

Objetivos

Agregación de los datos

Librerías utilizadas

Distribuciones de frecuencias de la variable Country code

Distribución de frecuencias para todos los tipos de interacciones

Distribución de frecuencias para los mensajes

Distribución de frecuencias para los mensajes entrantes

Distribución de frecuencias para los mensajes salientes

Distribución de frecuencias para las llamadas

Distribución de frecuencias para las llamadas entrantes

Distribución de frecuencias para las llamadas salientes

Distribución de frecuencias para el tráfico de Internet

Representación cartográfica de las distribuciones de frecuencias

Tráfico general

Tráfico de SMSs

Tráfico de llamadas

Tráfico de Internet

Análisis por celdas

Generación del tráfico por celdas

Tráfico total

Tráfico promedio

Medidas de posición, dispersión y forma

Transformación logarítmica

Tráfico total

Tráfico promedio

Medidas de posición, dispersión y forma (e. l.)

Dinámica de generación del tráfico

Comunicaciones en Milan

Ekaterina Mitiashkina, Adriana Cecilia Nguema Mbang

14.05.2019

Objetivos

El objetivo general del trabajo es realizar un informe estadístico de tipo descriptivo que permita comprender las características de las variables recogidas en la muestra. Como objetivos específicos, se establecen los siguientes:

1. Construir una nueva hoja de datos que agregue la información de la hoja de datos original por celdas. Incluir en ella, para cada celda, la suma total del tráfico generado en la celda y el promedio del tráfico generado por cada interacción.
2. Proporcionar distribuciones de frecuencias de la variable Country code para los SMSs entrantes, para los salientes, para las llamadas entrantes, para las salientes y para el tráfico de Internet. Interpretar los resultados en términos de los países que más y menos tráfico de mensajes SMS, llamadas e Internet generan.
3. Proporcionar una distribución de frecuencias de la variable Square id, que permita visualizar la dinámica de generación de interacciones en la ciudad de Milán.
4. Proporcionar una distribución de frecuencias del tipo de interacción, clasificando cada una de ellas como de SMS, llamada o Internet.
5. En referencia a la hoja de datos que agrega la información por celdas, realizar un análisis descriptivo de las distribuciones de frecuencias de las siguientes variables:
 - Tráfico total de SMSs recibidos y tráfico promedio de SMSs recibidos por interacción.
 - Tráfico total de SMSs enviados y tráfico promedio de SMSs enviados por interacción.
 - Tráfico total de llamadas recibidas y tráfico promedio de llamadas recibidas por interacción.
 - Tráfico total de llamadas realizadas y tráfico promedio de llamadas realizadas por interacción.
 - Tráfico total de Internet y tráfico promedio de Internet por interacción.
6. Para las variables indicadas en el punto anterior, realizar un análisis descriptivo comparativo que incluya medidas de posición, dispersión y

forma, así como la identificación de celdas atípicas.

7. Realizar una transformación logarítmica de las variables señaladas en el punto 5 y analizar estas nuevas variables como en los puntos 5 y 6. Comparar los análisis realizados en la escala original con los realizados en escala logarítmica.

Agregación de los datos

Primero tenemos que importar los datos. En nuestro caso los datos originales vienen dados en el formato de texto (.txt). Así la importación se hace utilizando la función correspondiente que se encuentra en la pantalla de herramientas en la ruta File>Import Dataset>From Text (base) o con una función read.delim()

```
datos <- read.delim('sms-call-internet-mi-2013-11-01.txt')
names(datos) <- c("Square id", "Time interval", "Country code", "SMS-in activity", "SMS-out activity", "Call-in activity", "Call-out activity", "Internet traffic activity") #nombramos las variables necesarias
summary(datos)

##      Square id      Time interval      Country code      SMS-in activity
##  Min.   : 1   Min.   :1.383e+12   Min.   : 0.0   Min.   : 0.0
##  1st Qu.: 2989  1st Qu.:1.383e+12  1st Qu.: 1.0   1st Qu.: 0.1
##  Median : 5291  Median :1.383e+12  Median : 39.0  Median : 0.5
##  Mean   : 5194  Mean   :1.383e+12  Mean   : 174.6  Mean   : 1.6
##  3rd Qu.: 7439  3rd Qu.:1.383e+12  3rd Qu.: 46.0   3rd Qu.: 1.6
##  Max.   :10000  Max.   :1.383e+12  Max.   :97259.0  Max.   :234.0
##                                         NA's   :1981177
##      SMS-out activity  Call-in activity  Call-out activity
##  Min.   : 0   Min.   : 0   Min.   : 0.0
##  1st Qu.: 0   1st Qu.: 0   1st Qu.: 0.1
##  Median : 1   Median : 0   Median : 0.3
##  Mean   : 2   Mean   : 2   Mean   : 1.4
##  3rd Qu.: 2   3rd Qu.: 2   3rd Qu.: 1.2
##  Max.   :296  Max.   :191  Max.   :187.9
##  NA's   :3210069  NA's   :3329422  NA's   :2611015
##      Internet traffic activity
##  Min.   : 0.0
##  1st Qu.: 0.1
##  Median : 7.8
##  Mean   : 35.0
##  3rd Qu.: 36.3
##  Max.   :4995.6
##  NA's   :2488625
```

Librerías utilizadas

Distribuciones de frecuencias de la variable Country code

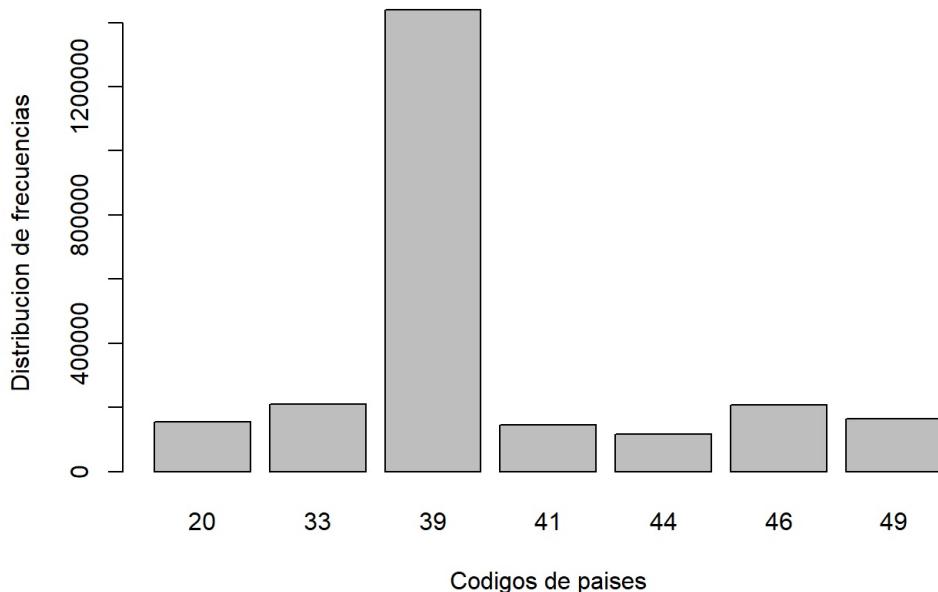
Distribución de frecuencias para todos los tipos de interacciones

El valor de "Country code" nulo no contiene la información representativa y encima hace la distribución de frecuencias muy asimétrica y se complica su análisis por eso hemos puesto un límite correspondiente. Para ver los códigos de los países más y menos frecuentes utilizamos los diagramas de barras poniendo los límites de tal manera que se vean 6-7 países.

La cantidad máxima de las interacciones en Milán sucedían con los números pertenecidos a Italia (cod. 39). Le siguen los códigos de Suecia (cod. 46) y de Francia (cod. 33), luego - los de Alemania (cod. 49) y de Egipto (cod. 20). Al final, el ranking se cierra el código de Escocia (cod. 44).

```
code <- table(datos$`Country code`[datos$`Country code`>0])
barplot(code[code > 100000], main = 'DF maximas para todos los tipos de interacciones',
        xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```

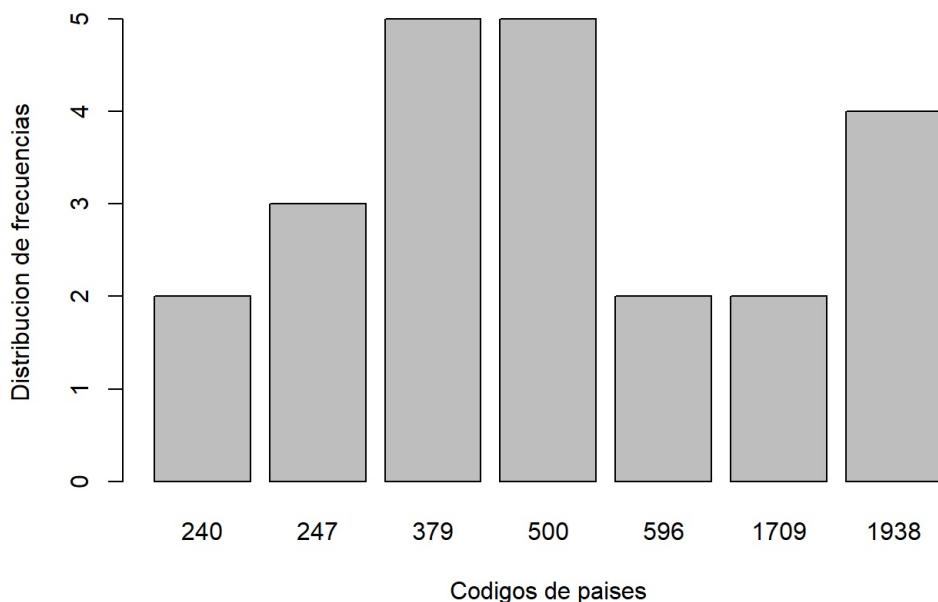
DF maximas para todos los tipos de interacciones



Había dos interacciones con Guinea Ecuatorial (cod. 240), la isla de Martinica (cod. 596), Canada: Newfoundland and Labrador (cod. 1709); tres interacciones con la Isla Ascensión (cod. 247); cuatro - con Estados Unidos: Alabama (cod. 1938) y cinco con Vaticano (cod. 379) y las Islas Malvinas (cod. 500).

```
barplot(code[code < 6], main = 'DF minimas para todos los tipos de interacciones',
       xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```

DF minimas para todos los tipos de interacciones



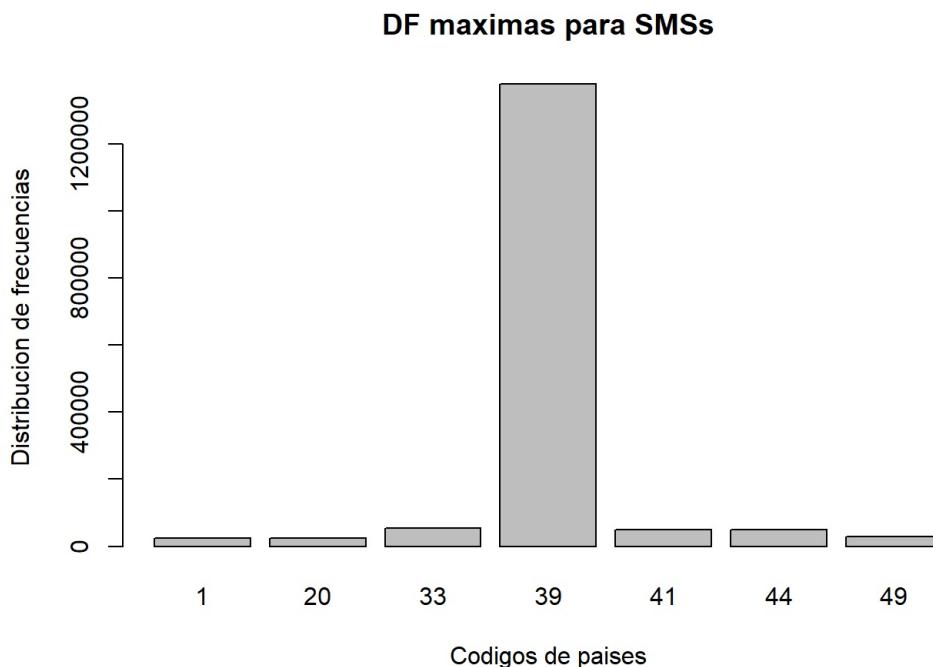
Para hacer un análisis más detallado añadimos a la hoja de datos las tres columnas del formato lógico que nos señalan que tipo de interacción había en cada intervalo del tiempo.

```
datos$SMS <- !is.na(datos$'SMS-in activity') | !is.na(datos$'SMS-out activity')
datos$Calls <- !is.na(datos$'Call-in activity') | !is.na(datos$'Call-out activity')
datos$Internet <- !is.na(datos$'Internet traffic activity')
```

Distribución de frecuencias para los mensajes

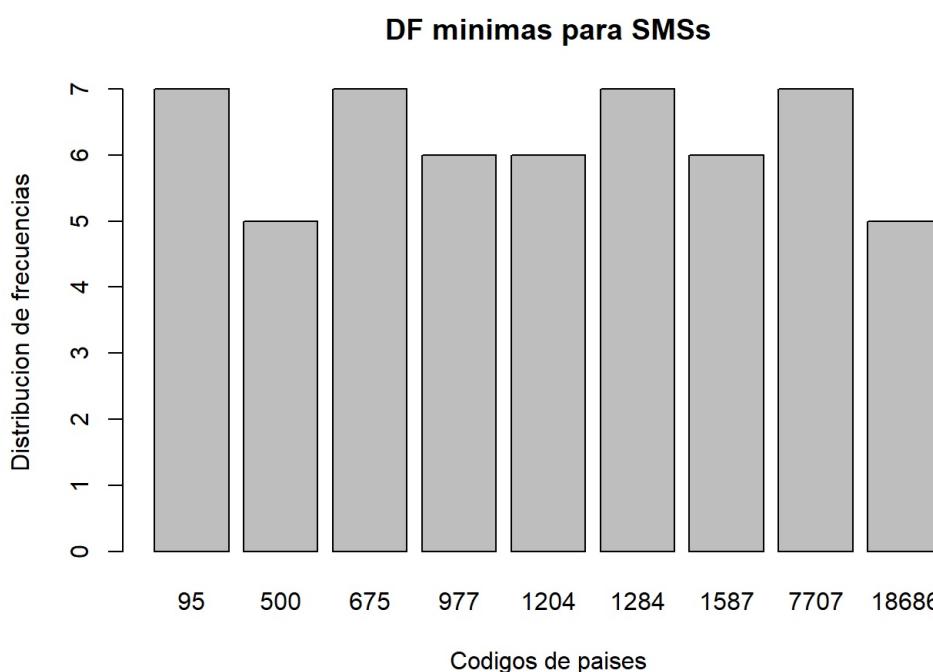
Podemos ver casi la misma proporción, pero no hay ninguna interacción con Suecia (cod. 46); aparecen interacciones con Estados Unidos sin especificar el régión (cod. 1).

```
codeSMS <- table(datos$`Country code`[(datos$SMS == TRUE) & (datos$`Country code`>0)])
barplot(codeSMS[codeSMS > 22000], main = 'DF maximas para SMSs',
        xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```



La distribución de las frecuencias menores de los SMSs no nos representa ninguna correlación con la distribución principal. 5 interacciones tenían lugar con las Islas Malvinas (cod. 500) y con Mexico (cod. 18686); seis - con Nepal (cod. 977), con Cánada: Manitoba (cod. 1204) y otra vez con Cánada: Alberta (cod. 1587); siete - con Birmania (cod. 95), con Papúa Nueva Guinea (cod. 675), con las Islas Vírgenes Británicas (cod. 1284) y con Kazahstán (cod. 7707).

```
barplot(codeSMS[codeSMS < 8], main = 'DF minimas para SMSs',
        xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```

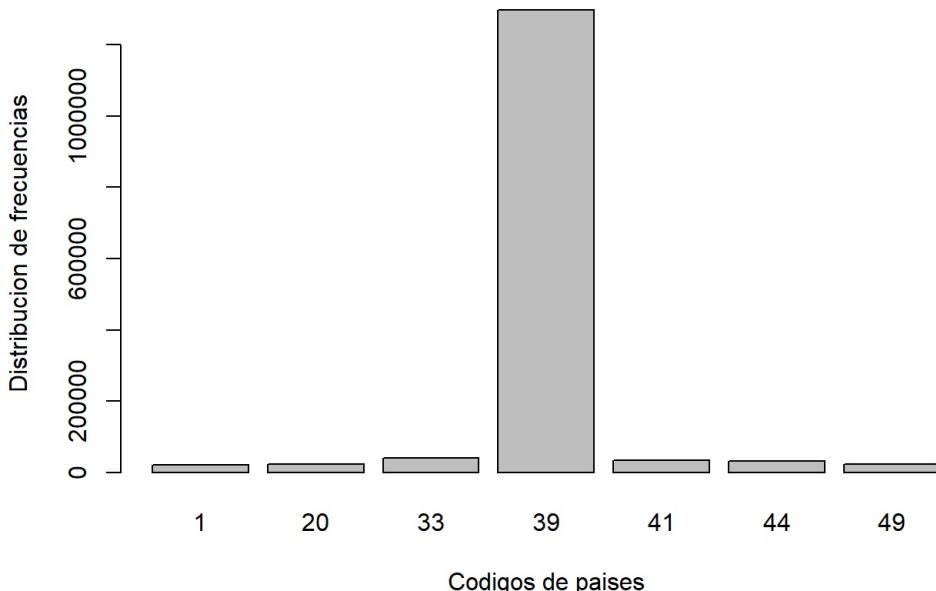


Distribución de frecuencias para los mensajes entrantes

Comparando con la representación de las frecuencias máximas de SMSs en general, no ha cambiado ni la lista de los países, ni su posición respecto uno a otro. Pero todo esto es aplicable solamente si bajamos nuestro filtro porque los SMSs entrantes forman solamente una parte de los SMSs en general.

```
codeSMSentr <- table(datos$`Country code`[!is.na(datos$'SMS-in activity') & (datos$`Country code`>0)])
barplot(codeSMSentr[codeSMSentr > 21000], main = 'DF maximas para SMSs entrantes',
        xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```

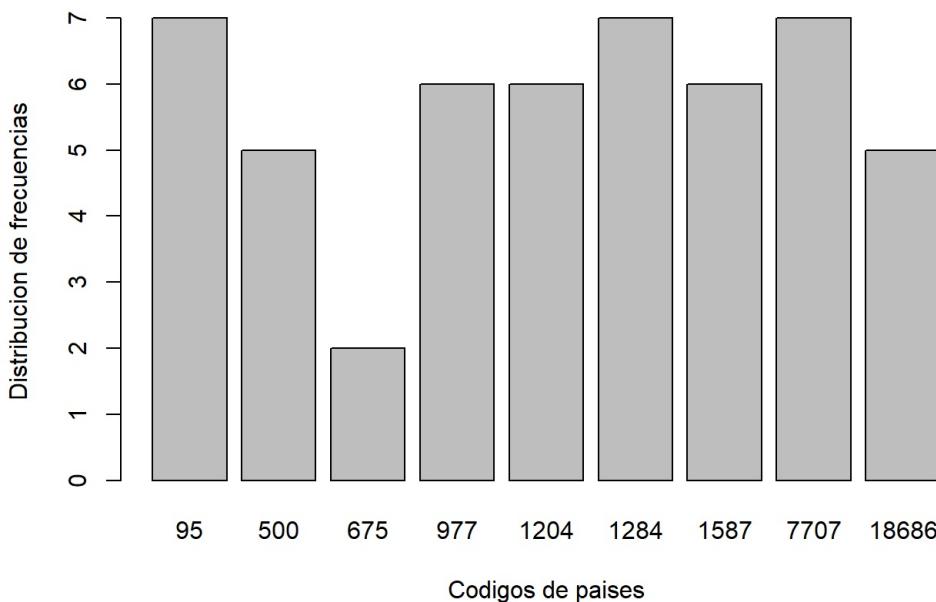
DF maximas para SMSs entrantes



En el diagrama que representa las frecuencias más bajas los mismos países que en la gráfica de la distribución de frecuencias mínimas de los SMSs en general. Las frecuencias también coinciden que nos permite hacer una conclusión que todas las interacciones (SMSs) menos frecuentes eran del tipo entrante.

```
barplot(codeSMSentr[codeSMSentr < 8], main = 'DF minimas para SMSs entrantes',
        xlab = 'Códigos de países', ylab = 'Distribucion de frecuencias')
```

DF minimas para SMSs entrantes

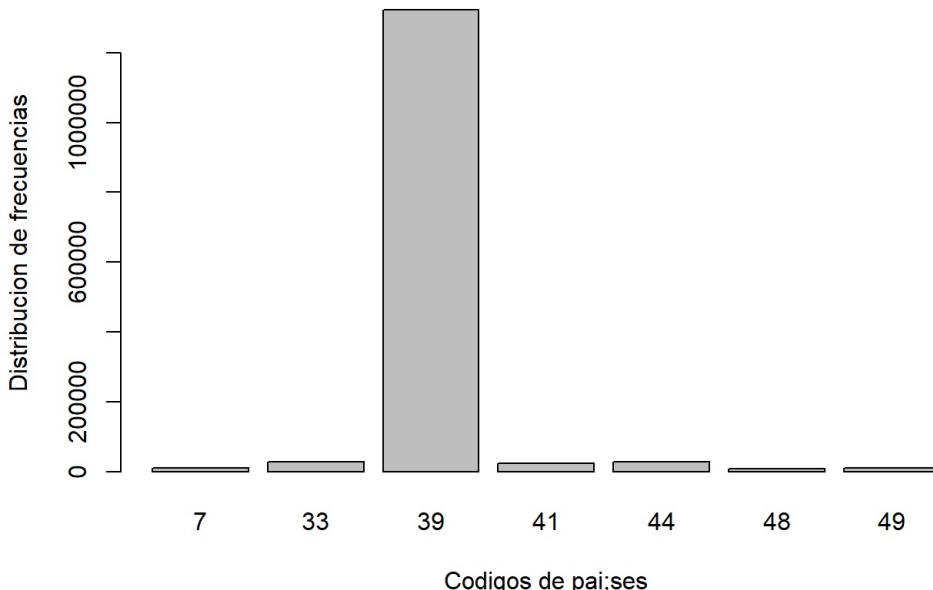


Distribución de frecuencias para los mensajes salientes

Con los mensajes salientes ponemos un límite más bajo de 8000. Aparecen Rusia (cod. 7) y Polonia (cod. 48), "desplazando" a los EE. UU. y Egipto.

```
codeSMSsal <- table(datos$`Country code`[!is.na(datos$`SMS-out activity`) & (datos$`Country code`>0)])
barplot(codeSMSsal[codeSMSsal > 8000], main = 'DF maximas para SMSs salientes',
        xlab = 'Códigos de países', ylab = 'Distribucion de frecuencias')
```

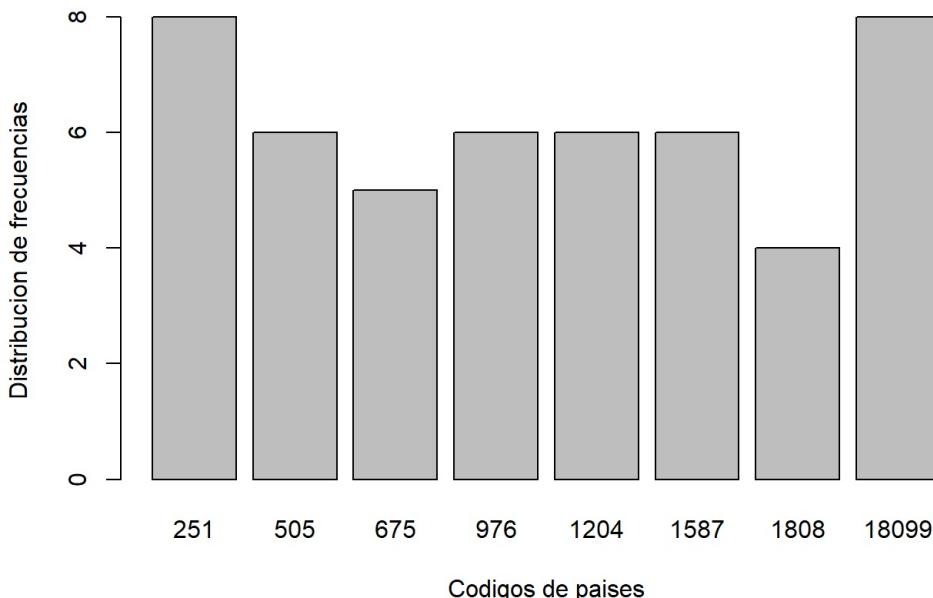
DF maximas para SMSs salientes



Los mensajes salientes menos frecuentes son: los 4 a Hawaii (cod. 1808); los 5 a Papúa Nueva Guinea (cod. 675); los 6 a Nicaragua (cod. 505), Mongolia (cod. 976), Canadá: Manítoba (cod. 1204), las islas Vírgenes Británicas (cod. 1587); los 8 a Etiopía (cod. 251) y a una región desconocida de México (cod. 18099).

```
barplot(codeSMSsal [codeSMSsal < 9], main = 'DF minimas para SMSs salientes',
        xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```

DF minimas para SMSs salientes

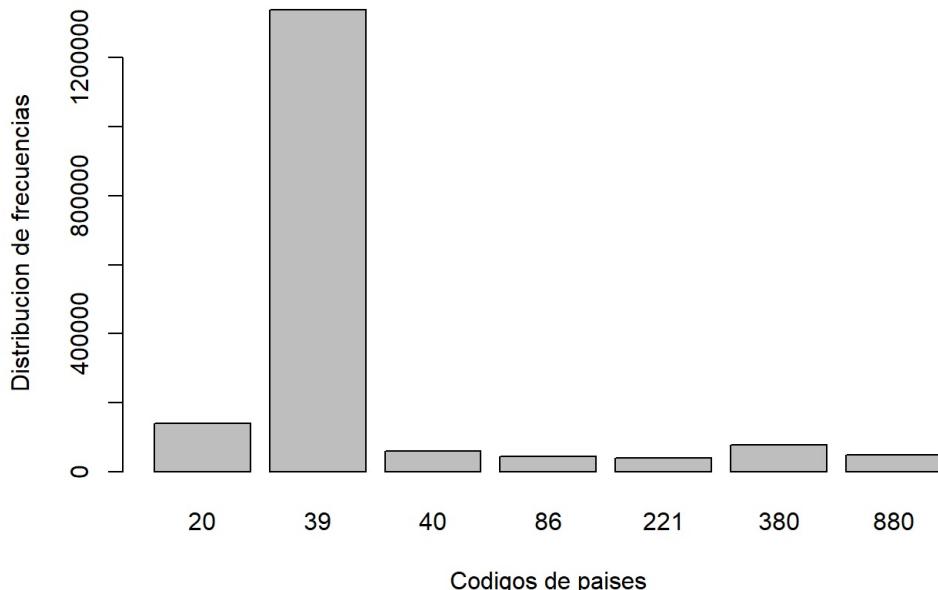


Distribución de frecuencias para las llamadas

Los países que ya conocemos son Italia (cod. 39) y Egipto (cod. 20). Mientras que los países "nuevos" son: Rumania (cod. 40), China (cod. 86), Senegal (cod. 221), Ucrania (cod. 380) y Bangladesh (cod. 880). Probablemente unos de estos países tienen el nivel de la cantidad de interacciones tan alto por no utilizar Internet.

```
codeCalls <- table(datos$`Country code`[(datos$Calls == TRUE) & (datos$`Country code`>0)])
barplot(codeCalls[codeCalls > 40000], main = 'DF maximas para llamadas',
        xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```

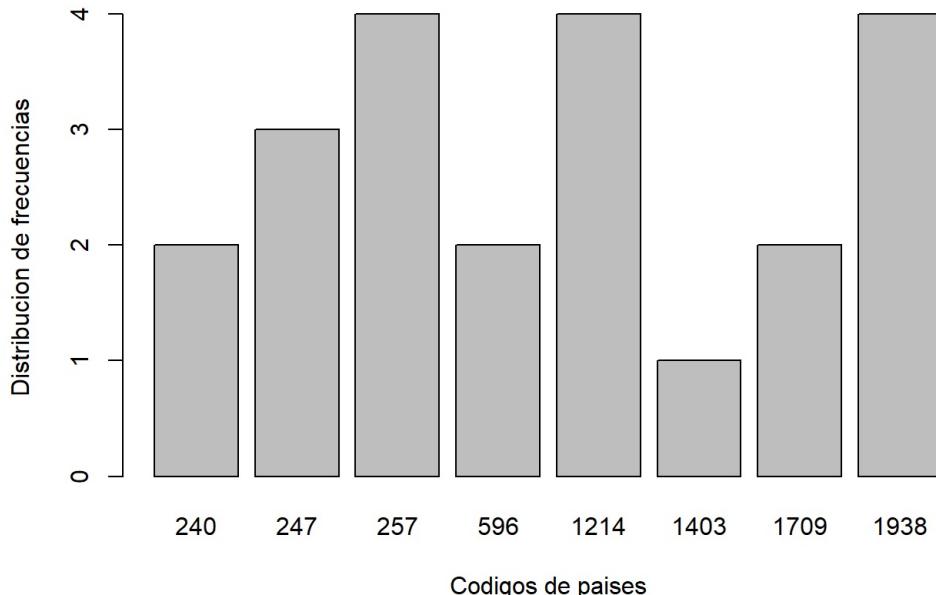
DF maximas para llamadas



Los códigos de las interacciones mínimas, de la frecuencia 1 hasta 4: Canada: Alberta (cod. 1403), Guinea Ecuatorial (cod. 240), Isla Martínica (cod. 596), Canada: Newfoundland and Labrador (cod. 1709), la Isla Ascension (cod. 247), Burundi (cod. 257), Estados Unidos: Texas (cod. 1214), Estados Unidos: Alabama (cod. 1938).

```
barplot(codeCalls[codeCalls < 5], main = 'DF minimas para llamadas',
        xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```

DF minimas para llamadas

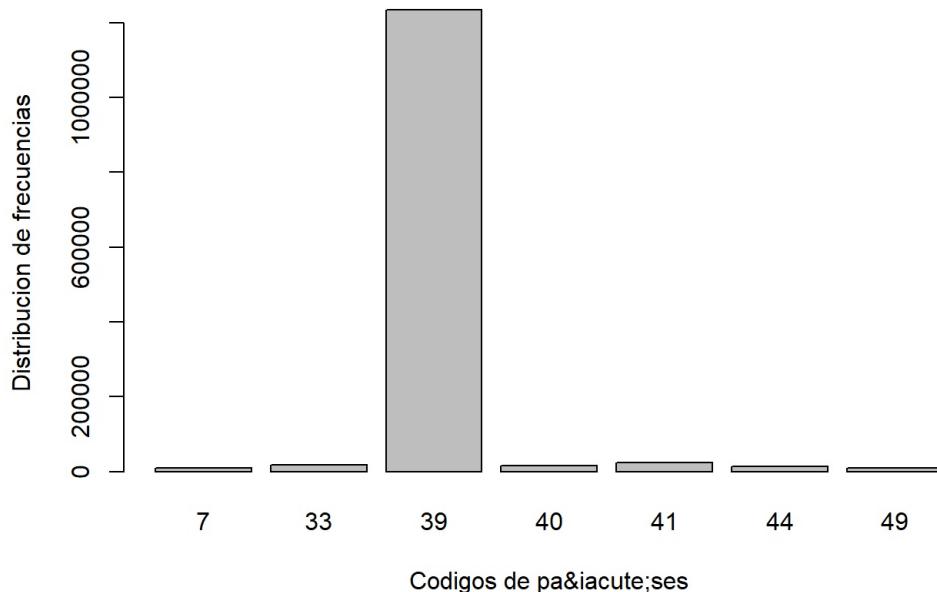


Distribución de frecuencias para las llamadas entrantes

Las frecuencias de las llamadas entrantes son más bajas que las de las llamadas en general y los países a que corresponde el código se han vuelto a cambiar y en este caso parecen más en los que hemos visto en las representaciones de las frecuencias de SMS y interacciones generales. El código más frecuente es el de Italia (cod. 39), le siguen Francia (cod. 33) y Suiza (cod. 41), luego - Rumanía (cod. 40) y Reino Unido (cod. 44). Los últimos dos códigos que apenas superan el límite de 9000 son de Alemania (cod. 49) y de Rusia (cod. 7).

```
codeCallsentr <- table(datos$`Country code`[!is.na(datos$`Call-in activity`) & (datos$`Country code`>0)])
barplot(codeCallsentr[codeCallsentr > 9000], main = 'DF maximas para llamadas entrantes',
        xlab = 'Códigos de países', ylab = 'Distribución de frecuencias')
```

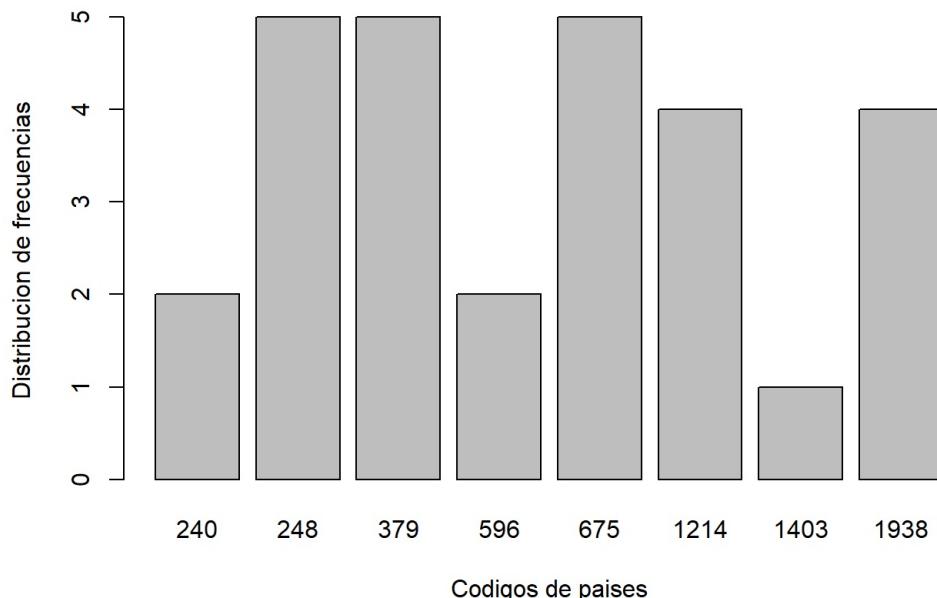
DF maximas para llamadas entrantes



En la distribución de frecuencias más bajas de llamadas entrantes están presentados casi los mismos países que en la distribución de frecuencias de las llamadas total con algunas diferencias: aparecen Seychelles (cod. 248), Papúa Nueva Guinea (cod. 675), Vaticano (cod. 379) y desaparecen la Isla Ascension (cod. 247), Burundi (cod. 257), Canada: Newfoundland and Labrador (cod. 1709).

```
barplot(codeCallsentr[codeCallsentr < 6], main = 'DF minimas para llamadas entrantes',
        xlab = 'Códigos de países', ylab = 'Distribucion de frecuencias')
```

DF minimas para llamadas entrantes

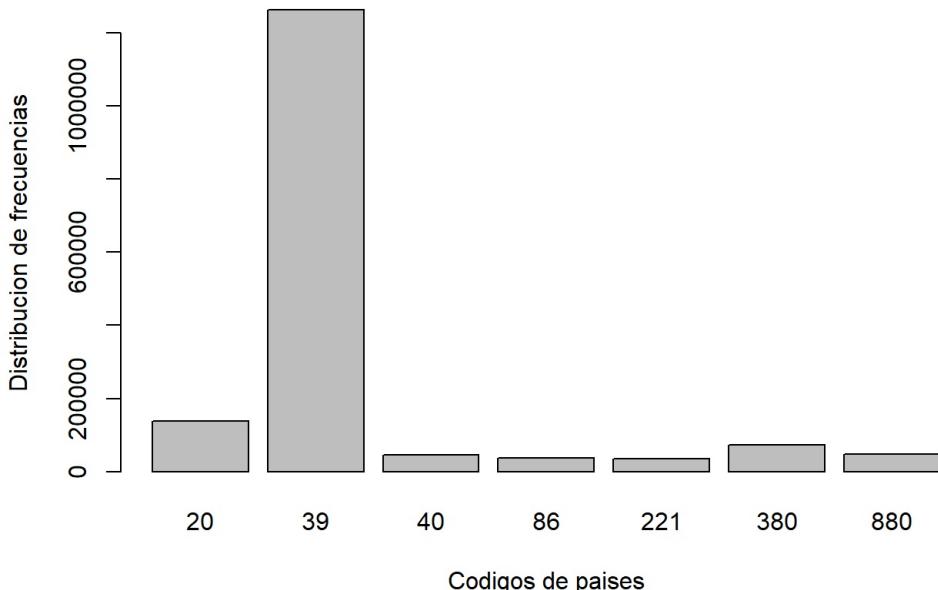


Distribución de frecuencias para las llamadas salientes

Se nota que las frecuencias de las llamadas salientes son más altas que las de las llamadas entrantes y los códigos coinciden con los códigos que hemos visto en la representación de las frecuencias de las llamadas en general.

```
codeCallssal <- table(datos$`Country code`[!is.na(datos$`Call-out activity`) & (datos$`Country code`>0)])
barplot(codeCallssal[codeCallssal > 34000], main = 'DF maximas para llamadas salientes',
        xlab = 'Códigos de países', ylab = 'Distribucion de frecuencias')
```

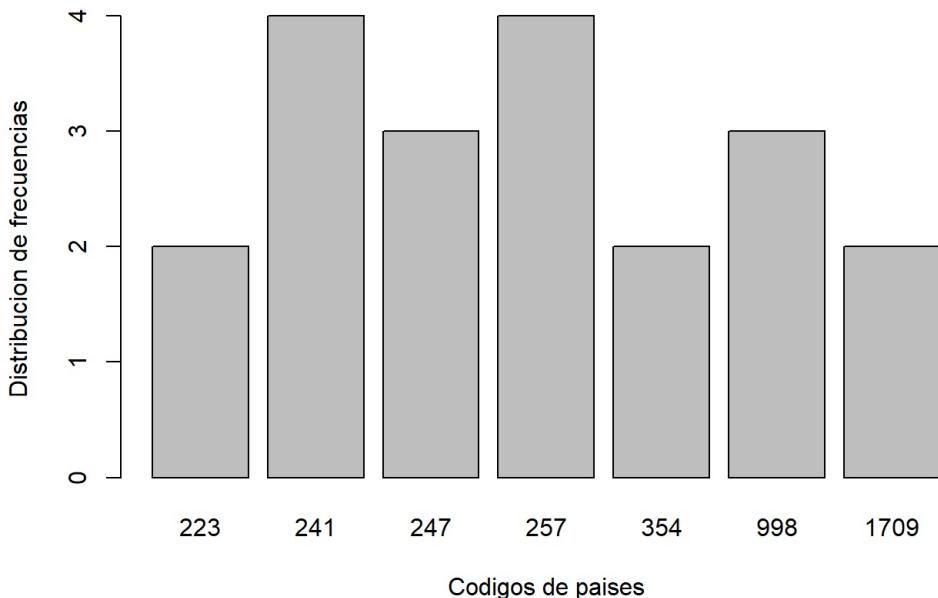
DF maximas para llamadas salientes



Los códigos de los países del nivel de interacciones minimal, en el orden creciendo: Malí (cod. 223), Islandia (cod. 354), Canada: Newfoundland and Labrador (cod. 1709), Uzbekistán (cod. 998), Gabón (cod. 241), Burundi (cod. 257).

```
barplot(codeCallssal[codeCallssal < 6], main = 'DF minimas para llamadas salientes',
        xlab = 'Códigos de países', ylab = 'Distribucion de frecuencias')
```

DF minimas para llamadas salientes

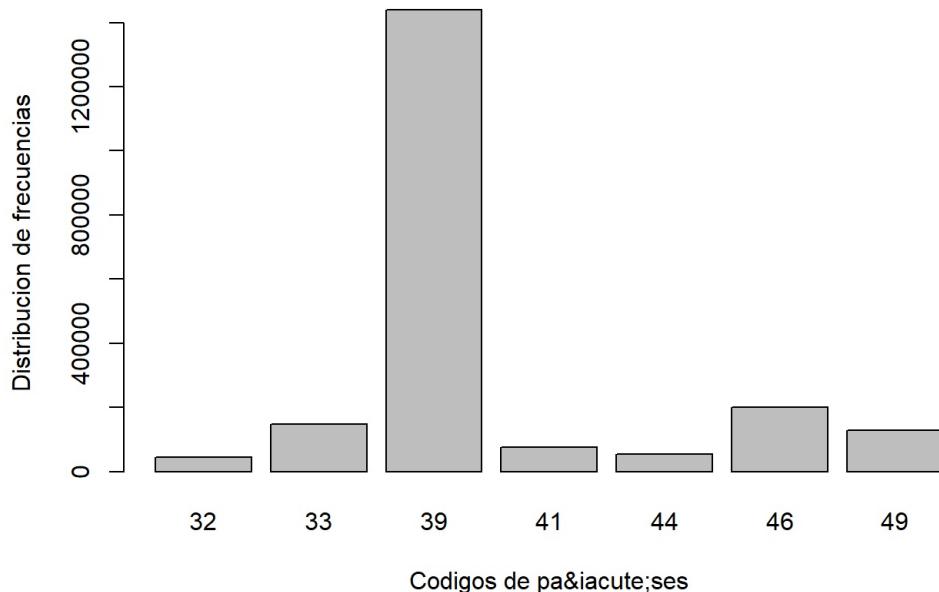


Distribución de frecuencias para el tráfico de Internet

Los teléfonos italianos han utilizado la cantidad máxima del tráfico (cod. 39), los de Francia (cod. 33), Suecia (cod. 46) y Alemania (cod. 49) han utilizado mucho menos, pero los códigos de estos países siguen formando el grupo de las frecuencias de interacción más alta con los códigos de Bélgica (cod. 32), Suiza (cod. 41) y Reino Unido (cod. 44).

```
codeInternet <- table(datos$`Country code`[(datos$Internet == TRUE) & (datos$`Country code`>0)])
barplot(codeInternet[codeInternet > 31000], main = 'DF maximas para trafico de Internet',
        xlab = 'Códigos de países', ylab = 'Distribucion de frecuencias')
```

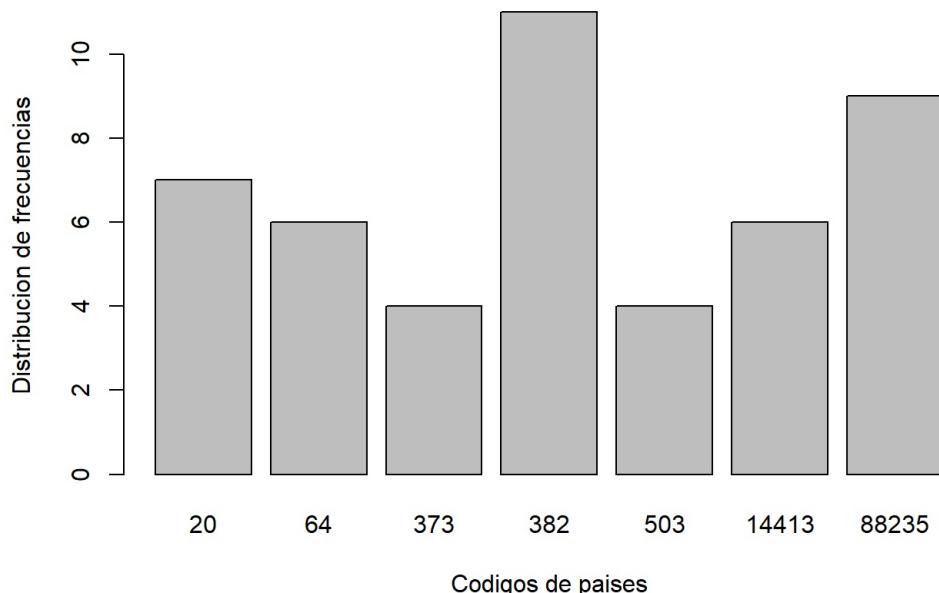
DF maximas para trafico de Internet



Aparece el código de Egipto, el país que era uno de los líderes de SMSs y llamadas. Los otros códigos pertenecen a: Moldavia (cod. 373), El Salvador (cod. 503) con la freq. 4, Nueva Zelanda (cod. 64) y Bermuda (cod. 14413) con freq. 6; código 88235 pertenece a "Jasper Wireless satellite services" (GPS?) con freq. 9, Montenegro (cod. 382) con freq. 10.

```
barplot(codeInternet[codeInternet < 12], main = 'DF minimas para trafico de Internet',
        xlab = 'Códigos de países', ylab = 'Distribucion de frecuencias')
```

DF minimas para trafico de Internet

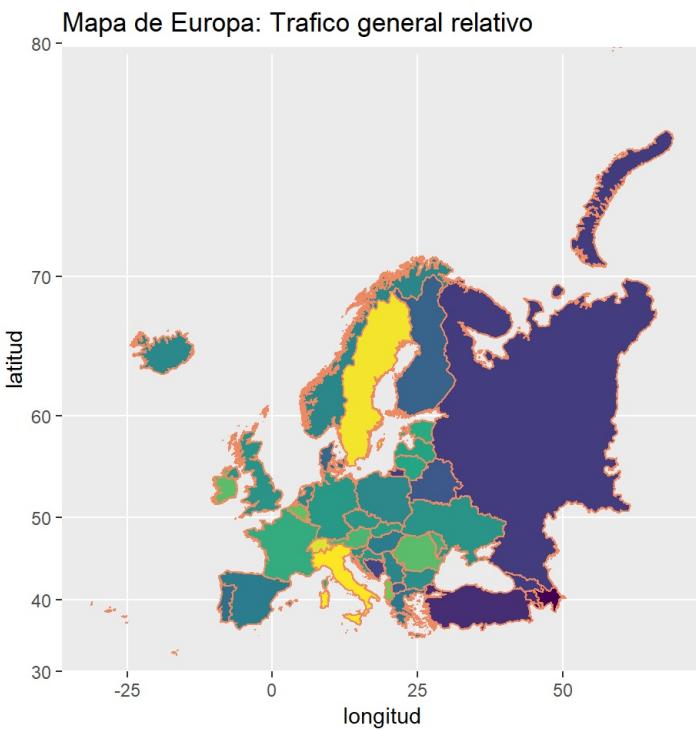
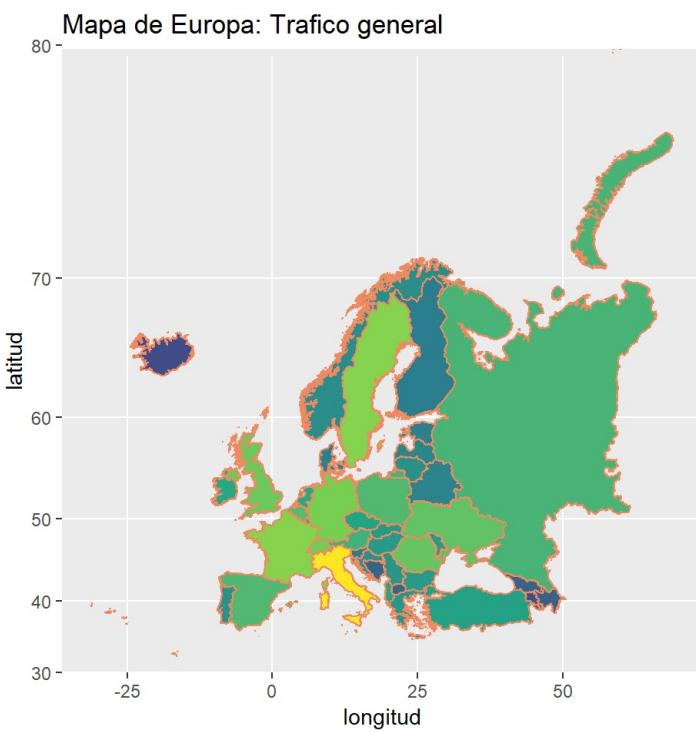


Representación cartográfica de las distribuciones de frecuencias

Para contribuir a que los colores de los mapas sean más informativos, hemos considerado las frecuencias en escala logarítmica. Como el rango que abarcan los valores es grande, una escala logarítmica proporciona un medio mejor de visualización de los datos. La transformación se hace mediante la función `log()`.

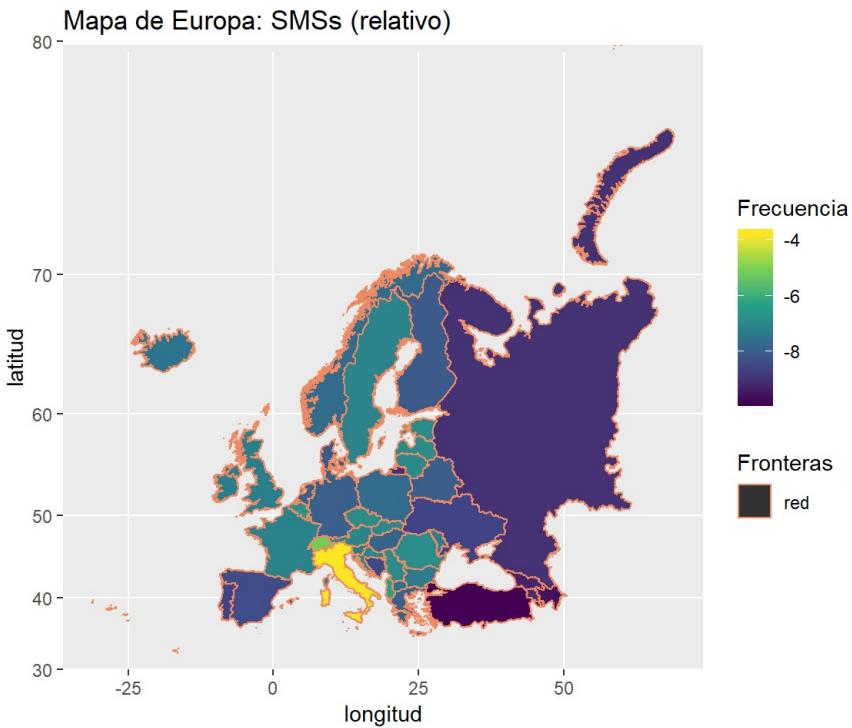
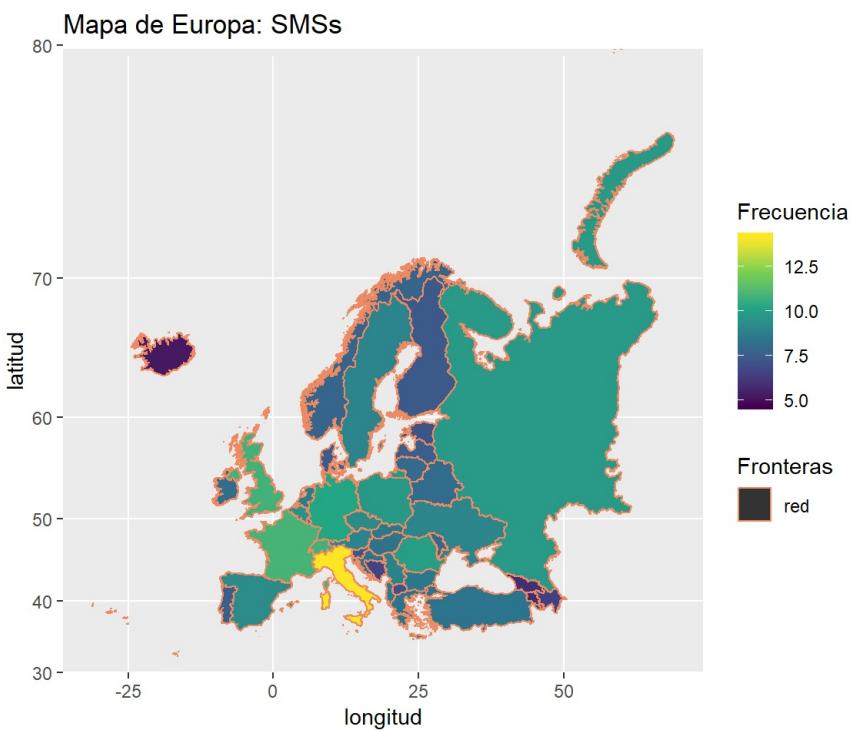
Tráfico general

Podemos ver que las interacciones más frecuentes que tenían lugar pertenecen a Italia. Los países de Europa Central tienen el nivel de la frecuencia parecido, ya que aparecen regiones con el nivel más bajo: Balcanes, Portugal, países del Báltico y países nórdicos excepto de Suecia. Además, este mapa representa que la cantidad de interacciones (por población) relativamente grandes pertenecen a: Rumanía, Austria, Bélgica, Irlanda, Albania y Letonia, Lituania, Estonia, Moldavia y Francia.



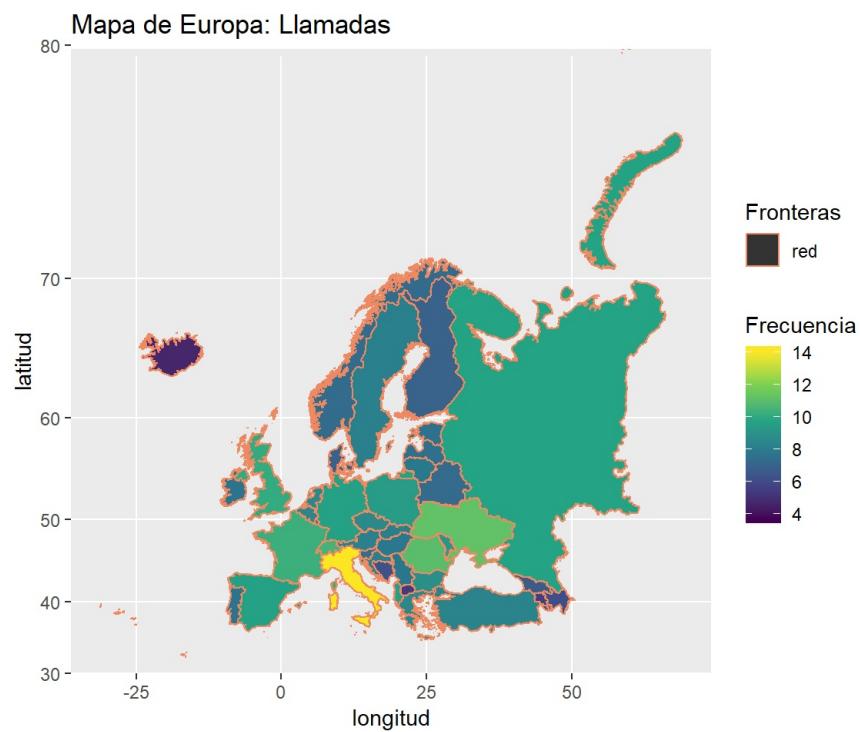
Tráfico de SMSs

Los países con números más activos en mandar mensajes son: Reino Unido, Francia y Suiza. A su vez, Suecia demuestra la actividad más alta entre todos los países nórdicos y entre toda la Europa si comparamos nuestros datos respecto a la población. En estos términos también sobresalen los países Bálticos (excepto de Polonia), Balcanes, República Checa, Austria, Eslovaquia, Francia, Bélgica y Reino Unido.

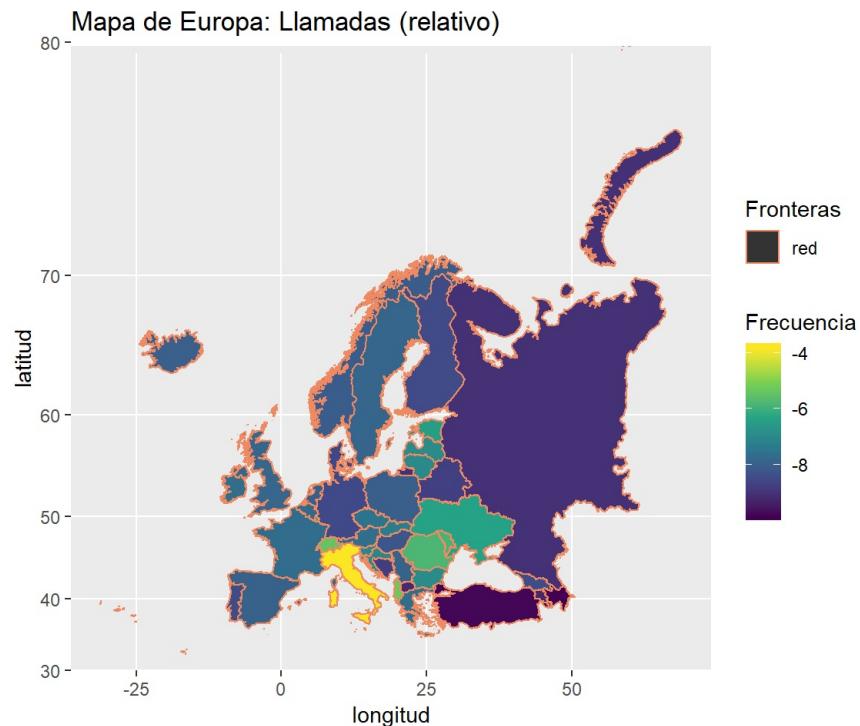


Tráfico de llamadas

El mapa representa que en general la gente hace más llamadas que manda los mensajes. Aparte de Italia, más llamadas fueron establecidas con Ucrania y Rumania, un poco menos - con Francia, Suiza, Reino Unido, Alemania, España, Polonia y Rusia.

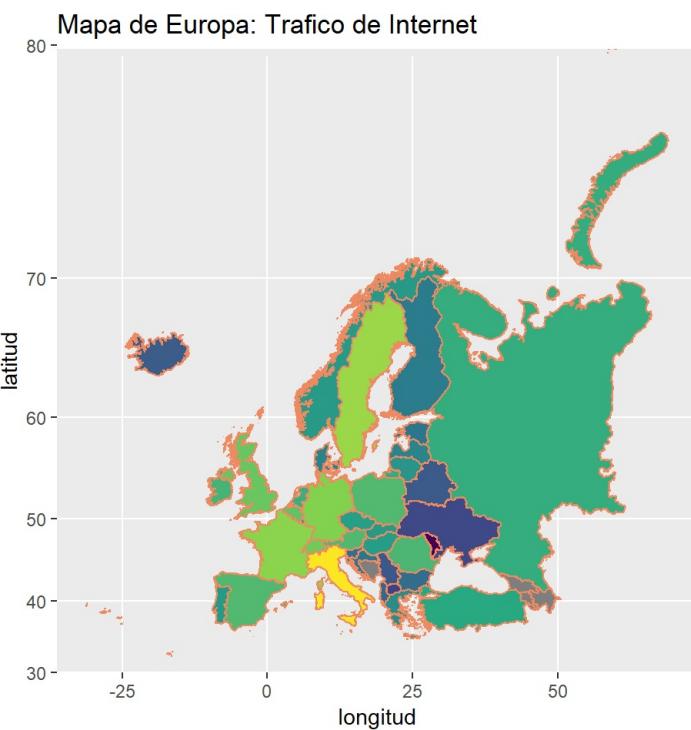


La representación con respecto a la población nos muestra que comparando con la cantidad de la gente que vive en cada país la gran cantidad hicieron los habitantes de Suiza, Rumanía, Ucrania (coincide con el mapa anterior) y también Albania, Moldavia y Letonia, Lituania, Estonia.

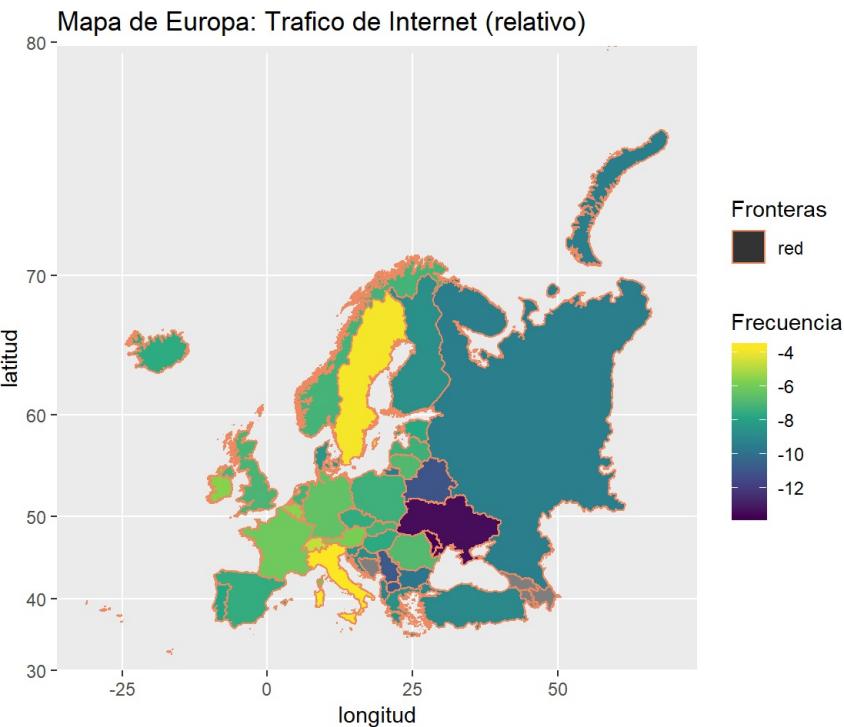


Tráfico de Internet

Las regiones que se diferencian son: Islandia, Balcanes, Moldavia, Ucrania, Bielorrusia.



El mapa que fue pintado utilizando los datos de población muestra que, comparando con la cantidad de habitantes total, el tráfico decrece moviendo del oeste al este de Europa. Como en la representación general, Suecia notablemente sobresale que significa que la cantidad del tráfico (número de interacciones) de Internet influye mucho en la cantidad (número de interacciones) general.



Análisis por celdas

Generación del tráfico por celdas

Creamos una nueva hoja de datos en la que incluimos la suma y la media del tráfico para cada tipo por celdas. Para añadir la información por celdas se creó una matriz y mediante un bucle la rellenamos con los vectores correspondientes a cada celda. El sumatorio se consigue con la función `colSums()` (en nuestro caso es la suma por columnas) y el promedio - con `colMeans()`.

```

matriz <- matrix(,0,11) #Creamos una matriz vacia para la suma de trafico

for (i in 1:10000)
{
  datos[datos$`Square id` == i,]
  vector <- c(i,colSums(datos[datos$`Square id` == i, 4:8], na.rm = TRUE),
             colMeans(datos[datos$`Square id` == i, 4:8], na.rm = TRUE))
  matriz <- rbind(matriz, vector)
} #rellenamos la matriz con la suma y el promedio del trafico
colnames(matriz) <- c("Square.id", "Suma.SMS.in", "Suma.SMS.out", "Suma.Call.in",
                      "Suma.Call.out", "Suma.Internet", "Media.SMS.in", "Media.SMS.out",
                      "Media.Call.in", "Media.Call.out", "Media.Internet")
suma <- data.frame(matriz) #hacemos una hoja de datos

```

```
summary(suma)
```

```

##   Square.id      Suma.SMS.in      Suma.SMS.out      Suma.Call.in
## Min. : 1       Min. : 0.00       Min. : 0.00       Min. : 0.0
## 1st Qu.: 2501  1st Qu.: 85.56     1st Qu.: 51.74     1st Qu.: 49.3
## Median : 5000  Median : 225.79    Median : 130.72    Median : 129.5
## Mean   : 5000  Mean   : 460.64    Mean   : 282.24    Mean   : 262.3
## 3rd Qu.: 7500  3rd Qu.: 506.43    3rd Qu.: 307.41    3rd Qu.: 300.6
## Max.  :10000   Max.  :16448.39    Max.  :8343.75    Max.  :7650.5
##
##   Suma.Call.out      Suma.Internet      Media.SMS.in
## Min. : 1.311       Min. : 47.79       Min. : 0.008189
## 1st Qu.: 60.013    1st Qu.: 1666.19    1st Qu.: 0.328543
## Median : 158.197   Median : 4224.11    Median : 0.830625
## Mean   : 319.676   Mean   : 8247.92    Mean   : 1.388215
## 3rd Qu.: 360.538   3rd Qu.: 9102.35    3rd Qu.: 1.741639
## Max.  :8974.389   Max.  :237230.39   Max.  :18.139612
## NA's   :2          NA's   :2          NA's   :2
##   Media.SMS.out      Media.Call.in      Media.Call.out
## Min. : 0.00611     Min. : 0.00967     Min. : 0.01016
## 1st Qu.: 0.35004   1st Qu.: 0.35915     1st Qu.: 0.33896
## Median : 0.84040   Median : 0.90410     Median : 0.78660
## Mean   : 1.53265   Mean   : 1.50409     Mean   : 1.19839
## 3rd Qu.: 1.88324   3rd Qu.: 1.91444     3rd Qu.: 1.51704
## Max.  :44.61333   Max.  :22.84349    Max.  :17.45990
## NA's   :2          NA's   :2          NA's   :2
##   Media.Internet
## Min. : 0.2158
## 1st Qu.: 7.2871
## Median : 18.9801
## Mean   : 34.1939
## 3rd Qu.: 41.4583
## Max.  :413.1590
## NA's   :2

```

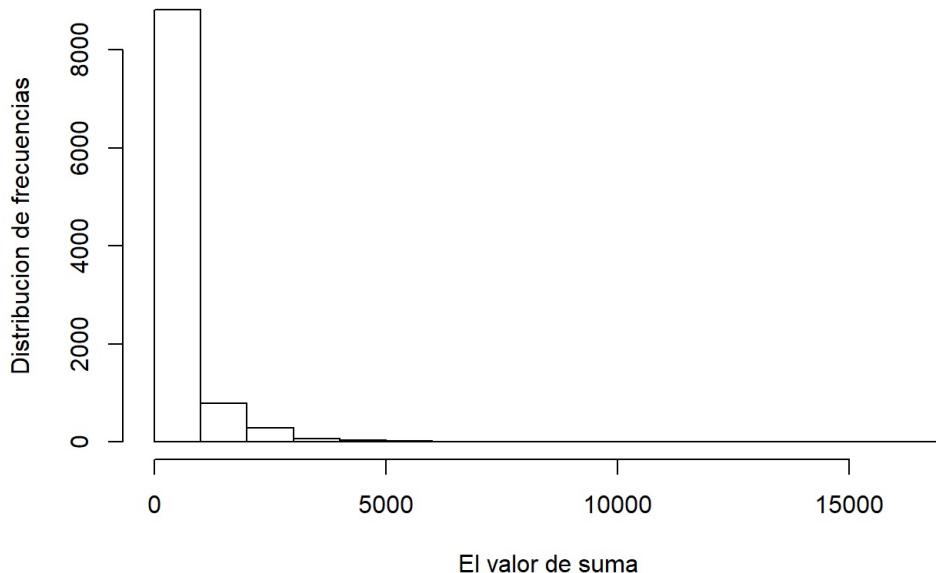
Tráfico total

Todos los histogramas salen muy asimétricos a la derecha dado que hay muchas celdas en la ciudad con la actividad de interacciones baja mientras que existen unas cuantas celdas (que casi no podemos ver en la gráfica) con el nivel de actividad extraordinariamente alto. Esto fenómeno es un ejemplo de un “Efecto Mateo”: “Porque a cualquiera que tiene, se le dará, y tendrá más; pero al que no tiene, aun lo que tiene le será quitado”.

Más probable que los valores de suma más altos pertenezcan a las celdas centrales de Milán. Como hemos visto antes, en referencia a la suma, el tráfico de Internet ocupa el primer lugar de cantidad (por otro tipo de codificación), luego la cantidad de SMSs entrantes avanza un poco la cantidad de otros tipos de interacciones.

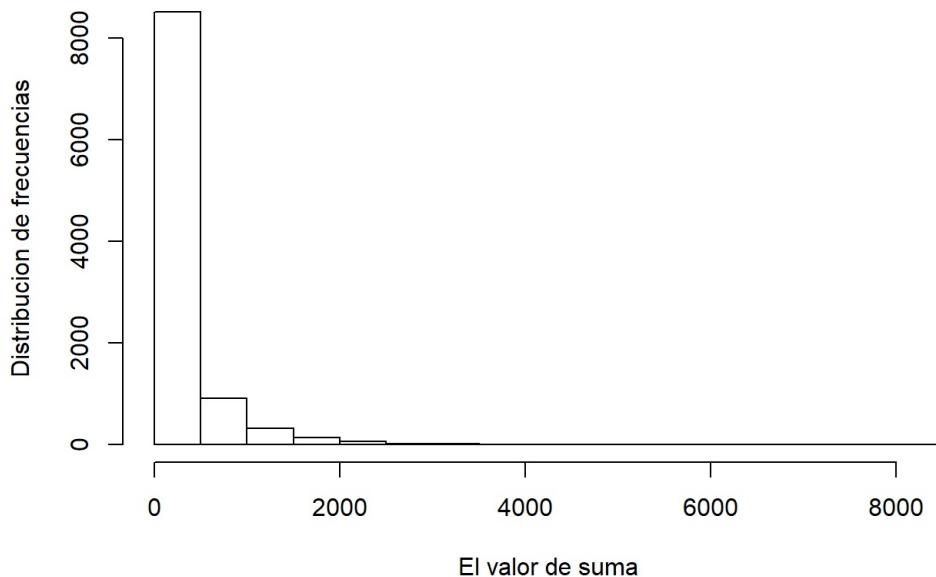
```
hist(suma$Suma.SMS.in, main = 'DF de total: SMSs entrantes', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

DF de total: SMSs entrantes



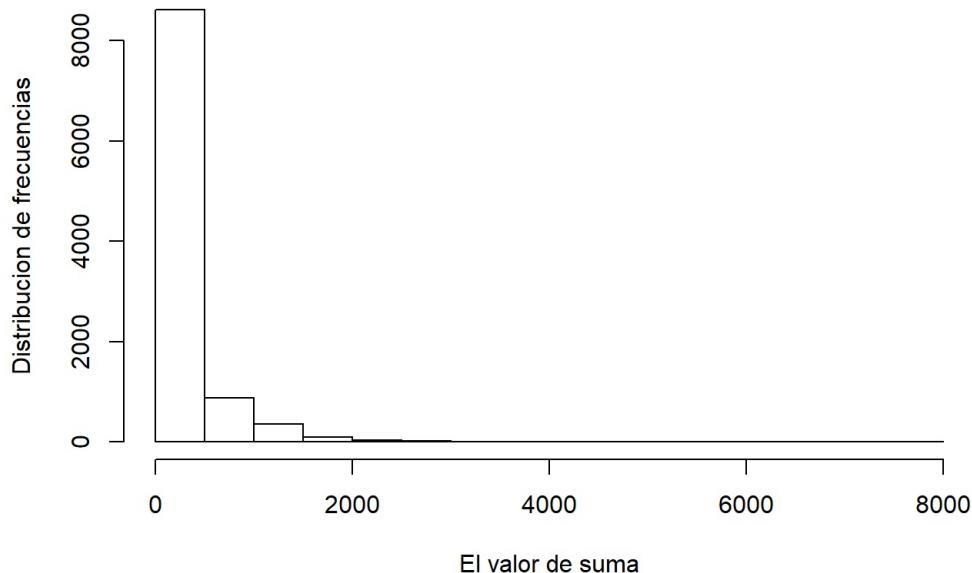
```
hist(suma$Suma.SMS.out, main = 'DF de total: SMSs salientes', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

DF de total: SMSs salientes



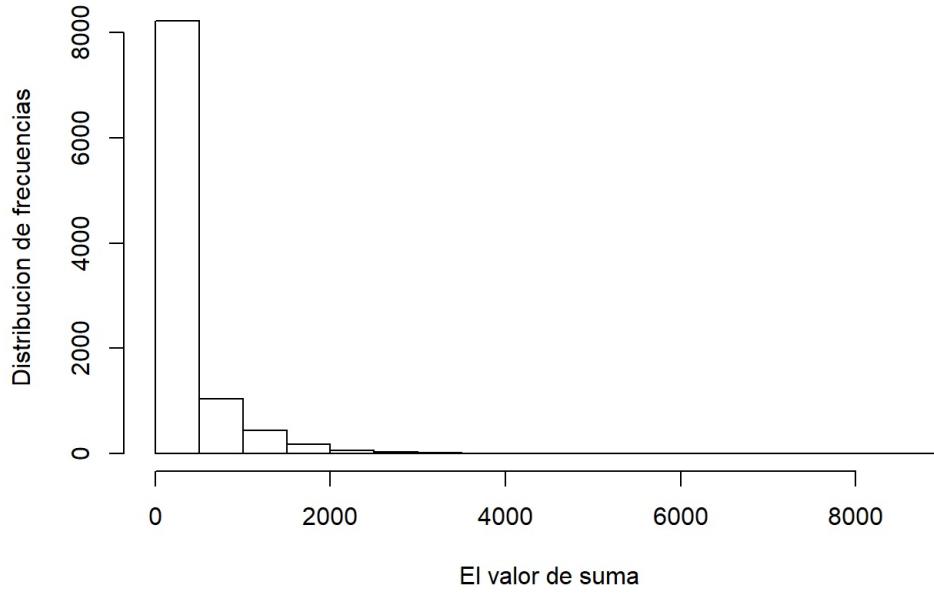
```
hist(suma$Suma.Call.in, main = 'DF de total: llamadas entrantes', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

DF de total: llamadas entrantes



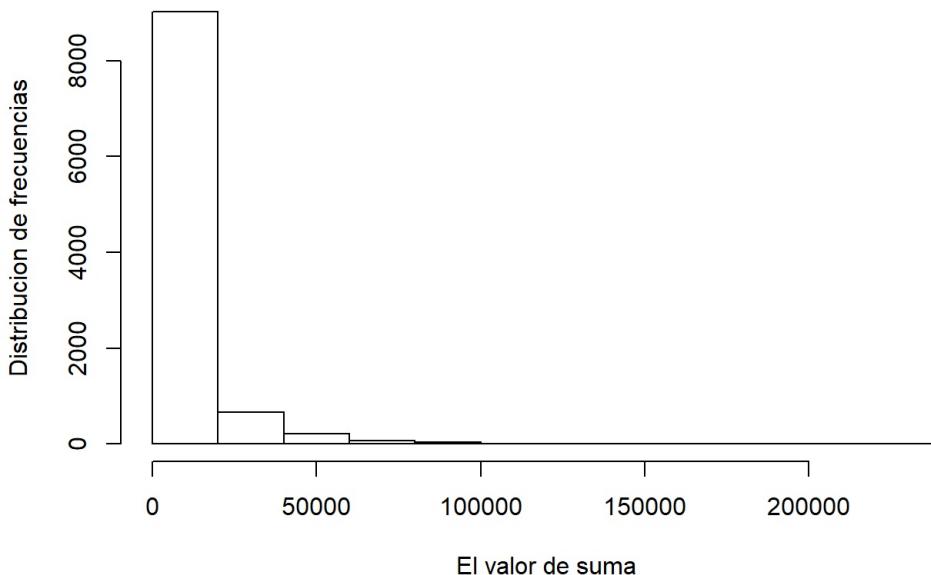
```
hist(suma$Suma.Call.out, main = 'DF de total: llamadas salientes', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

DF de total: llamadas salientes



```
hist(suma$Suma.Internet, main = 'DF de total: Internet', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

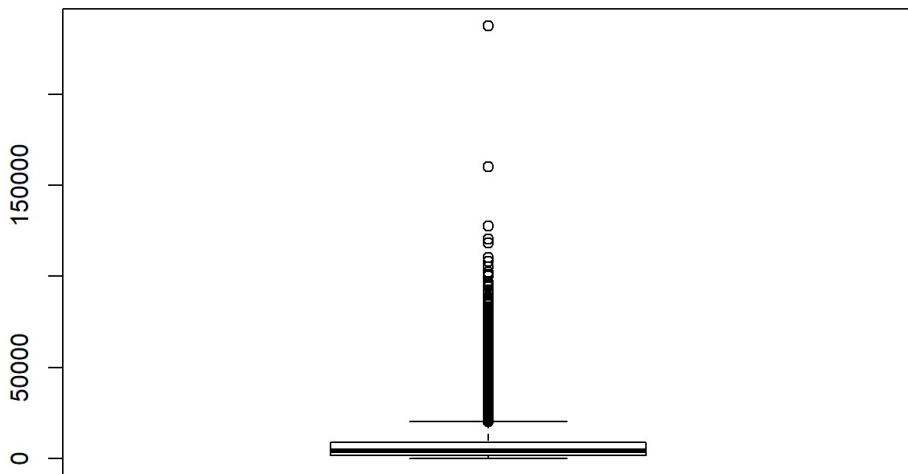
DF de total: Internet



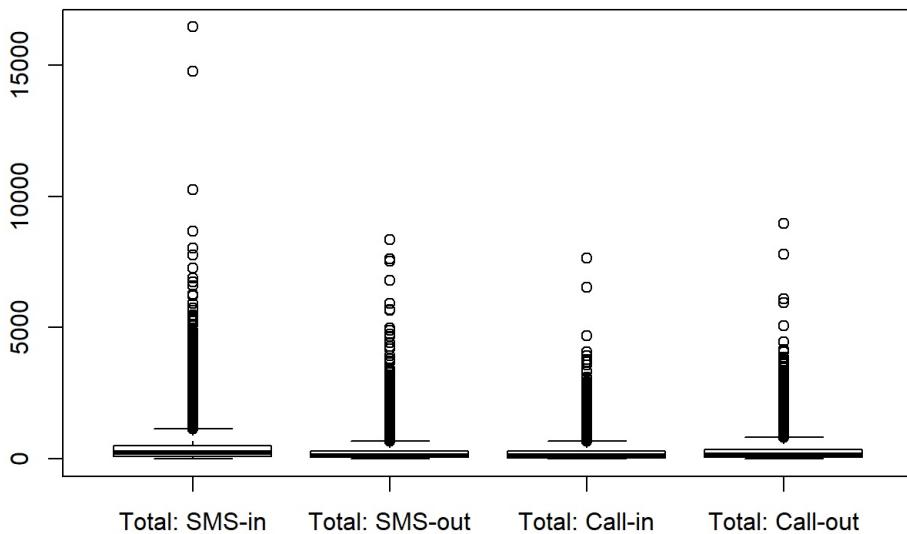
Los valores del tráfico total de mensajes y llamadas son comparables, por eso es posible juntar los cuatro diagramas en una figura del diagrama de cajas. Cómo en la gran mayoría de los datos se acerca al cero y por eso tanto la media, como los percentiles 25%, 50% y 75% se encuentran relativamente cerca al cero.

```
Ssmsin <- suma$Suma.SMS.in  
Ssmsout <- suma$Suma.SMS.out  
Scallin <- suma$Suma.Call.in  
Scallout <- suma$Suma.Call.out  
Sinter <- suma$Suma.Internet  
boxplot(Sinter, main= "Total: Internet")
```

Total: Internet



```
boxplot(Ssmsin, Ssmsout, Scallin, Scallout, names = c("Total: SMS-in", "Total: SMS-out", "Total: Call-in", "Total : Call-out"))
```



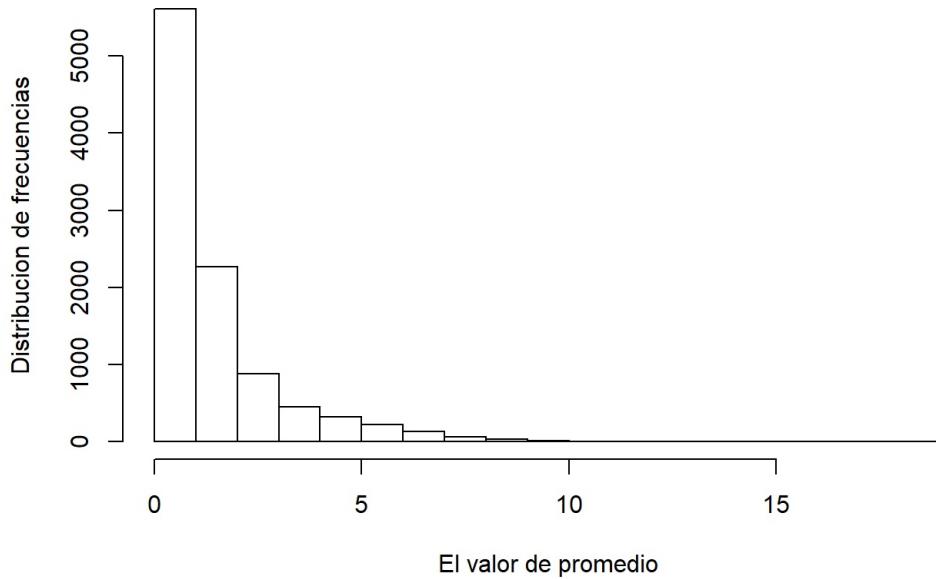
Tráfico promedio

Ahora analizamos la media del tráfico por interacción; es decir, si antes analizábamos la suma, Y, del tráfico de X interacciones, ahora analizamos el promedio, Y/X . En este caso, todas las gráficas siguen siendo asimétricas a la derecha pero algunas de ellas (SMSs entrantes, llamadas salientes, Internet) ya siguen a la distribución más homogénea debido a los valores de promedio o a los intervalos más pequeños.

Los valores del tráfico del Internet son visiblemente más altos que los otros, pero comparando los valores de llamadas y mensajes se nota que el valor promedio de SMSs salientes alcanza un nivel más alto que todos los demás aunque el valor de total no se diferencia tanto entre todos. Eso se puede explicar por la cantidad diferente de las interacciones por lo que la suma pueda ser igual. Así la mayor dispersión pertenece a los valores de SMSs salientes y a las llamadas entrantes.

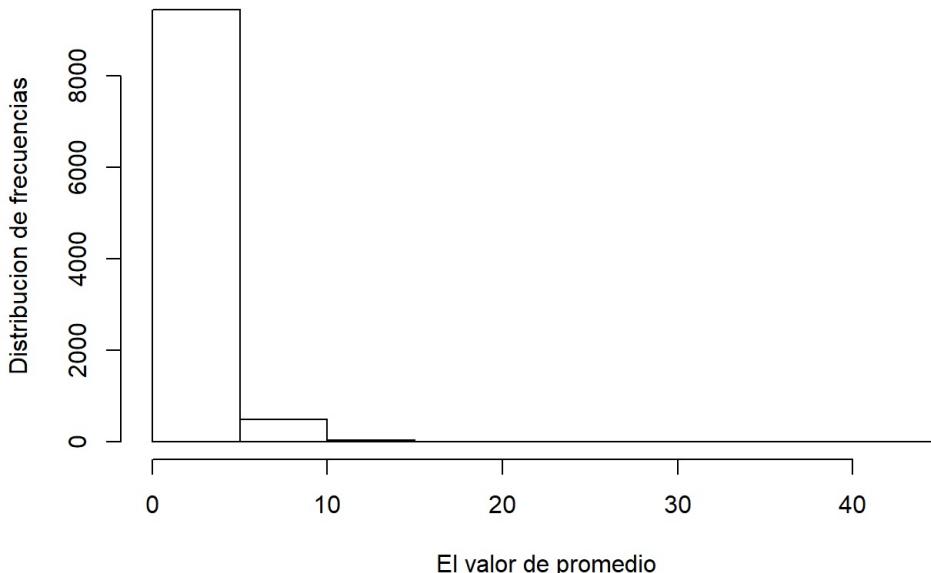
```
hist(suma$Media.SMS.in, main = 'DF de promedio: SMSs entrantes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio: SMSs entrantes



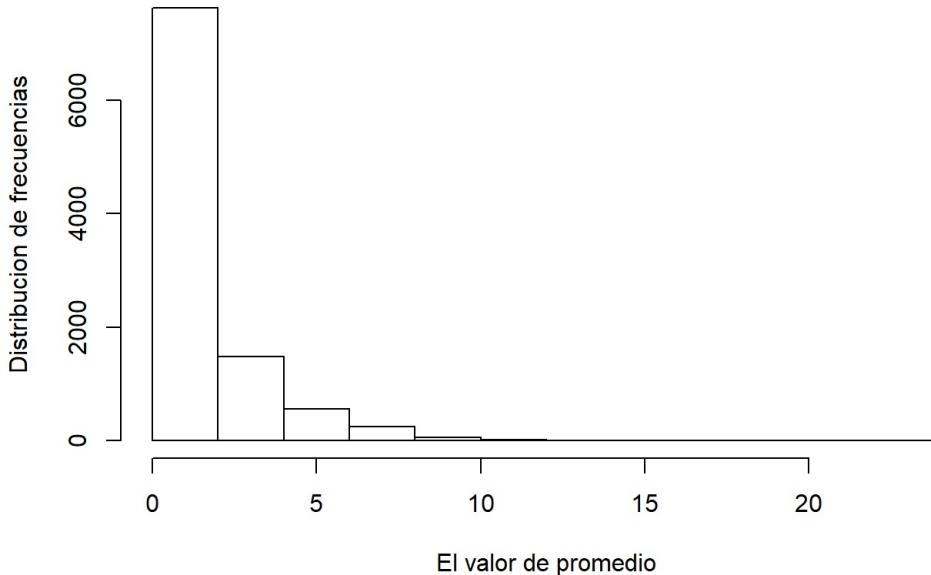
```
hist(suma$Media.SMS.out, main = 'DF de promedio: SMSs salientes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio: SMSs salientes



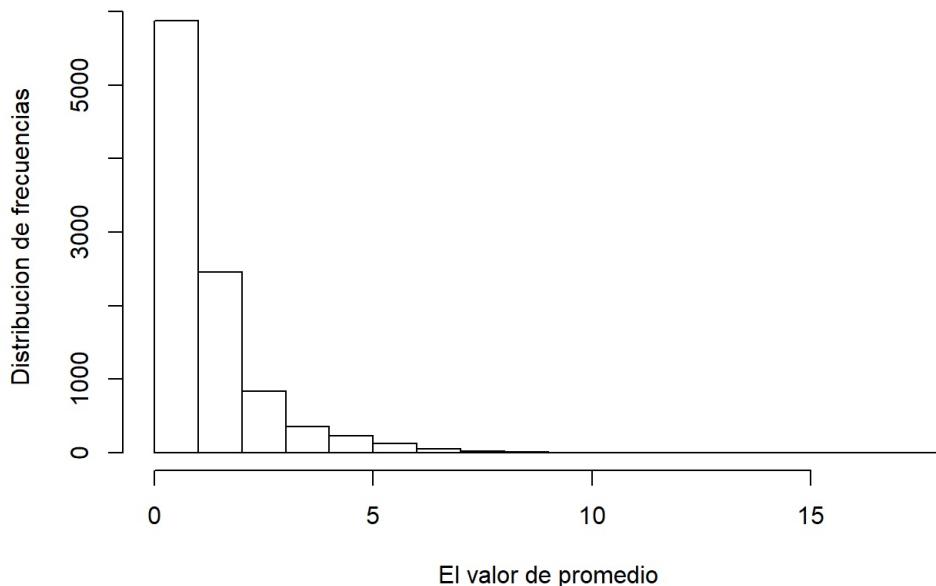
```
hist(suma$Media.Call.in, main = 'DF de promedio: llamadas entrantes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio: Llamadas entrantes



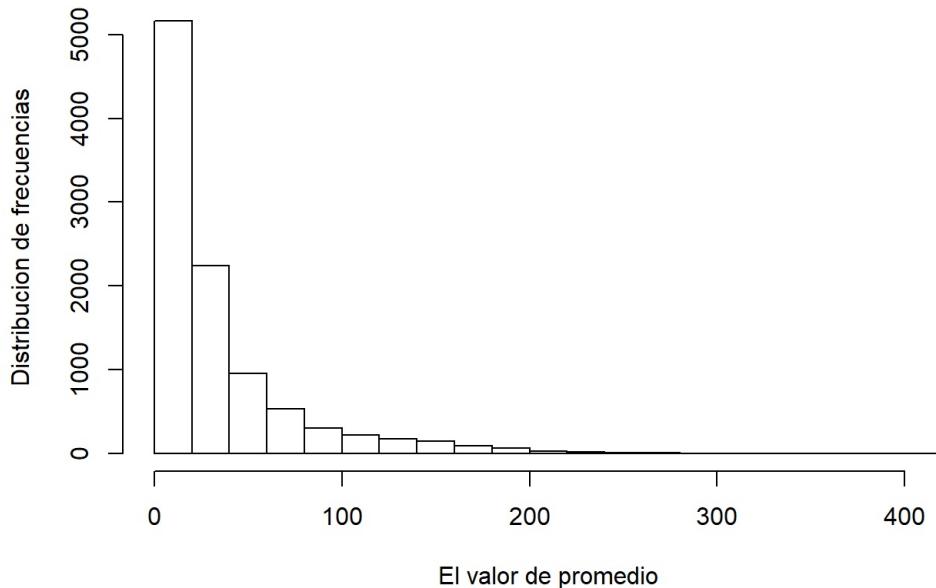
```
hist(suma$Media.Call.out, main = 'DF de promedio: llamadas salientes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio: llamadas salientes



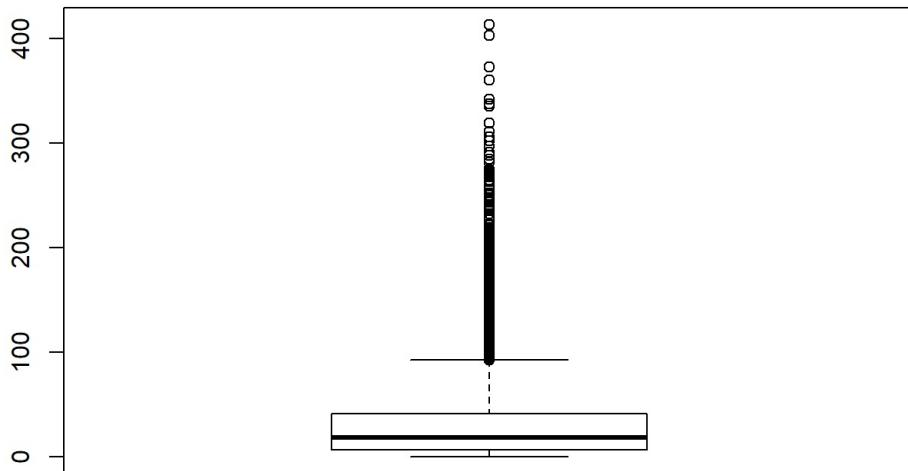
```
hist(suma$Media.Call.out, main = 'DF de promedio: llamadas salientes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio: Internet

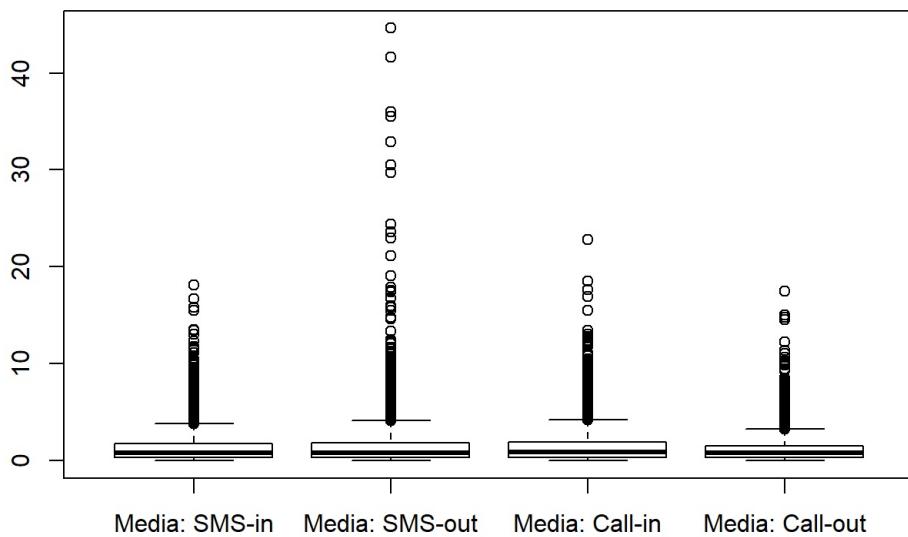


```
Msmsin <- suma$Media.SMS.in  
Msmsout <- suma$Media.SMS.out  
Mcallin <- suma$Media.Call.in  
Mcallout <- suma$Media.Call.out  
Minter <- suma$Media.Internet  
boxplot(Minter, main= "Media: Internet")
```

Media: Internet



```
boxplot(Msmsin, Msmsout, Mcallin, Mcallout, names = c("Media: SMS-in", "Media: SMS-out", "Media: Call-in", "Media: Call-out"))
```



Medidas de posición, dispersión y forma

Podemos ver que en todos los casos el coeficiente de variación supera a 100% que indica la variación muy grande y encima, siendo el coeficiente de asimetría positivo y mayor que 0.9, podemos concluir que todas las distribuciones varían mucho y son asimétricas a la derecha.

```

sumaSMSin <- c(mean(suma$Suma.SMS.in, na.rm= TRUE), quantile(suma$Suma.SMS.in, probs = c(0.25, 0.5, 0.75), na.rm = TRUE), sd(suma$Suma.SMS.in, na.rm= TRUE)/mean(suma$Suma.SMS.in, na.rm= TRUE), skewness(suma$Suma.SMS.in, na.rm = TRUE))
sumaSMSout <- c(mean(suma$Suma.SMS.out, na.rm= TRUE), quantile(suma$Suma.SMS.out, probs = c(0.25, 0.5, 0.75), na .rm= TRUE), sd(suma$Suma.SMS.out, na.rm= TRUE)/mean(suma$Suma.SMS.out, na.rm= TRUE), skewness(suma$Suma.SMS.out, na.rm= TRUE))
sumaCallin <- c(mean(suma$Suma.Call.in, na.rm= TRUE), quantile(suma$Suma.Call.in, probs = c(0.25, 0.5, 0.75), na .rm= TRUE), sd(suma$Suma.Call.in, na.rm= TRUE)/mean(suma$Suma.Call.in, na.rm= TRUE), skewness(suma$Suma.Call.in, na.rm= TRUE))
sumaCallout <- c(mean(suma$Suma.Call.out, na.rm= TRUE), quantile(suma$Suma.Call.out, probs = c(0.25, 0.5, 0.75), na .rm= TRUE), sd(suma$Suma.Call.out, na.rm= TRUE)/mean(suma$Suma.Call.out, na.rm= TRUE), skewness(suma$Suma.Call .out, na.rm= TRUE))
sumaInternet <- c(mean(suma$Suma.Internet, na.rm= TRUE), quantile(suma$Suma.Internet, probs = c(0.25, 0.5, 0.75) , na.rm= TRUE), sd(suma$Suma.Internet, na.rm= TRUE)/mean(suma$Suma.Internet, na.rm= TRUE), skewness(suma$Suma.In ternet, na.rm= TRUE))

```

```

mediaSMSin <- c(mean(suma$Media.SMS.in, na.rm= TRUE), quantile(suma$Media.SMS.in, probs = c(0.25, 0.5, 0.75), na .rm= TRUE), sd(suma$Media.SMS.in, na.rm= TRUE)/mean(suma$Media.SMS.in, na.rm= TRUE), skewness(suma$Media.SMS.in, na.rm= TRUE))
mediaSMSout <- c(mean(suma$Media.SMS.out, na.rm= TRUE), quantile(suma$Media.SMS.out, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(suma$Media.SMS.out, na.rm= TRUE)/mean(suma$Media.SMS.out, na.rm= TRUE), skewness(suma$Media.SMS .out, na.rm= TRUE))
mediaCallin <- c(mean(suma$Media.Call.in, na.rm= TRUE), quantile(suma$Media.Call.in, probs = c(0.25, 0.5, 0.75), na .rm= TRUE), sd(suma$Media.Call.in, na.rm= TRUE)/mean(suma$Media.Call.in, na.rm= TRUE), skewness(suma$Media.Cal l.in, na.rm= TRUE))
mediaCallout <- c(mean(suma$Media.Call.out, na.rm= TRUE), quantile(suma$Media.Call.out, probs = c(0.25, 0.5, 0.75) , na.rm= TRUE), sd(suma$Media.Call.out, na.rm= TRUE)/mean(suma$Media.Call.out, na.rm= TRUE), skewness(suma$Med ia.Call.out, na.rm= TRUE))
mediaInternet <- c(mean(suma$Media.Internet, na.rm= TRUE), quantile(suma$Media.Internet, probs = c(0.25, 0.5, 0.75) , na.rm= TRUE), sd(suma$Media.Internet, na.rm= TRUE)/mean(suma$Media.Internet, na.rm= TRUE), skewness(suma$Me dia.Internet, na.rm= TRUE))

```

	Media	X25.	X50.	X75.	Coef..var
## sumaSMSin	460.638725	85.5585016	225.7939315	506.425245	1.555811
## sumaSMSout	282.243008	51.7381656	130.7169181	307.413729	1.622400
## sumaCallin	262.279988	49.3044257	129.4945718	300.559202	1.467289
## sumaCallout	319.676231	60.0130739	158.1971844	360.537532	1.455416
## sumaInternet	8247.924659	1666.1924573	4224.1119226	9102.350796	1.490477
## mediaSMSin	1.388215	0.3285429	0.8306251	1.741639	1.160543
## mediaSMSout	1.532654	0.3500380	0.8404048	1.883241	1.354348
## mediaCallin	1.504089	0.3591457	0.9041036	1.914439	1.142012
## mediaCallout	1.198387	0.3389586	0.7865976	1.517039	1.097699
## mediaInternet	34.193875	7.2871431	18.9801420	41.458299	1.240050
## Coef..asim					
## sumaSMSin	5.300328				
## sumaSMSout	5.486482				
## sumaCallin	4.278257				
## sumaCallout	4.174853				
## sumaInternet	4.112729				
## mediaSMSin	2.465640				
## mediaSMSout	5.575920				
## mediaCallin	2.517272				
## mediaCallout	2.804049				
## mediaInternet	2.601676				

Transformación logarítmica

La razón de este paso consiste en la asimetría de los datos que hemos encontrado al representarlos en la forma de histogramas. Para poder hacerlo teníamos que quitar la información de las celdas 5239 y 5339, que por alguna razón tenían el valor del tráfico total igual a 0 (que al pasar a la escala logarítmica nos da ???Infinito).

```

sumal <- suma[-5339,]
sumal <- sumal[-5239,]
logsum <- log(sumal[, 2:11])
colnames(logsum) <- c("logSuma.SMS.in", "logSuma.SMS.out", "logSuma.Call.in", "logSuma.Call.out", "logSuma.Internet", "logMedia.SMS.in", "logMedia.SMS.out", "logMedia.Call.in", "logMedia.Call.out", "logMedia.Internet")

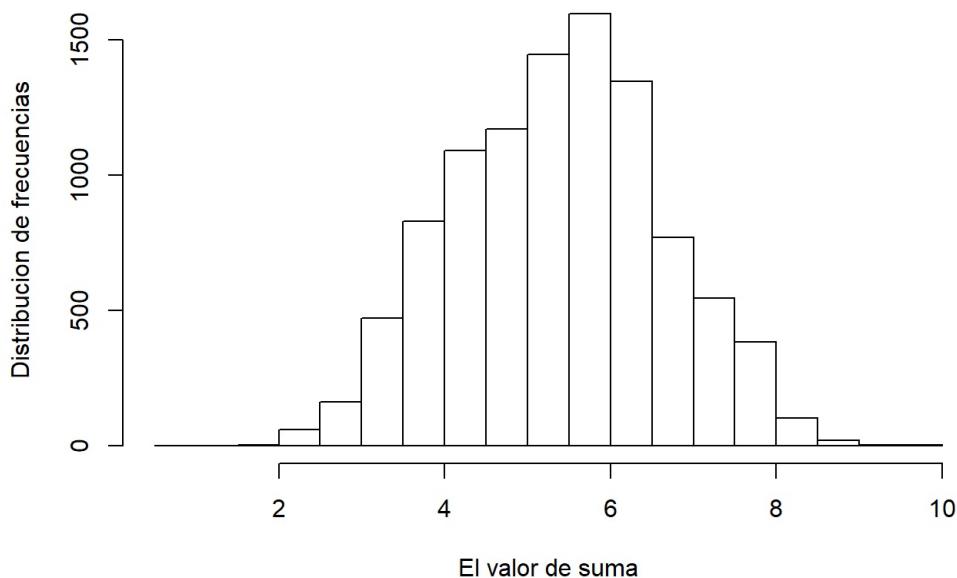
```

Tráfico total

Siendo simétricas las distribuciones de mensajes reflejan baja asimetría a la derecha mientras que todos los demás poseen de pequeña asimetría a la izquierda. La máxima cantidad de interacciones por celdas pertenece al tráfico de Internet y el tráfico de SMSs entrantes.

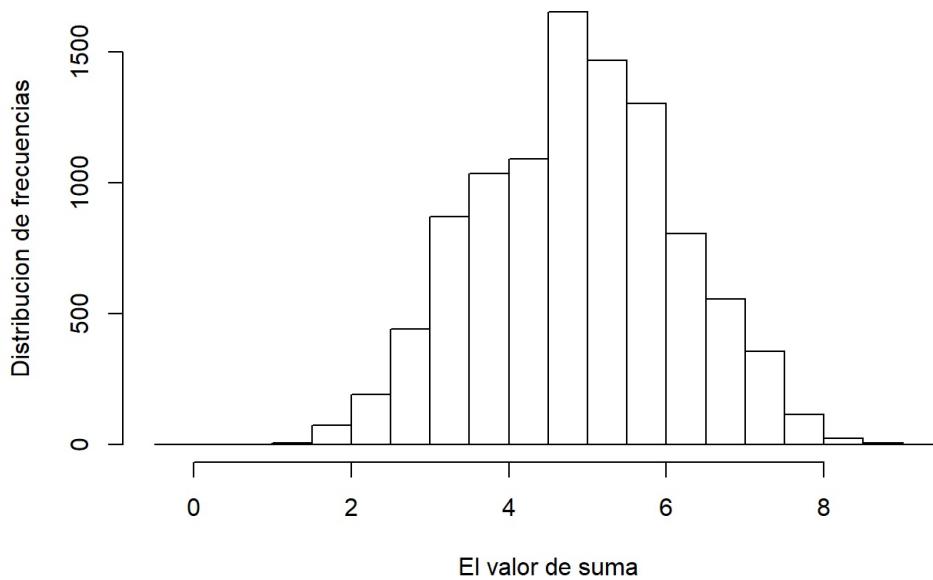
```
hist(logsum$logSuma.SMS.in, main = 'DF de total (e. l.): SMSs entrantes', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

DF de total (e. l.): SMSs entrantes



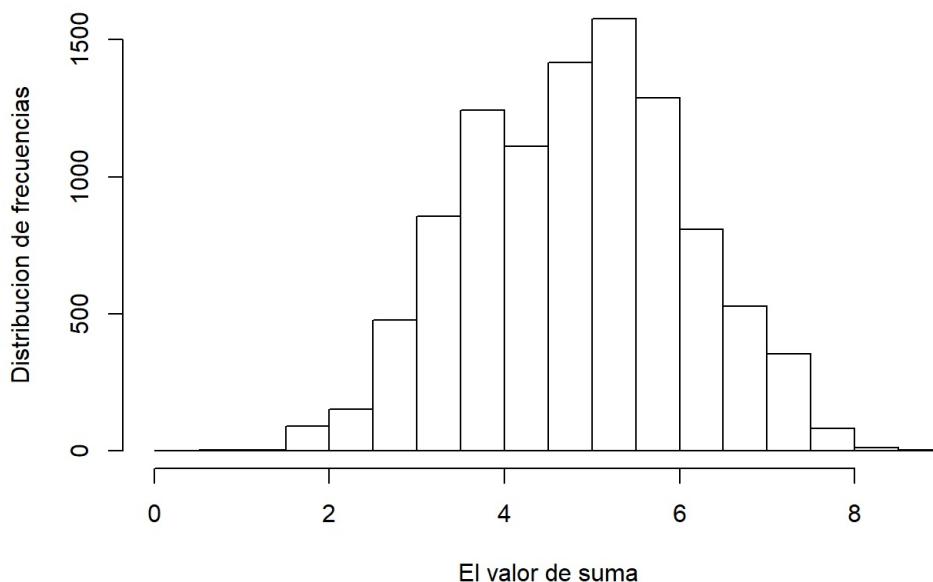
```
hist(logsum$logSuma.SMS.out, main = 'DF de total (e. l.): SMSs salientes', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

DF de total (e. l.): SMSs salientes



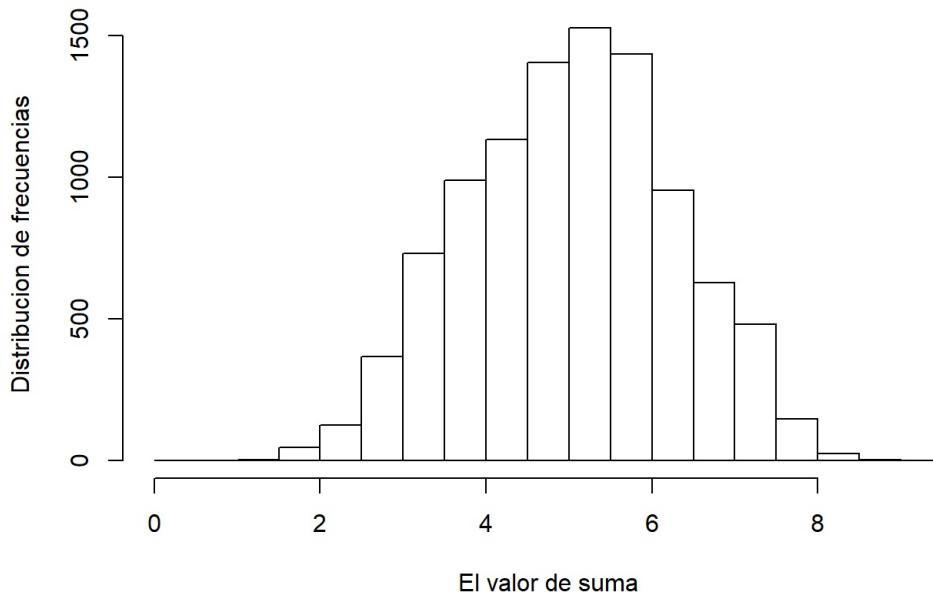
```
hist(logsum$logSuma.Call.in, main = 'DF de total (e. l.): llamadas entrantes', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

DF de total (e. l.): llamadas entrantes



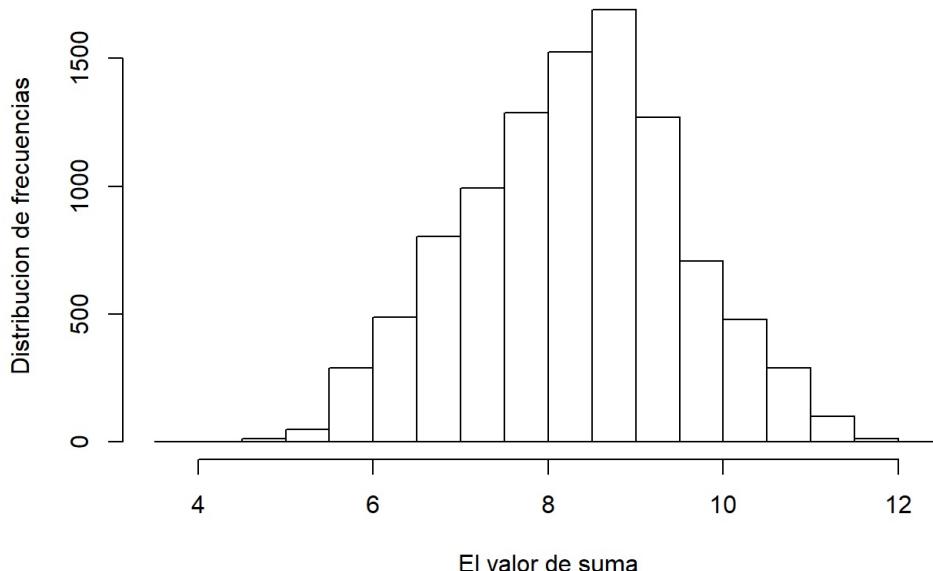
```
hist(logsum$logSuma.Call.out, main = 'DF de total (e. l.): llamadas salientes', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

DF de total (e. l.): llamadas salientes



```
hist(logsum$logSuma.Internet, main = 'DF de total (e. l.): Internet', xlab = 'El valor de suma', ylab = 'Distribucion de frecuencias')
```

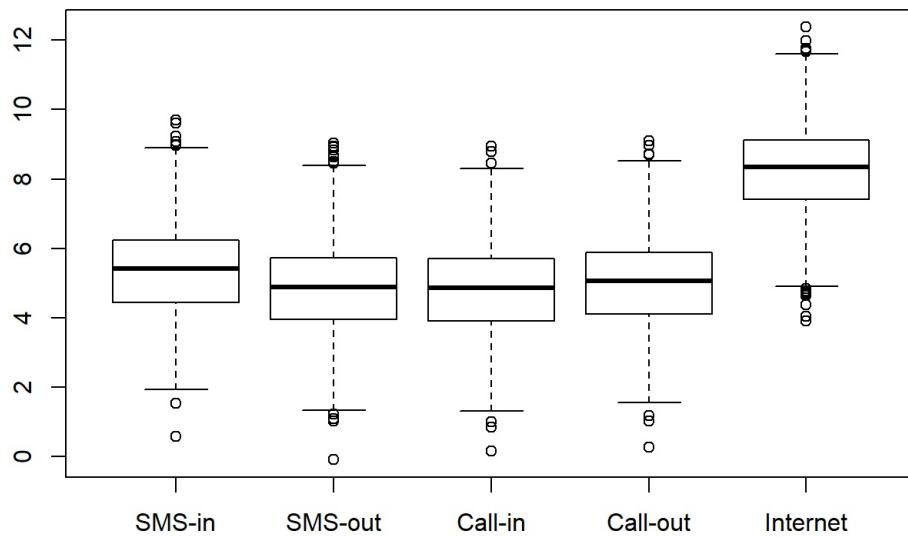
DF de total (e. l.): Internet



Los valores atípicos (disminuyendo en su cantidad) aparecen tanto en valores altos como en bajos. Eso se explica por la mayor simetría y menor variación de los datos en la escala logarítmica.

```
logSsmsin <- logsum$logSuma.SMS.in
logSsmsout <- logsum$logSuma.SMS.out
logScallin <- logsum$logSuma.Call.in
logScallout <- logsum$logSuma.Call.out
logSinter <- logsum$logSuma.Internet
boxplot(logSsmsin, logSsmsout, logScallin, logScallout, logSinter, main= "Total (e. l.)", names = c("SMS-in", "SMS-out", "Call-in", "Call-out", "Internet"))
```

Total (e. l.)

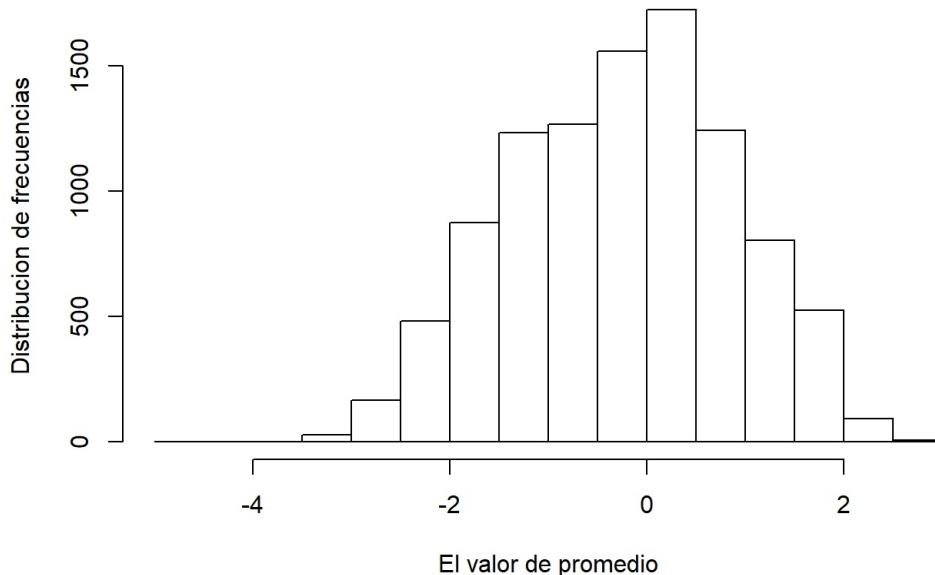


Tráfico promedio

Los valores obtenidos del promedio son tanto positivos como negativos, dado que alguna parte de los datos del promedio se encuentra en un intervalo [0,1]. En este caso todas las figuras demuestran la simetría con algo de inclinación a la izquierda.

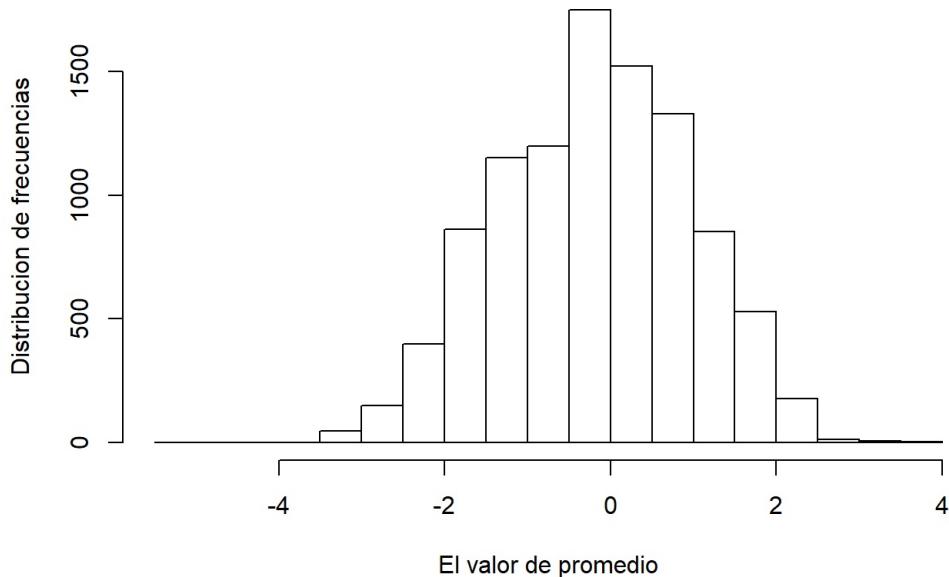
```
hist(logsum$logMedia.SMS.in, main = 'DF de promedio (e. l.): SMSs entrantes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio (e. l.): SMSs entrantes



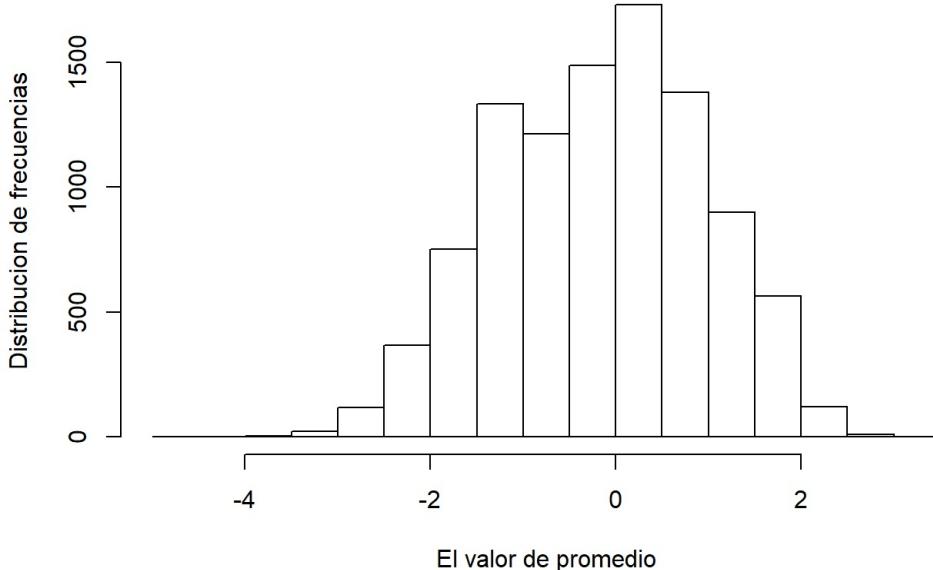
```
hist(logsum$logMedia.SMS.out, main = 'DF de promedio (e. l.): SMSs salientes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio (e. l.): SMSs salientes



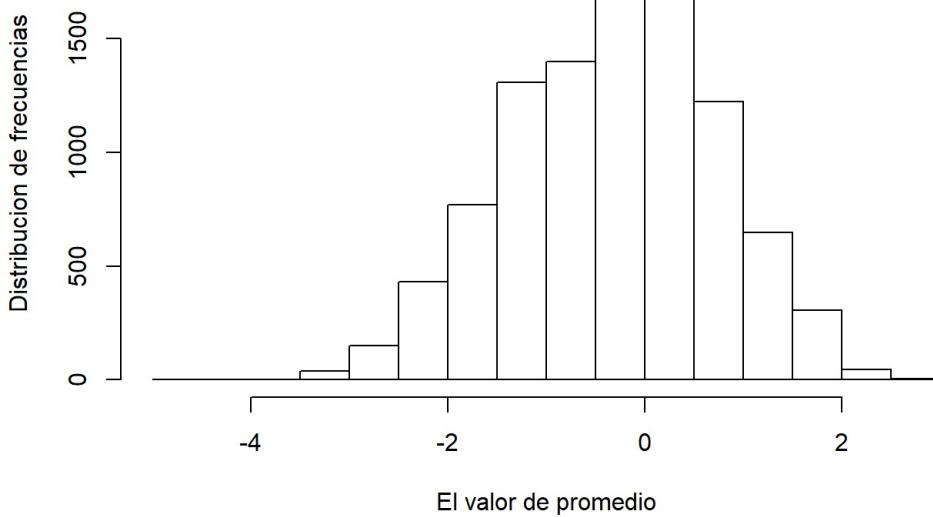
```
hist(logsum$logMedia.Call.in, main = 'DF de promedio (e. l.): llamadas entrantes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio (e. l.): llamadas entrantes



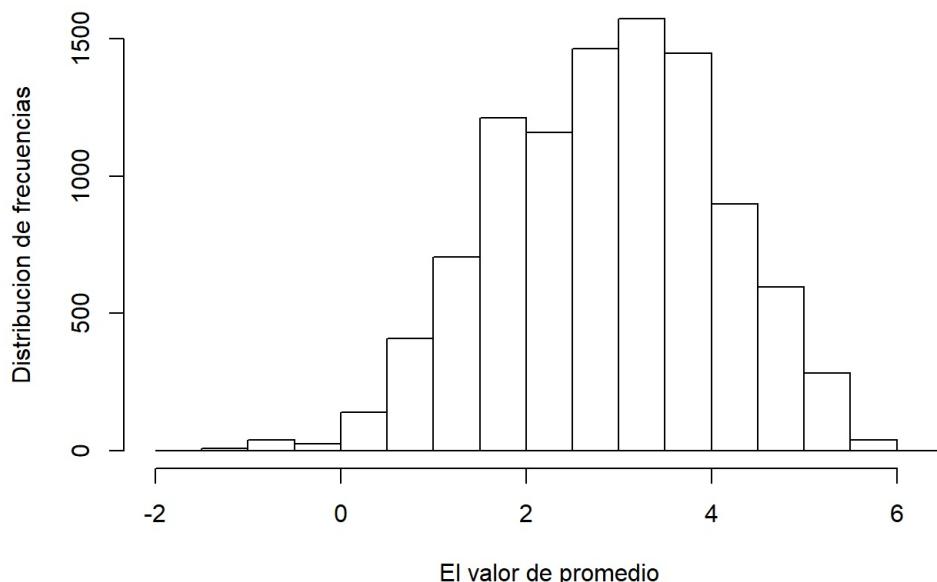
```
hist(logsum$logMedia.Call.out, main = 'DF de promedio (e. l.): llamadas salientes', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio (e. l.): llamadas salientes



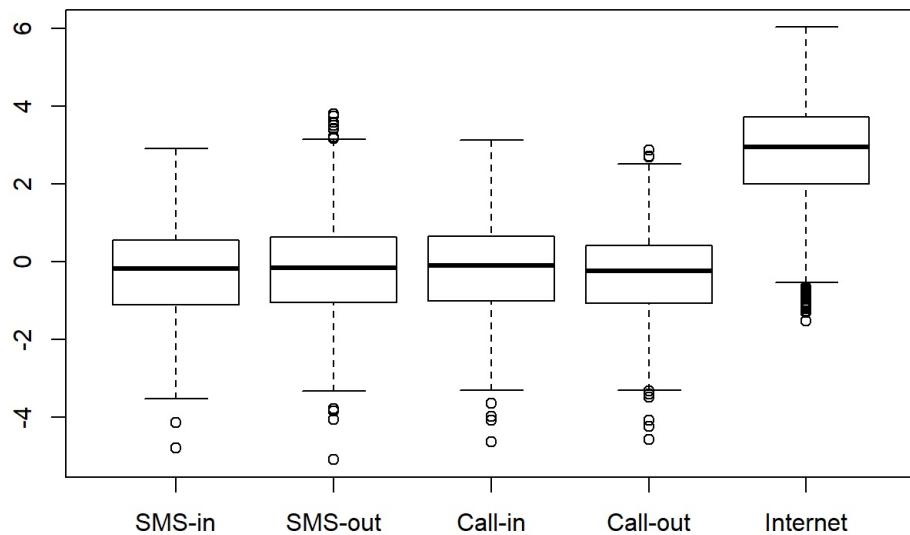
```
hist(logsum$logMedia.Internet, main = 'DF de promedio (e. l.): Internet', xlab = 'El valor de promedio', ylab = 'Distribucion de frecuencias')
```

DF de promedio (e. l.): Internet



```
logMsmsin <- logsum$logMedia.SMS.in  
logMsmsout <- logsum$logMedia.SMS.out  
logMcallin <- logsum$logMedia.Call.in  
logMcallout <- logsum$logMedia.Call.out  
logMinter <- logsum$logMedia.Internet  
boxplot(logMsmsin, logMsmsout, logMcallin, logMcallout, logMinter, main= "Promedio (e. l.)", names = c("SMS-in",  
"SMS-out", "Call-in", "Call-out", "Internet"))
```

Promedio (e. l.)



Medidas de posición, dispersión y forma (e. l.)

```

LOGsumaSMSin <- c(mean(logsum$logSuma.SMS.in, na.rm= TRUE), quantile(logsum$logSuma.SMS.in, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logSuma.SMS.in, na.rm= TRUE)/mean(logsum$logSuma.SMS.in, na.rm= TRUE), skewness(logsum$logSuma.SMS.in, na.rm= TRUE))
LOGsumaSMSout <- c(mean(logsum$logSuma.SMS.out, na.rm= TRUE), quantile(logsum$logSuma.SMS.out, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logSuma.SMS.out, na.rm= TRUE)/mean(logsum$logSuma.SMS.out, na.rm= TRUE), skewness(logsum$logSuma.SMS.out, na.rm= TRUE))
LOGsumaCallin <- c(mean(logsum$logSuma.Call.in, na.rm= TRUE), quantile(logsum$logSuma.Call.in, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logSuma.Call.in, na.rm= TRUE)/mean(logsum$logSuma.Call.in, na.rm= TRUE), skewness(logsum$logSuma.Call.in, na.rm= TRUE))
LOGsumaCallout <- c(mean(logsum$logSuma.Call.out, na.rm= TRUE), quantile(logsum$logSuma.Call.out, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logSuma.Call.out, na.rm= TRUE)/mean(logsum$logSuma.Call.out, na.rm= TRUE), skewness(logsum$logSuma.Call.out, na.rm= TRUE))
LOGsumaInternet <- c(mean(logsum$logSuma.Internet, na.rm= TRUE), quantile(logsum$logSuma.Internet, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logSuma.Internet, na.rm= TRUE)/mean(logsum$logSuma.Internet, na.rm= TRUE), skewness(logsum$logSuma.Internet, na.rm= TRUE))

```

```

LOGmediaSMSin <- c(mean(logsum$logMedia.SMS.in, na.rm= TRUE), quantile(logsum$logMedia.SMS.in, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logMedia.SMS.in, na.rm= TRUE)/mean(logsum$logMedia.SMS.in, na.rm= TRUE), skewness(logsum$logMedia.SMS.in, na.rm= TRUE))
mediaSMSout <- c(mean(logsum$logMedia.SMS.out, na.rm= TRUE), quantile(logsum$logMedia.SMS.out, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logMedia.SMS.out, na.rm= TRUE)/mean(logsum$logMedia.SMS.out, na.rm= TRUE), skewness(logsum$logMedia.SMS.out, na.rm= TRUE))
LOGmediaCallin <- c(mean(logsum$logMedia.Call.in, na.rm= TRUE), quantile(logsum$logMedia.Call.in, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logMedia.Call.in, na.rm= TRUE)/mean(logsum$logMedia.Call.in, na.rm= TRUE), skewness(logsum$logMedia.Call.in, na.rm= TRUE))
LOGmediaCallout <- c(mean(logsum$logMedia.Call.out, na.rm= TRUE), quantile(logsum$logMedia.Call.out, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logMedia.Call.out, na.rm= TRUE)/mean(logsum$logMedia.Call.out, na.rm= TRUE), skewness(logsum$logMedia.Call.out, na.rm= TRUE))
LOGmediaInternet <- c(mean(logsum$logMedia.Internet, na.rm= TRUE), quantile(logsum$logMedia.Internet, probs = c(0.25, 0.5, 0.75), na.rm= TRUE), sd(logsum$logMedia.Internet, na.rm= TRUE)/mean(logsum$logMedia.Internet, na.rm= TRUE), skewness(logsum$logMedia.Internet, na.rm= TRUE))

```

	Media	X25.	X50.	X75.	Coef..var
## sumaSMSin	460.6387250	85.5585016	225.7939315	506.425245	1.555811
## sumaSMSout	282.2430083	51.7381656	130.7169181	307.413729	1.622400
## sumaCallin	262.2799882	49.3044257	129.4945718	300.559202	1.467289
## sumaCallout	319.6762311	60.0130739	158.1971844	360.537532	1.455416
## sumaInternet	8247.9246590	1666.1924573	4224.1119226	9102.350796	1.490477
## mediaSMSin	1.3882146	0.3285429	0.8306251	1.741639	1.160543
## mediaSMSout	-0.1940575	NA	NA	NA	-5.945380
## mediaCallin	1.5040886	0.3591457	0.9041036	1.914439	1.142012
## mediaCallout	1.1983869	0.3389586	0.7865976	1.517039	1.097699
## mediaInternet	34.1938749	7.2871431	18.9801420	41.458299	1.240050
##					
##	Coef..asim				
## sumaSMSin	5.30032839				
## sumaSMSout	5.48648213				
## sumaCallin	4.27825675				
## sumaCallout	4.17485333				
## sumaInternet	4.11272870				
## mediaSMSin	2.46563997				
## mediaSMSout	-0.08563387				
## mediaCallin	2.51727179				
## mediaCallout	2.80404928				
## mediaInternet	2.60167577				

Dinámica de generación del tráfico

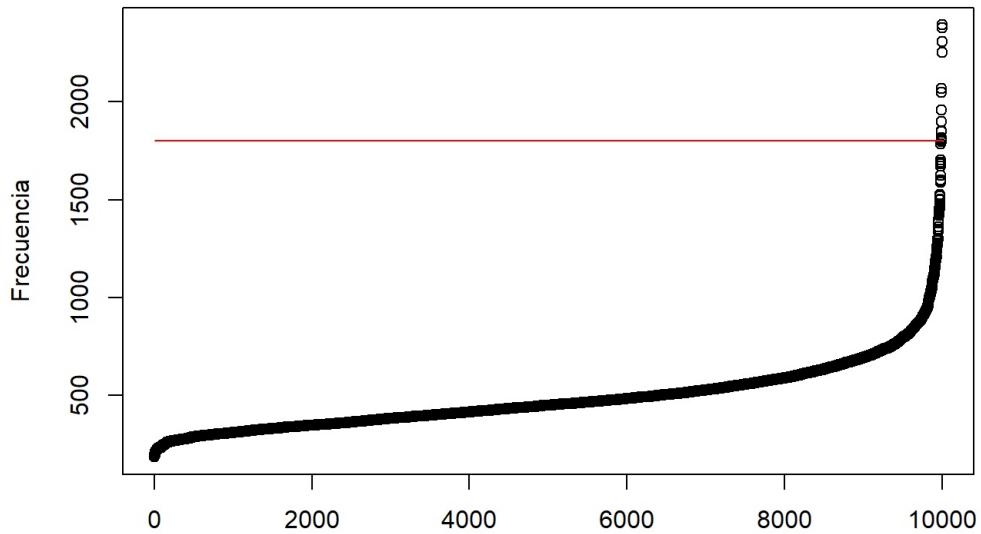
Las celdas con la máxima frecuencia de las interacciones se encuentran en un intervalo de la celda 5059 a la 6165 que est%aacute; por la mitad de la numeración asignada para Milán. Eso significa que las celdas representadas deben coincidir con el centro de la ciudad y as?? se puede explicar la actividad tan alta que sufren estas celdas. El mejor modo de comprobar nuestra hipótesis es juntar el mapa del Milán con el mapa de celdas (por las coordenadas dadas) y ver con qué barrios o lugares de interés coinciden las celdas representadas en el diagrama.

```

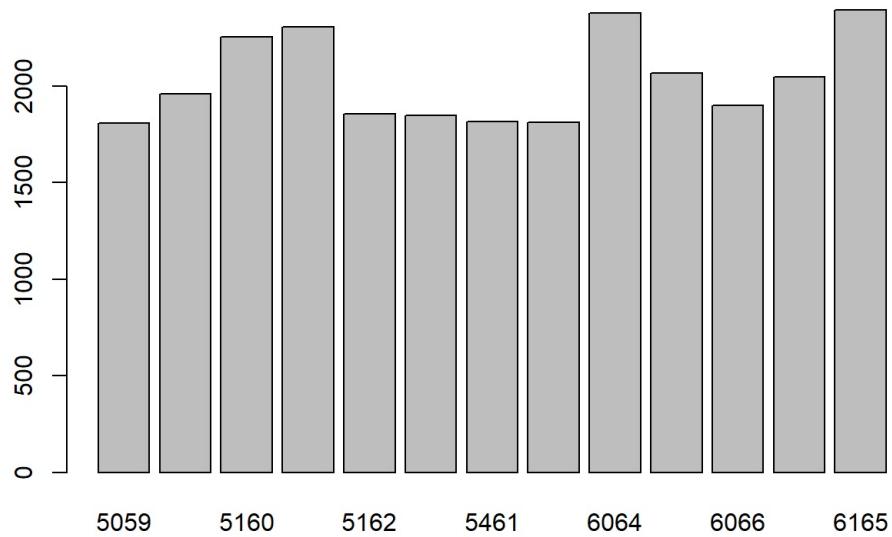
celda <- table(datos$`Square id`)
celdasort <- sort(as.vector(celda))
plot(celdasort, main = "DF de las celdas", xlab = "", ylab = "Frecuencia")
lines(c(0,10000), c(1800,1800), col = 2)

```

DF de las celdas

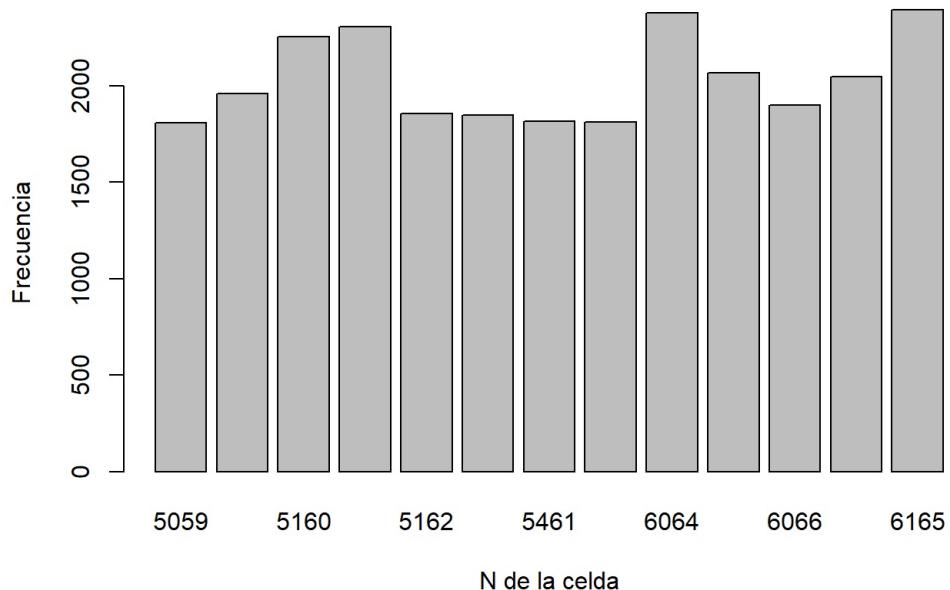


```
barplot(celda[celda > 1800])
```



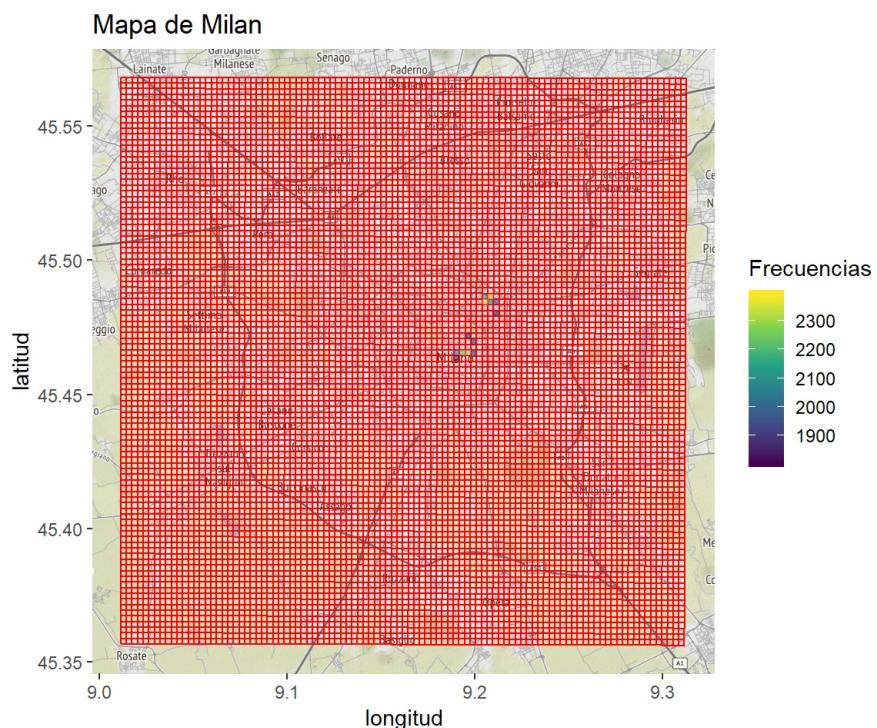
```
celdamax <- celda[celda > 1800]
barplot(celdamax, main = "Celdas con la frecuencia maxima", xlab = "N de la celda", ylab = "Frecuencia")
```

Celdas con la frecuencia maxima

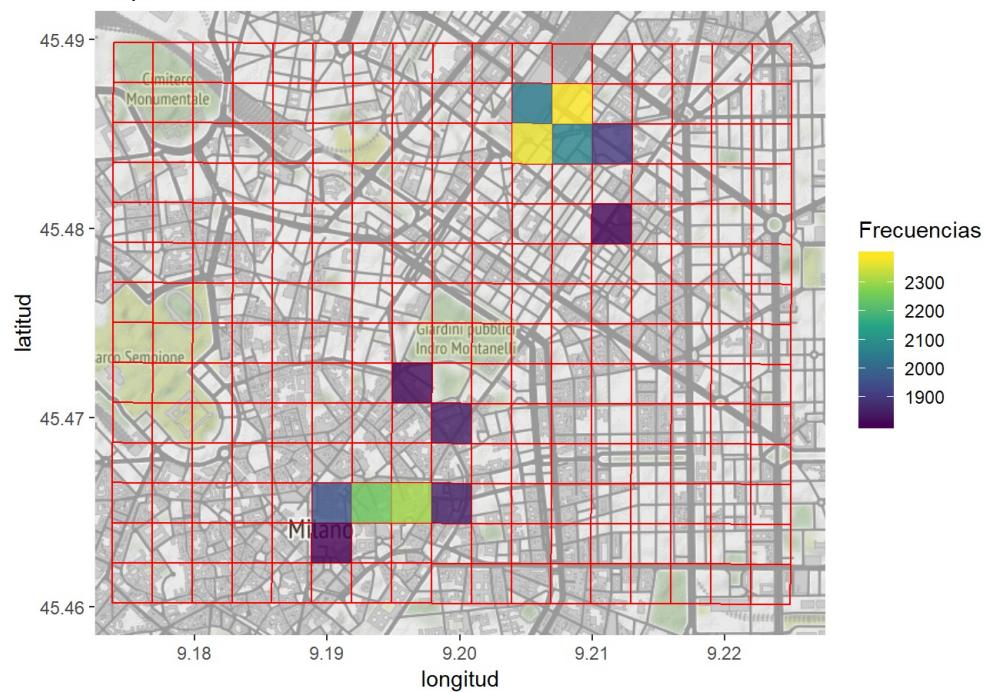


```
print(celdamax)
```

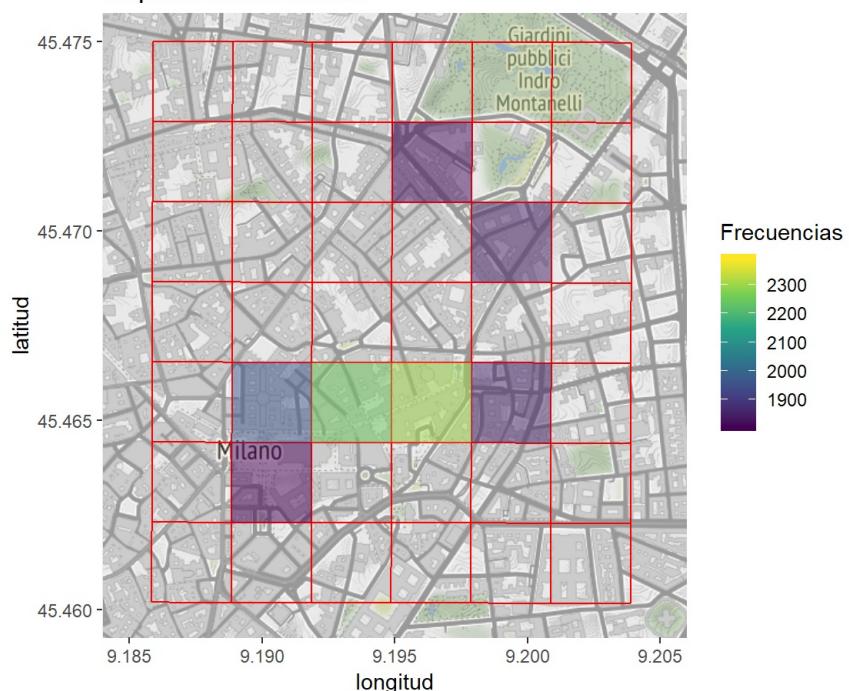
```
##  
## 5059 5159 5160 5161 5162 5362 5461 5866 6064 6065 6066 6164 6165  
## 1806 1960 2252 2307 1854 1847 1817 1810 2377 2067 1900 2048 2393
```



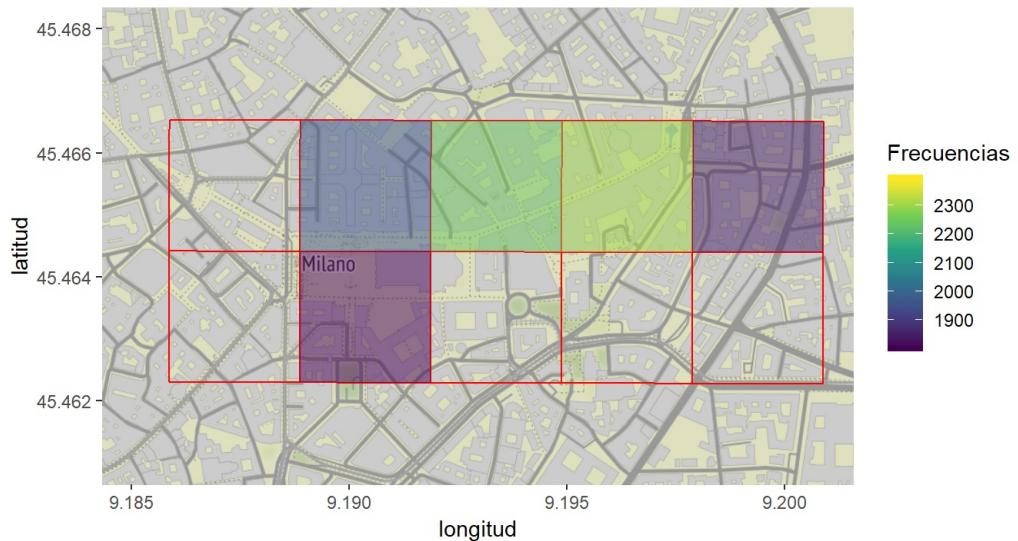
Mapa de Milan: Centro



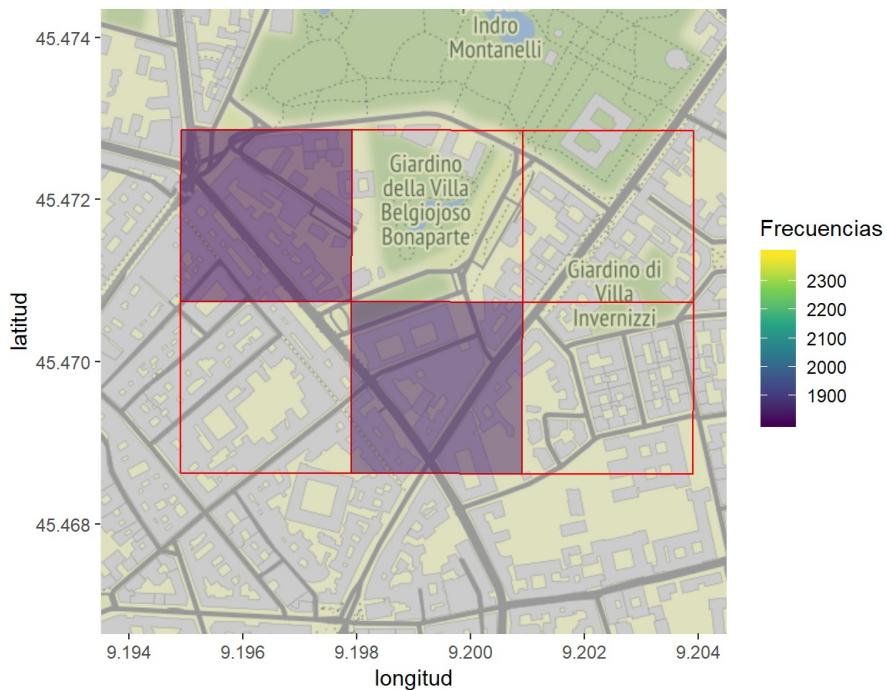
Mapa de Milan: Centro



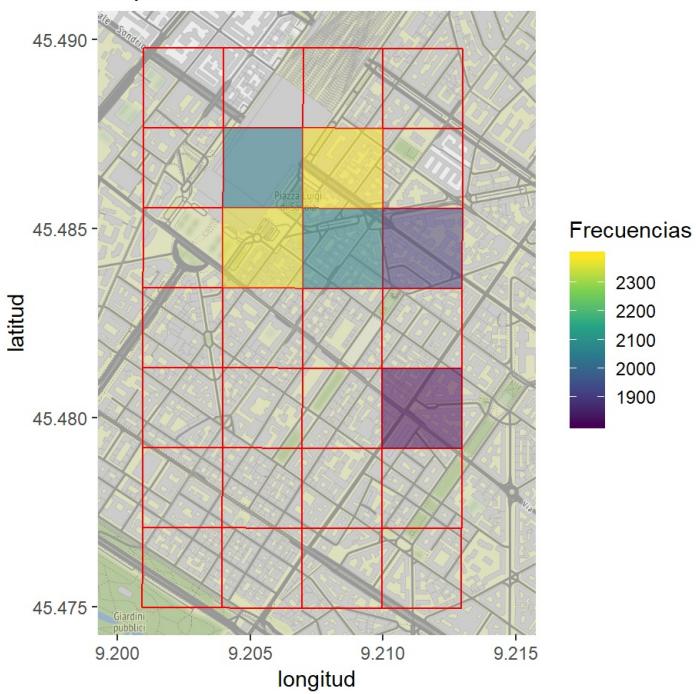
Mapa de Milan: Duomo di Milano



Mapa de Milan: metro



Mapa de Milán: Estación central



Las celdas con la cantidad máxima de las interacciones coinciden con la Estación Central de Milán y la famosa Catedral de Milán. Las celdas azules y violetas coinciden con siguientes estaciones del metro: Loreto, Lima, Centrale, Duomo. En el último mapa no hemos podido sacar algún edificio o elemento de la red de transporte preciso porque las celdas indicadas contienen muchas lugares de interés.