum(tabla_risk),
tablar_risk[1],tablar_risk[2],tablar_risk[3], sum(tablar_risk),
    tablar_risk[1], sum(tablar_risk[1:2]),  sum(tablar_risk[1:3]), sum(tablar_risk)
    ), byrow=T, nrow = 4, ncol =4 )


rownames(tabla.frec_risk) <- c("fi", "Fi","pi", "Pi")
colnames(tabla.frec_risk) <-c("Light",  "Moderate" ,  "Strong", "Total" )
tabla.frec_risk
```

```
##      Light Moderate  Strong Total
## fi 484.000  492.000 2.4e+01  1000
## Fi 484.000  976.000 1.0e+03  1000
## pi   0.484    0.492 2.4e-02     1
## Pi   0.484    0.976 1.0e+00     1
```

b) Realiza el analisis grafico adecuado para esta variable

```
barplot(tablar_risk, col = 1:3, ylim = c(0,0.5))
title("Sismos registrados en Fiji")
```



Con base en la tabla podemos observar que practicamente la distribución de los sismos light y moderado son iguales, existe muy poca diferencia. Mientras que para los sismos fuertes son realmente pocos en esta muestra de 1,000 observaciones. El 48.4% de los datos corresponden a riesgo bajo, el %49.2 a riesgo moderado y solamente el 2.4% a un riesgo alto.

c) Elabora una tabla de los estadísticos de resumen que sólo son aplicables a esta variable

Variable	Descripción
Cualitativa nominal	Métodos no parametricos (pruebas de hipotesis de frecuencias esperadas, pruebas de bondad de ajuste y analisis de tablas de contingencia)

Su categorias sirven como clasificación. Las operaciones aritmeticas no son aplicables, carecen de significado o interpretacion

3: Análisis descriptivo de las variables cuantitativas

a) Para cada variable realiza una tabla de frecuencias que contenga las frecuencias absolutas, frecuencias relativas y sus respectivas acumuladas; además de los totales. A partir de estas tablas identifica e indica claramente si la distribucíon para cada variable es unimodal o multimodal. Interpreta los resultados en el contexto de tu problema.

Las variables cuantitativas a analizar son: Magnitud y Estaciones

Magnitud

```
mag <- x$mag

tabla_mag <- table(mag)
tablar_mag <- prop.table(tabla_mag)

range(na.omit(mag))
```

```
## [1] 4.0 6.4
```

```
hist_mag <- hist(mag,breaks=seq(4,6.5,0.5),plot=F)
n <- length(hist_mag$breaks)
tab_mag <- cbind(hist_mag$breaks[-n],hist_mag$breaks[-1],
                 hist_mag$counts,
                  hist_mag$counts/sum(hist_mag$counts),
                  cumsum(hist_mag$counts),
                  cumsum(hist_mag$counts/sum(hist_mag$counts)))
dimnames(tab_mag)[[2]]<-c("Linf","Lsup","f","fr","F","Fr")
print(tab_mag)
```

```
##      Linf Lsup   f    fr    F    Fr
## [1,]  4.0  4.5 484 0.484  484 0.484
## [2,]  4.5  5.0 365 0.365  849 0.849
## [3,]  5.0  5.5 127 0.127  976 0.976
## [4,]  5.5  6.0  22 0.022  998 0.998
## [5,]  6.0  6.5   2 0.002 1000 1.000
```

En este caso la distribución para la magnitud es multimodal.

Estaciones

Para el caso de esta variable y hacer que sea una variable discreta reasignare los datos de tal manera que solamente puedan haber multiplos de 10 correspondientes al numero entero de estaciones.

```
range(x$stations)
```

```
## [1]  10 132
```

```
x$st <- NA
x$st[x$stations >= 0 & x$stations <= 15] <- 10
x$st[x$stations > 15 & x$stations <= 25] <- 20
x$st[x$stations > 25 & x$stations <= 35] <- 30
x$st[x$stations > 35 & x$stations <= 45] <- 40
x$st[x$stations > 45 & x$stations <= 55] <- 50
x$st[x$stations > 55 & x$stations <= 65] <- 60
x$st[x$stations > 65 & x$stations <= 75] <- 70
x$st[x$stations > 75 & x$stations <= 85] <- 80
x$st[x$stations > 85 & x$stations <= 95] <- 90
x$st[x$stations > 95 & x$stations <= 105] <- 100
x$st[x$stations > 105 & x$stations <= 115] <- 110
x$st[x$stations > 115 & x$stations <= 125] <- 120
x$st[x$stations > 125 & x$stations <= 135] <- 130
est <- x$st

tabla_est <- table(est)
tablar_est <- prop.table(tabla_est)

range(na.omit(est))
```

```
## [1]  10 130
```

```
hist_est <- hist(est,breaks=seq(0,140,20),plot=F)
n <- length(hist_est$breaks)
tab_est <- cbind(hist_est$breaks[-n],hist_est$breaks[-1],
                 hist_est$counts,
                  hist_est$counts/sum(hist_est$counts),
                  cumsum(hist_est$counts),
                    cumsum(hist_est$counts/sum(hist_est$counts)))
dimnames(tab_est)[[2]]<-c("Linf","Lsup","f","fr","F","Fr")
print(tab_est)
```

```
##      Linf Lsup   f    fr    F    Fr
## [1,]    0   20 473 0.473  473 0.473
## [2,]   20   40 313 0.313  786 0.786
## [3,]   40   60 113 0.113  899 0.899
## [4,]   60   80  61 0.061  960 0.960
## [5,]   80  100  27 0.027  987 0.987
## [6,]  100  120  11 0.011  998 0.998
## [7,]  120  140   2 0.002 1000 1.000
```

```
library(fdth)
```

```
## Warning: package 'fdth' was built under R version 4.1.3
```

```
##
## Attaching package: 'fdth'
```

```
## The following objects are masked from 'package:stats':
##
##     sd, var
```

```
mfv(na.omit(est))
```

```
## [1] 20
```

La distribución para las estaciones es unimodal.

b) Elabora una gráfica adecuada con la que puedas describir el comportamiento de las observaciones para cada variable en terminos de la dispersión, simetría y la posible existencia de

valores atipicos

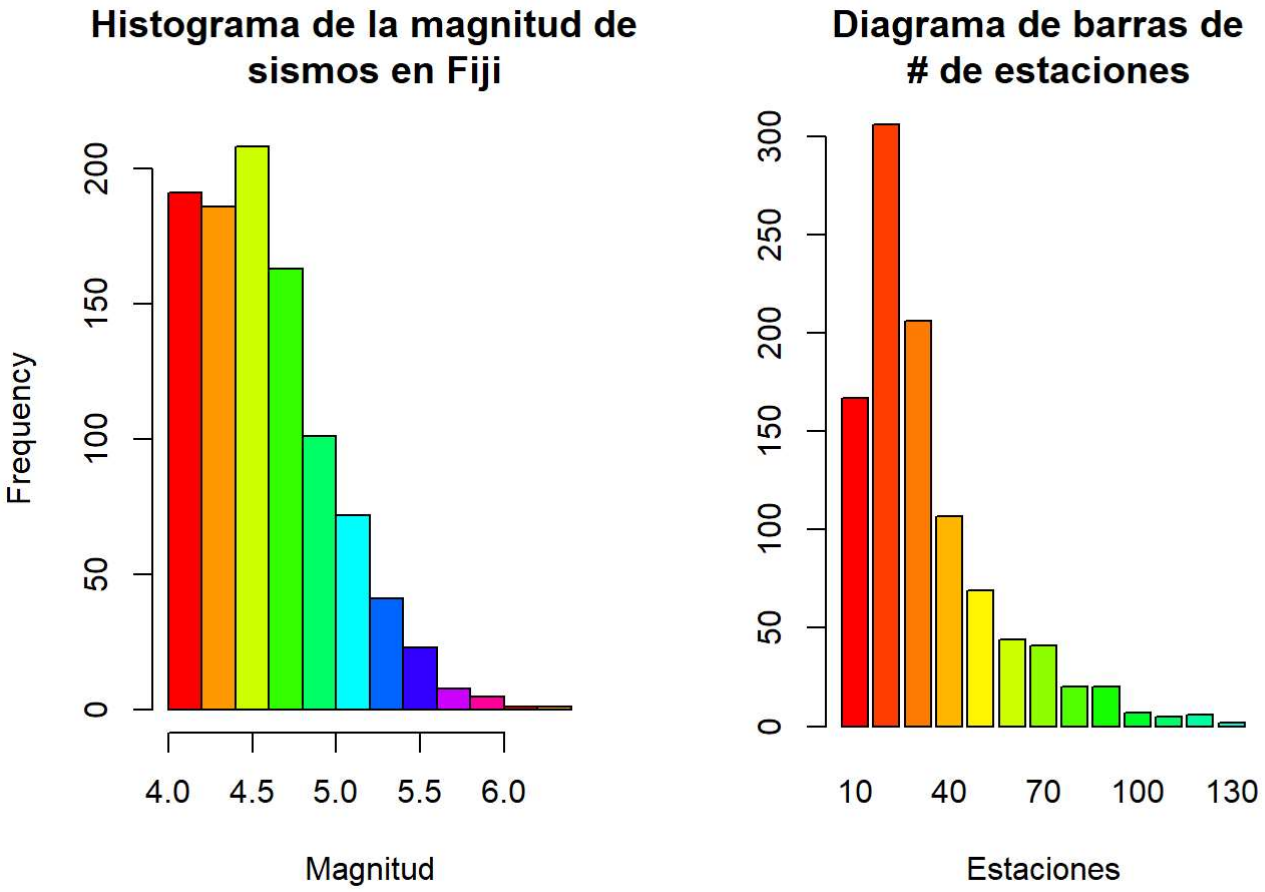
```
par(mfrow=c(1,2))
hist(mag, col=rainbow(10), main="Histograma de la magnitud de \n sismos en Fiji", prob=F, xlab =
"Magnitud")
barplot(tabla_est, col=rainbow(25), main=" Diagrama de barras de  \n # de estaciones", prob=F, x
lab = "Estaciones")
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "prob" is not a graphical
## parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "prob" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "prob"
## is not a graphical parameter
```

```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "prob" is not a
## graphical parameter
```




```
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.1.3
```

```
## Loading required package: ggplot2
```

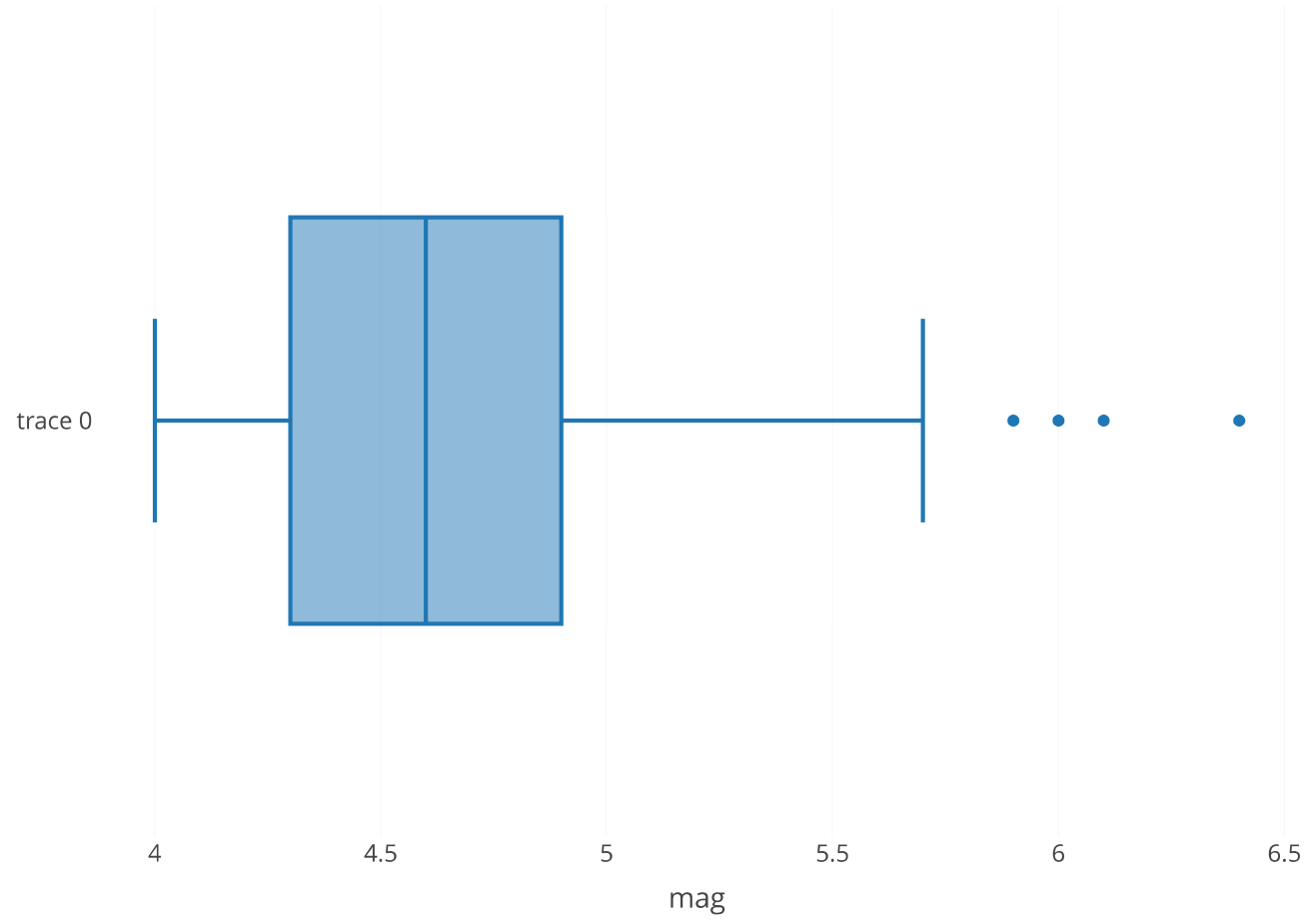
```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

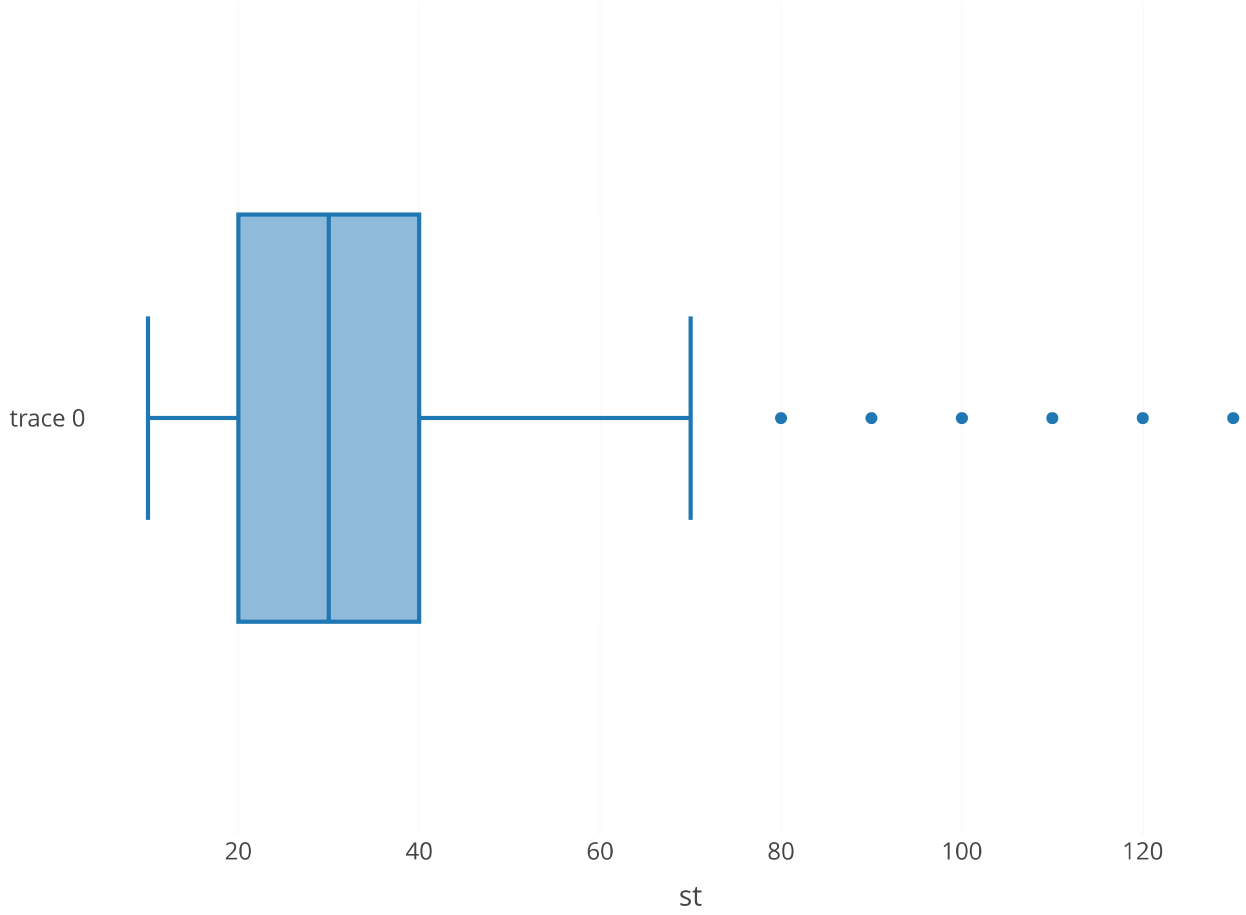
```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
plot_ly(x,x = ~mag, type = "box")
```



```
plot_ly(x,x = ~st, type = "box")
```



Para el caso de la magnitud, notamos que no hay presencia de valores atipicos, aunque son pocos. El valor minimo se encuentra muy cerca del primer cuantil. A diferencia de los valores máximos con el cuantil 3 hay una gran diferencia. La media de la magnitud es de 4.6204 en escala Richter.

En el caso de las estaciones se presentan algunos valores atipicos que comienzan a presentarse arriba de las 90 estaciones. La media de las estaciones es de 32.84. Quiero empezar a inferir que cuando más de 90 estaciones registran un sismos es porque tal vez sea muy fuerte el sismo.

```
(j3<- moments::skewness(mag,na.rm=T))
```

```
## [1] 0.7685997
```

```
(j4<- moments::skewness(est,na.rm=T))
```

```
## [1] 1.549075
```

```
(j5<- moments::kurtosis(mag,na.rm=T))
```

```
## [1] 3.510299
```

```
(j6<- moments::kurtosis(est,na.rm=T))
```

```
## [1] 5.498408
```

```
summary(mag)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    4.30    4.60    4.62    4.90    6.40
```

```
summary(est)
```

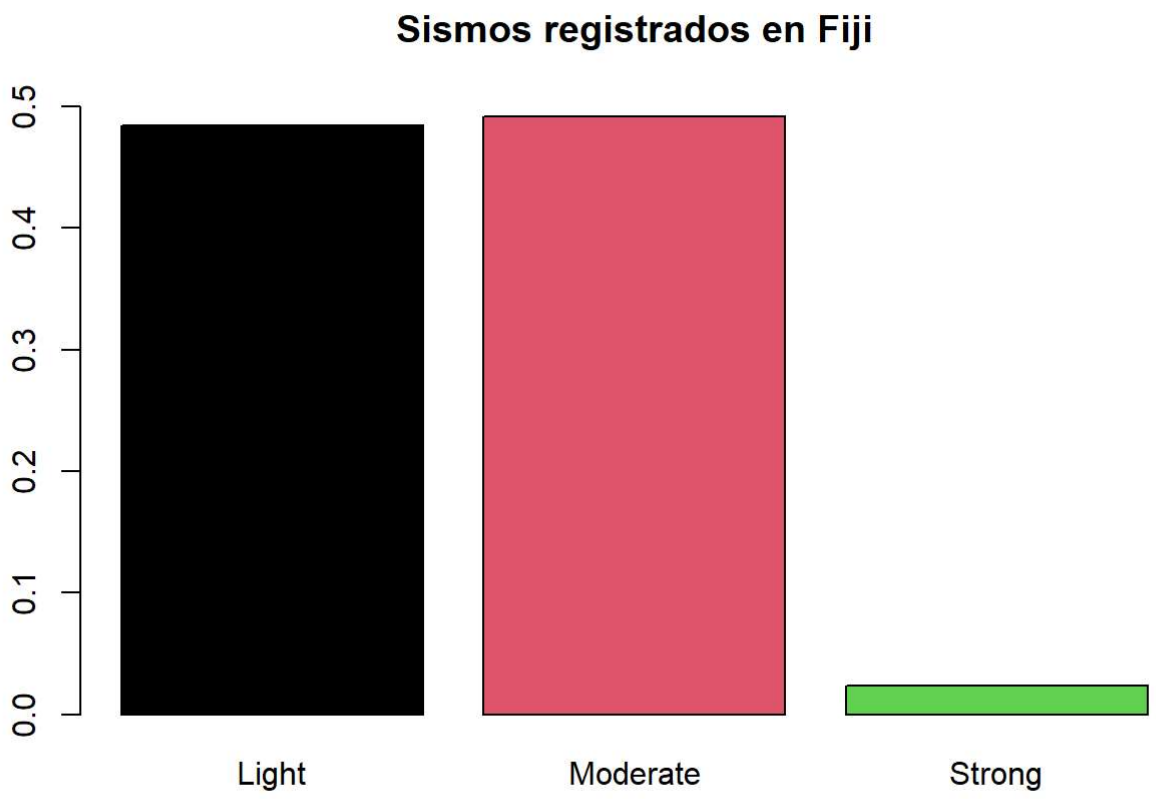
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   20.00   30.00   32.84   40.00  130.00
```

Tomando en cuenta la información anterior y el análisis previo, los datos de estaciones es una distribucion leptocurtica, sin embargo para el caso de la magnitud no me atreveria a decir con tanta seguridad que se trata de una leptocurtica. La distribución de las estaciones muestra un sesgo positivo en sus datos mientras que los datos de Magnitud parece de mucho mayor la amplitud de la curva a diferencia de las estaciones.

4: El problema de comparación y asosiación entre variables

###a) Considerando a la variable cualitativa, compara estadısticamente las distintas categorıas. Utiliza las herramientas adecuadas que se estudiaron en clase. Explica tus resultados lo mas detallado posible y en el contexto de tu problema.

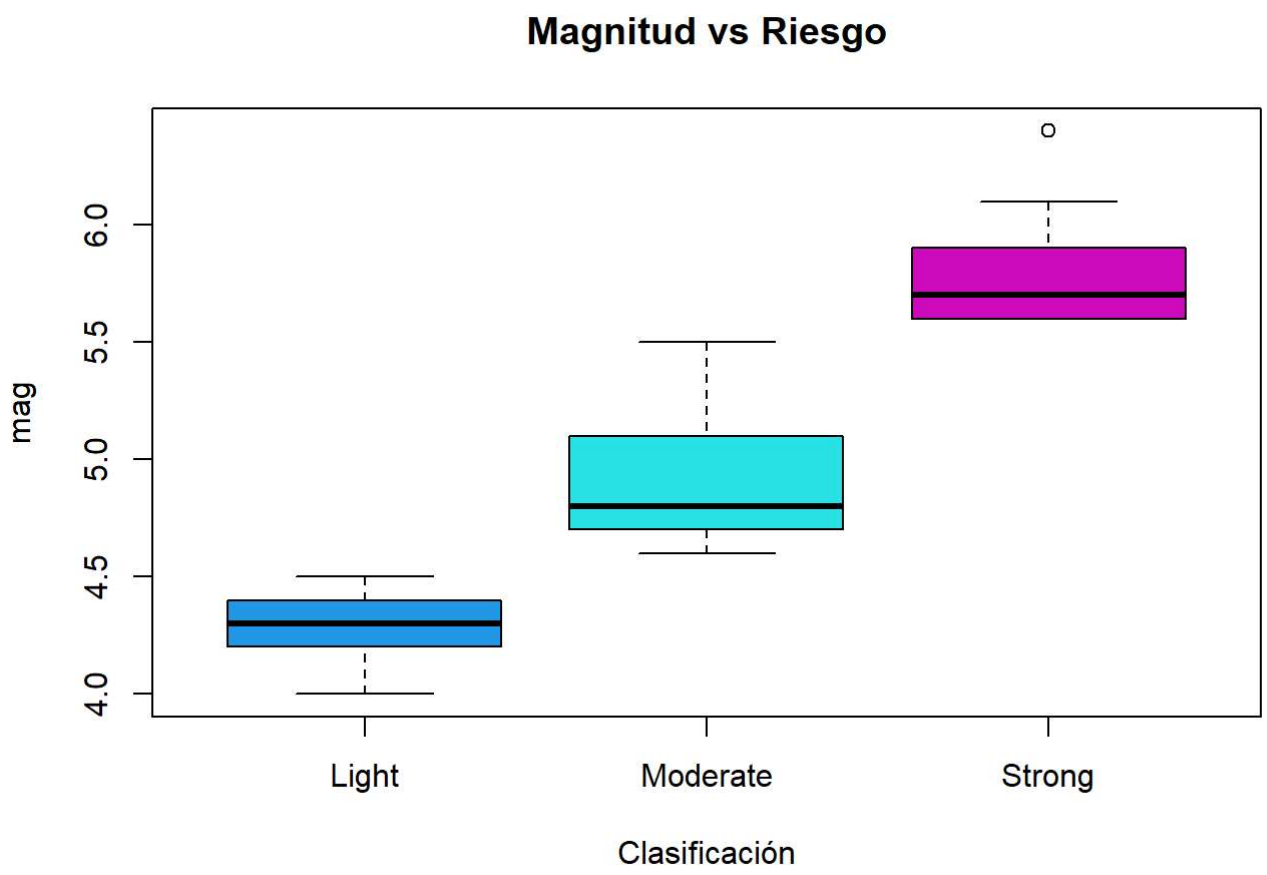
```
par(mfrow=c(1,1))
barplot(tablar_risk, col = 1:3, ylim = c(0,0.5))
title("Sismos registrados en Fiji")
```



Con base en el analisis anterior observamos que practicamente la distribución de los sismos light y moderado son iguales. Los sismos fuertes son realmente pocos en esta muestra de 1,000 observaciones. El 48.4% de los datos corresponden a riesgo bajo, el %49.2 a riesgo moderado y solamente el 2.4% a un riesgo alto.

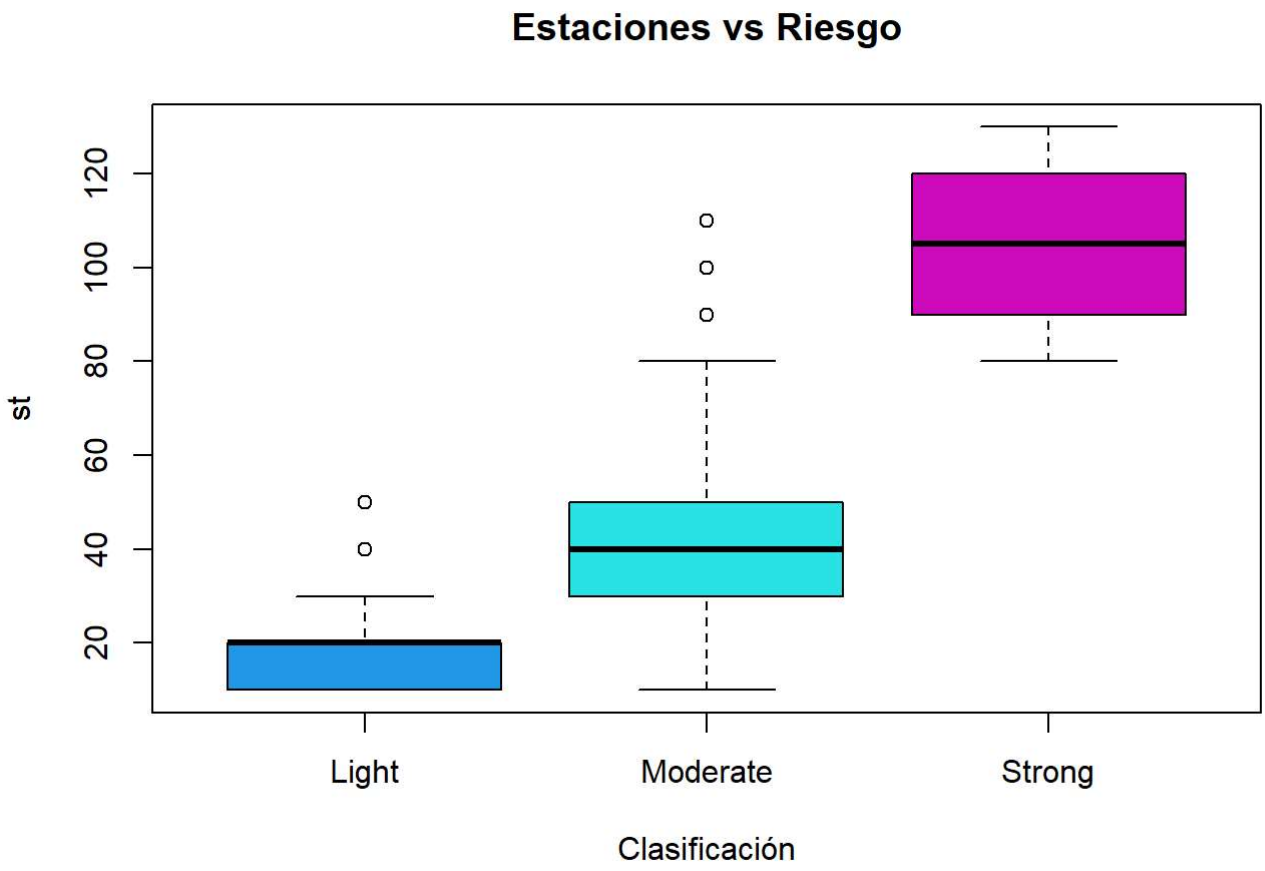
b) Con una de las variables cuantitativas, forma grupos a partir de la variable cualitativa. Realiza el analisis estadıstico correspondiente para comparar el comportamiento de la variable cuantitativa en cada uno de los grupos. Explica tus resultados lo mas detallado posible y en el contexto de tu problema.

```
boxplot(mag ~ risk, data = x, col=20:22, xlab = "Clasificación")
title("Magnitud vs Riesgo")
```



Como era de esperarse, dependiendo la magnitud del sismo cambia la clasificación. Y justo a partir de los 4.5 y 5.5 en escala Richter es que se da el cambio de clasificación. Y aunque el rango para el sismo light y moderado es claro, para la clasificación de sismo alto no tiene un límite aunque despues de los 6 en escala richter, notamos que se presenta un dato atipico porque seguramente las veces que se han dado sismos de esa magnitud son muy pocos o en este caso, solamente uno.

```
boxplot(st ~ risk, data = x, col=20:22, xlab = "Clasificación")
title("Estaciones vs Riesgo")
```



En este caso, tomando en cuenta solamente la caja y no los brazos, podriamos decir que depenendiendo la clasificacion del sismos es el numero de estaciones que reportan el sismo. Si es un sismo light entonces pocas estaciones lo reportan, en cambio si es un sismo fuerte muchas estaciones lo reportaran. De esa manera tiene mucho sentido lo que nos refleja la gráfica. Ya tomando en cuenta todos los datos podemos observa que los brazos se alcanzana a traslapar unos con otros; al grado de que casi el brazo superior se acerca a la media de la siguiente caja. Los riesgos lights y moderados presentan datos atipicos que se acerca al tercer cuantil de la siguiente caja.

c) Analiza el problema de asociación utilizando las dos variables cuantitativa. Explica tus resultados.

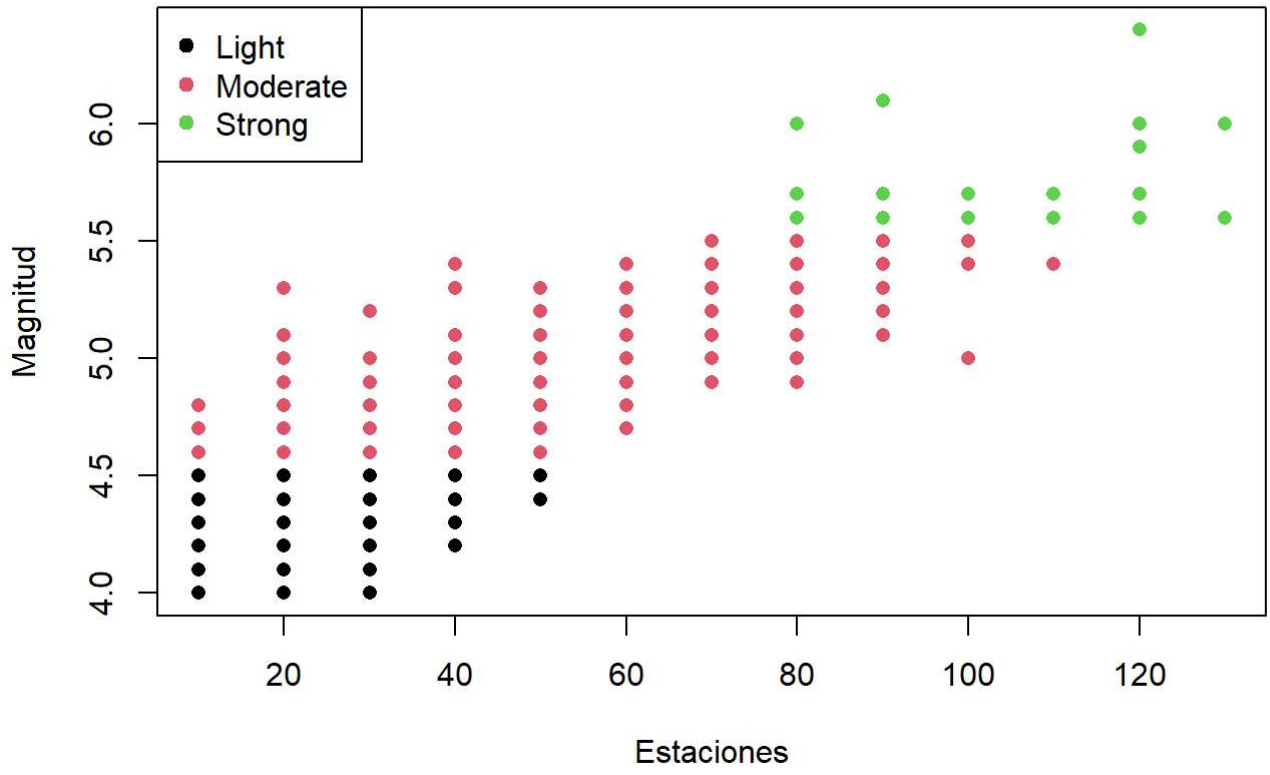
```
names(x)
```

```
## [1] "mag"      "stations" "risk"     "st"
```

```
round(cor(x[c(1,4)]), digits = 2)
```

```
##      mag   st
## mag 1.00 0.85
## st  0.85 1.00
```

```
plot(x$st, x$mag,col= factor(x$risk), pch= 16, xlab= "Estaciones", ylab= "Magnitud")
legend("topleft", legend = levels(factor(x$risk)),pch=19, col= factor(levels(factor(x$risk))))
```



La correlación entre la magnitud y las estaciones es alta.

En la gráfica se observa una relación aparentemente lineal entre los datos.

```
cv_mag <- sd(x$mag)/mean(x$mag); cv_mag
```

```
## [1] 0.08717275
```

```
cv_est <- sd(x$stations)/mean(x$stations); cv_est
```

```
## [1] 0.655347
```

De los coeficientes de variación observo que los datos de las estaciones son bastante heterogeneos.

Despues de este analisis y ver el comportamiento de los datos me parece interesante la relacion que se da entre el riesgo del sismo y el numero de las estaciones que lo reportan. De igual manera la asociación entre la magnitud del sismo y su riesgo era de esperarse.