Programación en R

Gabriel Ledda[*,1], Agustina Lopez[1], Lucas Pizarro[**,1], Florencia Puebla[**,1], Pilar Cortiñas[1]

[a]*Department, Street, City, State, Zip*
[b]*Department, Street, City, State, Zip*

Abstract

This is the abstract.

It consists of two paragraphs.

Elementos básicos de programación

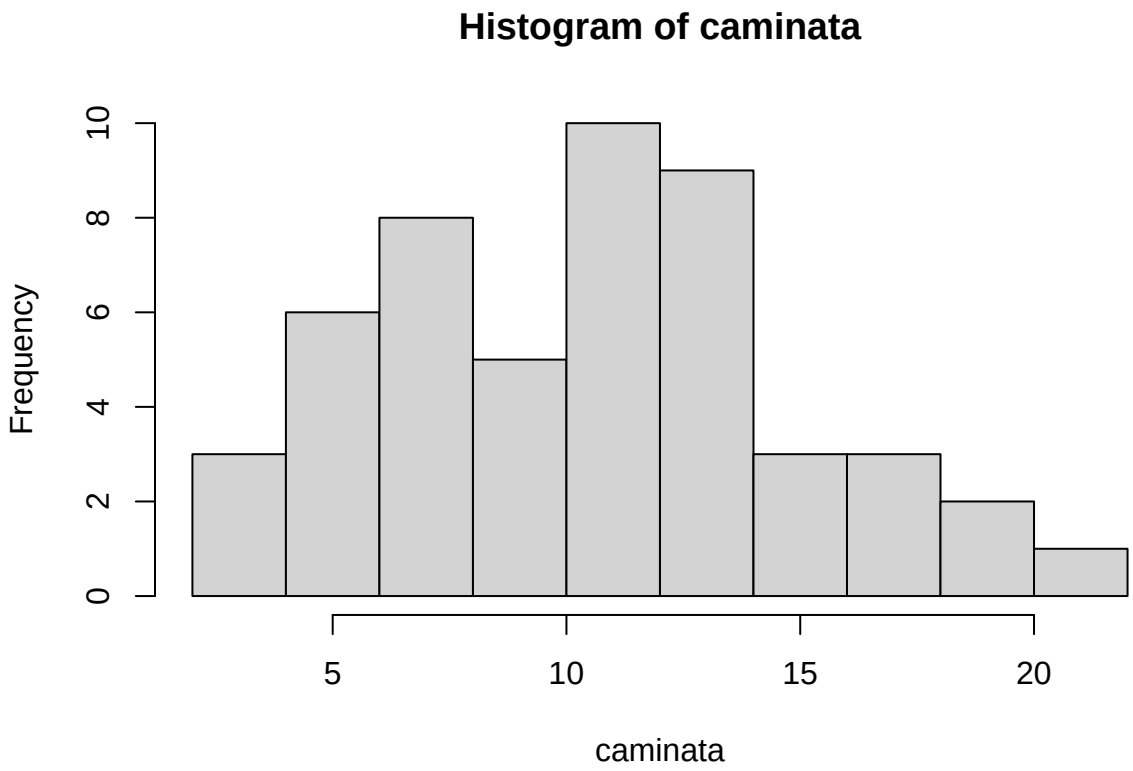
En este breve tutorial examinaremos algunos elementos del lenguaje de programacion R y como valernos de ello para resolver problemas de la vida cotidiana. Apelaremos a ejemplos bien conocidos, pero además mostraremos las soluciones que desarrollaremos contra las misma que ya están implementadas en R. Comparando el **costo computacional**, medido como tiempo de ejecución. Esto nos permitirá entender la calidad del algoritmo que implementemos. Como excusa para introducirnos propondremos realizar tres experimentos y medir el tiempo ejecución. Veremos:

- Generar un vector secuencia
- Implementación de una serie Fibonacci
- Ordenación de un vector por método burbuja
- Progresión geométrica del COVid-19

Como ejecutar comandos de R

```
caminata <- rnorm(50, 10, 5)
hist(caminata)
```

*Corresponding Author
**Equal contribution
Email addresses: gabinledda@gmail.com (Gabriel Ledda), agustinaayelenlopez@gmail.com (Agustina Lopez), lucas.pizarro137@gmail.com (Lucas Pizarro), pueblaflor.fp@gmail.com (Florencia Puebla), pilarcrt9@gmail.com (Pilar Cortiñas)



Algunas ideas de como medir el tiempo de ejecución

Es muy frecuente la pregunta sobre cuál algoritmo es el mejor para realizar una tarea específica o resolver un problema. Una de las técnicas que suele utilizarse para esto es el benchmarking. No necesariamente un algoritmo es siempre mejor que otro. Por el contrario existen métricas para hacer el benchmarking que comparan el uso del procesador, la memoria requerida, el tiempo que tarda en resolver el problema, la exactitud de la solución, etc. El benchmarking no me dice cual es el mejor, pero me da información sobre cuál es el más conveniente según las características del problema y el recurso físico (máquina) que tengo.

Muchos de ustedes están familiarizados con Octave o Matlab. Algunos recordarán que para invertir matrices y saber que método era más eficiente su utilizaban los comandos tic y tac. Por ejemplo se generaba una matriz A, se ejecutaba el comando tic que disparaba una especie de cronómetro interno, luego se invertía siguiendo una algoritmo de determinante y finalmente se ejecutaba el tomando toc que detenía el reloj y entregaba el tiempo de ejecución. Luego se repetía el mismo procedimiento, pero en lugar de hacerlo con determinante se

usaba un algoritmo de matriz LU.

Una búsqueda rápida en línea nos revela al menos tres paquetes R para comparar performance del código R (rbenchmark, microbenchmark y tictoc). Estos además de medir el tiempo nos indican porcentaje de memoria y microprocesador utilizados.

Además, la base R proporciona al menos dos métodos para medir el tiempo de ejecución del código R (Sys.time y system.time), que es una aproximación bastante útil para un curso como el que desarrollamos. A continuación, paso brevemente por la sintaxis del uso de cada una de las cinco opciones, y presento mis conclusiones al final.

A continuación, paso brevemente por la sintaxis del uso de cada una de las cinco opciones, y presento mis conclusiones al final.

Usando Sys.time

El tiempo de ejecución de un fragmento de código se puede medir tomando la diferencia entre el tiempo al inicio y al final del fragmento de código leyendo los registros del RTC (Real Time Clock. Simple pero flexible: como un relojito de arena :.

```
sleep_for_a_minute <- function() { Sys.sleep(14) }
start_time <- Sys.time()
sleep_for_a_minute()
end_time <- Sys.time()
end_time - start_time
```

```
## Time difference of 14.19799 secs
```

Hemos generado una función que antes no existía y la hemos usado. Deficiencias: Si usas el comando dentro de un documento en R-Studio te demorarás mucho tiempo cuando compiles un PDF o una presentación.

Biblioteca tictoc

Esto de usar una biblioteca es llamar u cargar una procedimientos que generará comando nuevos en R. Como ya fue comentado, cargar una biblioteca implica ejecutar el comando install.packages() o usar en r-studio el menú de Herramientas y Luego Instalar paquetes. Las funciones tic y toc son de la misma biblioteca de Octave/Matlab y se usan de la misma manera para la evaluación comparativa que el tiempo de sistema recién demostrado. Sin embargo, tictoc agrega mucha más comodidad al usuario y armonía al conjunto. La versión de desarrollo más reciente de tictoc se puede instalar desde github: install packages (tictoc)

```
library(tictoc)
tic("sleeping")
A<-20
print("dormire una siestita...")
```

```
## [1] "dormire una siestita..."
```

```
Sys.sleep(2)
print("...suena el despertador")
```

```
## [1] "...suena el despertador"
```

```
toc()
```

```
## sleeping: 2.05 sec elapsed
```

Uno puede cronometrar solamente un fragmento de código a la vez.

#Biblioteca rbenchmark

La documentación de la función benchmark del paquete rbenchmark R lo describe como "un simple contenedor alrededor de system.time." Sin embargo, agrega mucha conveniencia en comparación con las llamadas simples a system. time. Por ejemplo, requiere solo una llamada de referencia para cronometrar múltiples repeticiones de múltiples expresiones. Además, los resultados devueltos se organizan convenientemente en un marco de datos. Recuerda antes de ejecutar

##library (cualquiercosa)##

debes haber cargado en tu máquina la biblioteca que quieres invocar usando install.packages (cualquiercosa).

«bench_mark,echo=TRUE»=

```
library(rbenchmark)
# lm crea una regresión lineal
benchmark("lm" = {
X <- matrix(rnorm(1000), 100, 10)
y <- X %*% sample(1:10, 10) + rnorm(100)
b <- lm(y ~ X + 0)$coef
},
"pseudoinverse" = {
X <- matrix(rnorm(1000), 100, 10)
y <- X %*% sample(1:10, 10) + rnorm(100)
b <- solve(t(X) %*% X) %*% t(X) %*% y
},
"linear system" = {
X <- matrix(rnorm(1000), 100, 10)
y <- X %*% sample(1:10, 10) + rnorm(100)
b <- solve(t(X) %*% X, t(X) %*% y)
```

```
},
replications = 1000,
columns = c("test", "replications", "elapsed",
"relative", "user.self", "sys.self"))
```

```
##            test replications elapsed relative user.self sys.self
## 3 linear system         1000    0.14    1.000      0.14     0.00
## 1            lm         1000    0.87    6.214      0.86     0.01
## 2 pseudoinverse         1000    0.17    1.214      0.17     0.00
```

En el informe de salida nos dice que cantidad de tiempo consume cada parte del código.

Biblioteca Microbenchmark

La versión de desarrollo más reciente de microbenchmark se puede instalar desde github:

Al igual que el punto de referencia del paquete rbenchmark, la función microbenchmark se puede usar para comparar tiempos de ejecuci?n de múltiples fragmentos de cdigo R. Pero ofrece una gran comodidad y funcionalidad adicional. Es más "beta" (inestable), pero como todo lo que hoy es nuevo poco a poco se hará más estable y no complicará tanto las cosas para el usuario final. Una cosa interesante es que se puede ver la salida gráfica del uso de recursos. Ver líneas finales del código. Me parece que una característica particularmente agradable de microbenchmark es la capacidad de verificar automáticamente los resultados de las expresiones de referencia con una función especificada por el usuario. Esto se demuestra a continuación, donde nuevamente comparamos tres métodos que computan el vector de coeficientes de un modelo lineal.

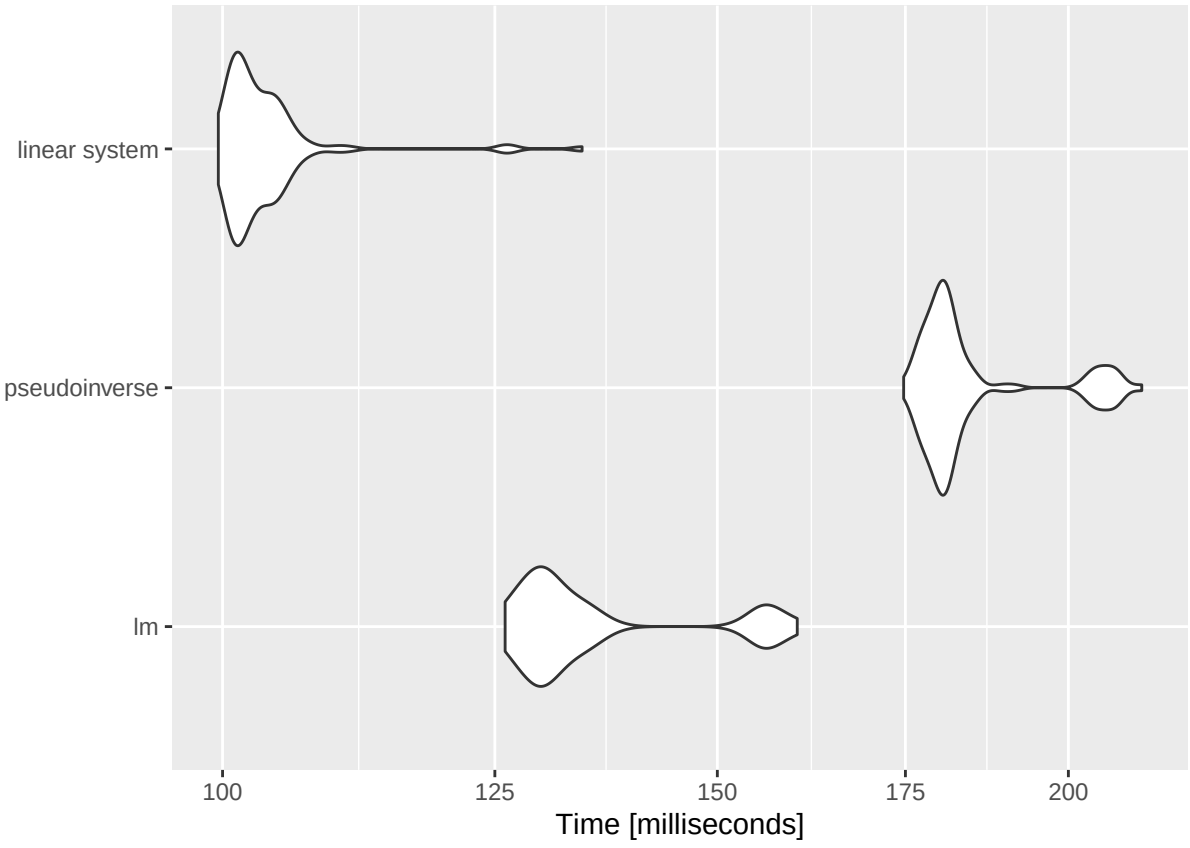
```
library(microbenchmark)
set.seed(2017)
n <- 10000
p <- 100
X <- matrix(rnorm(n*p), n, p)
y <- X %*% rnorm(p) + rnorm(100)
check_for_equal_coefs <- function(values) {
tol <- 1e-12
max_error <- max(c(abs(values[[1]] - values[[2]]),
abs(values[[2]] - values[[3]]),
abs(values[[1]] - values[[3]])))
max_error < tol
}
mbm <- microbenchmark("lm" = { b <- lm(y ~ X + 0)$coef },
"pseudoinverse" = {
b <- solve(t(X) %*% X) %*% t(X) %*% y
```

```
},
"linear system" = {
b <- solve(t(X) %*% X, t(X) %*% y)
},
check = check_for_equal_coefs)
mbm
```

```
## Unit: milliseconds
##           expr      min       lq     mean   median       uq      max neval
##             lm 126.0558 129.2762 135.7460 130.9651 135.5945 160.1525   100
##  pseudoinverse 174.7469 179.1562 184.7844 180.8130 183.9194 212.4011   100
##  linear system  99.6521 101.0528 103.4529 102.1950 104.2553 134.2659   100
```

```
library(ggplot2)
autoplot(mbm)
```

```
## Coordinate system already present. Adding new coordinate system, which will replace the e
```



Consigna del trabajo

El trabajo de hay que presentar implica revisar los algoritmos que se presentan a continuación. Deberá ejecutarlos primero en la línea de comando de la consola. Luego deberá elegir alguno de los métodos vistos para medir la performance y comparar los resultados con otros compañeros que hayan usado otros métodos para medir la performance. Luego todo deberá entregarse en un informe en formato pdf construido con RStudio, archivo RMarkdow.

Generar un vector secuencia

De echo R. tiene un comando para generar secuencias llamado "seq." Recomendamos ejecutar la ayuda del comando en RStudio. Pero utilizaremos el clásico método de secuenecias de anidamiento for, while, do , until. Generaremos una secuencia de números que de dos en dos entre 1 y 100.000.

Secuencias generada con for

```
for (i in 1:50000) { A[i] <- (i*2)}
head (A)
```

```
## [1]  2  4  6  8 10 12
```

```
tail (A)
```

```
## [1]  99990  99992  99994  99996  99998 100000
```

Secuencia generada con R

```
A <- seq(1,1000000, 2)
head (A)
```

```
## [1]  1  3  5  7  9 11
```

```
tail (A)
```

```
## [1] 999989 999991 999993 999995 999997 999999
```

CONSIGNA: Comparar la performance con systime
A continuación evaluamos la performance de estos algoritmos:
Performance secuencia generada con for:

```
start_time <- Sys.time()

for (i in 1:50000) { A[i] <- (i*2)}
head (A)
```

```
## [1]  2  4  6  8 10 12
```

```
tail (A)
```

```
## [1] 999989 999991 999993 999995 999997 999999
```

```
end_time <- Sys.time()
end_time - start_time
```

```
## Time difference of 0.01202488 secs
```

Performance secuancia generada con R

```
start_time <- Sys.time()

A <- seq(1,1000000, 2)
head (A)
```

```
## [1]  1  3  5  7  9 11
```

```
tail (A)
```

```
## [1] 999989 999991 999993 999995 999997 999999
```

```
end_time <- Sys.time()
end_time - start_time
```

```
## Time difference of 0.01196384 secs
```

Implementación de una serie Fibonachi o Fibonacci

En matemáticas, la sucesión o serie de Fibonacci es la siguiente sucesión infinita de números naturales: 0,1,1,2,3,5,8 ... 89,144,233 ... La sucesión comienza con los números 0 y 1,2 a partir de estos, «cada término es la suma de los dos anteriores», es la relación de recurrencia que la define. A los elementos de esta sucesión se les llama números de Fibonacci. Esta sucesión fue descrita en Europa por Leonardo de Pisa, matemático italiano del siglo XIII también conocido como Fibonacci. Tiene numerosas aplicaciones en ciencias de la computación, matemática y teoría de juegos. También aparece en configuraciones biológicas,

como por ejemplo en las ramas de los árboles, en la disposición de las hojas en el tallo, en las flores de alcachofas y girasoles, en las inflorescencias del brécol romanesco, en la configuración de las piñas de las coníferas, en la reproducción de los conejos y en como el ADN codifica el crecimiento de formas orgánicas complejas. De igual manera, se encuentra en la estructura espiral del caparazón de algunos moluscos, como el nautilus. Original de la Biblioteca Uninersidad de Florencia. Liber Abachi - Autor Fibonacci

==========

Fibonacci

```
a<-0
b <-1
print(a)
```

```
## [1] 0
```

```
print(b)
```

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

```
for(i in 0:5)
{
c <- a+b
# comentar esta línea para conocer el número más grande hallado
print(c)
a <- b
b <- c
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 5
## [1] 8
## [1] 13
```

```
#Descomentar esta línea para saber el número más grande hallado
# print(c)
```

CONSIGNA: ¿Cuantas iteraciones se necesitan para generar un número de la serie mayor que 1.000.000 ?

Debido al gran crecimiento que tiene la sucesión podemos probar el algoritmo algunas veces hasta alcanzar el valor deseado.

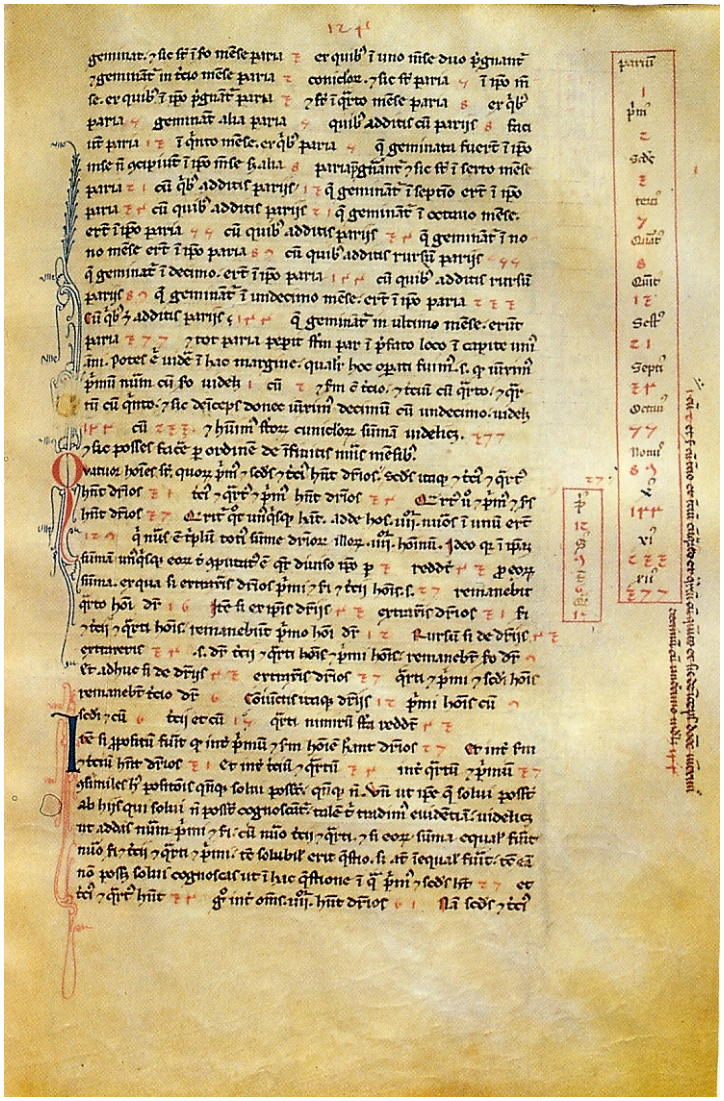


Figure 1: Fig 3

```r
a<-0
b <-1
print(a)
```

```
## [1] 0
```

```r
print(b)
```

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

```r
for(i in 0:28)
{
c <- a+b
# comentar esta línea para conocer el número más grande hallado
# print(c)
a <- b
b <- c
}
#Descomentar esta l?nea para saber el número más grande hallado
print(c)
```

```
## [1] 832040
```

```r
a<-0
b <-1
print(a)
```

```
## [1] 0
```

```r
print(b)
```

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

```r
for(i in 0:29)
{
c <- a+b
# comentar esta línea para conocer el número más grande hallado
# print(c)
a <- b
b <- c
}
#Descomentar esta l?nea para saber el número más grande hallado
print(c)
```

```
## [1] 1346269
```

Observamos que el número de iteraciones para generar un número mayor a 1.000.000 es 29.

Ordenación de un vector por método burbuja

La Ordenación de burbuja (**Bubble Sort en inglés**) es un sencillo algoritmo de ordenamiento. Funciona revisando cada elemento de la lista que va a ser ordenada con el siguiente, intercambiándolos de posición si están en el orden equivocado. Es necesario revisar varias veces toda la lista hasta que no se necesiten más intercambios, lo cual significa que la lista está ordenada. Este algoritmo obtiene su nombre de la forma con la que suben por la lista los elementos durante los intercambios, como si fueran pequeñas ***burbujas***. También es conocido como el método del intercambio directo. Dado que solo usa comparaciones para operar elementos, se lo considera un algoritmo de comparación, siendo uno de los más sencillos de implementada.

```
# Tomo una muestra de 10 números ente 1 y 100
x<-sample(1:100,10)
# Creo una función para ordenar
burbuja <- function(x){
n<-length(x)
for(j in 1:(n-1)){
for(i in 1:(n-j)){
if(x[i]>x[i+1]){
temp<-x[i]
x[i]<-x[i+1]
x[i+1]<-temp
}
}
}
return(x)
}
res<-burbuja(x)
#Muestra obtenida
x
```

```
##  [1] 43 54 59 99 79 72 76 24 45 62
```

```
#Muestra Ordenada
res
```

```
##  [1] 24 43 45 54 59 62 72 76 79 99
```

Con el comando sort de R

```
#Ordanación con el comando SORT de R-Cran
sort(x)
```

```
##  [1] 24 43 45 54 59 62 72 76 79 99
```

CONSIGNA: Compara la performance de ordenación del método burbuja vs el método sort de R Usar método microbenchmark para una muestra de tamaño 20.000

Primero compararemos ambos algoritmos usando el método tictoc

```
library(tictoc)
tic("Performance burbuja: ")

# Tomo una muestra de 10 números ente 1 y 100
x<-sample(1:100,10)
# Creo una función para ordenar
burbuja <- function(x){
n<-length(x)
for(j in 1:(n-1)){
for(i in 1:(n-j)){
if(x[i]>x[i+1]){
temp<-x[i]
x[i]<-x[i+1]
x[i+1]<-temp
}
}
}
return(x)
}
res<-burbuja(x)

toc()
```

```
## Performance burbuja: : 0.01 sec elapsed
```

```
library(tictoc)
tic("Performance sort: ")
sort(x)
```

```
##  [1] 20 38 42 48 51 53 62 70 73 97
```

```
toc()
```

```
## Performance sort: : 0 sec elapsed
```

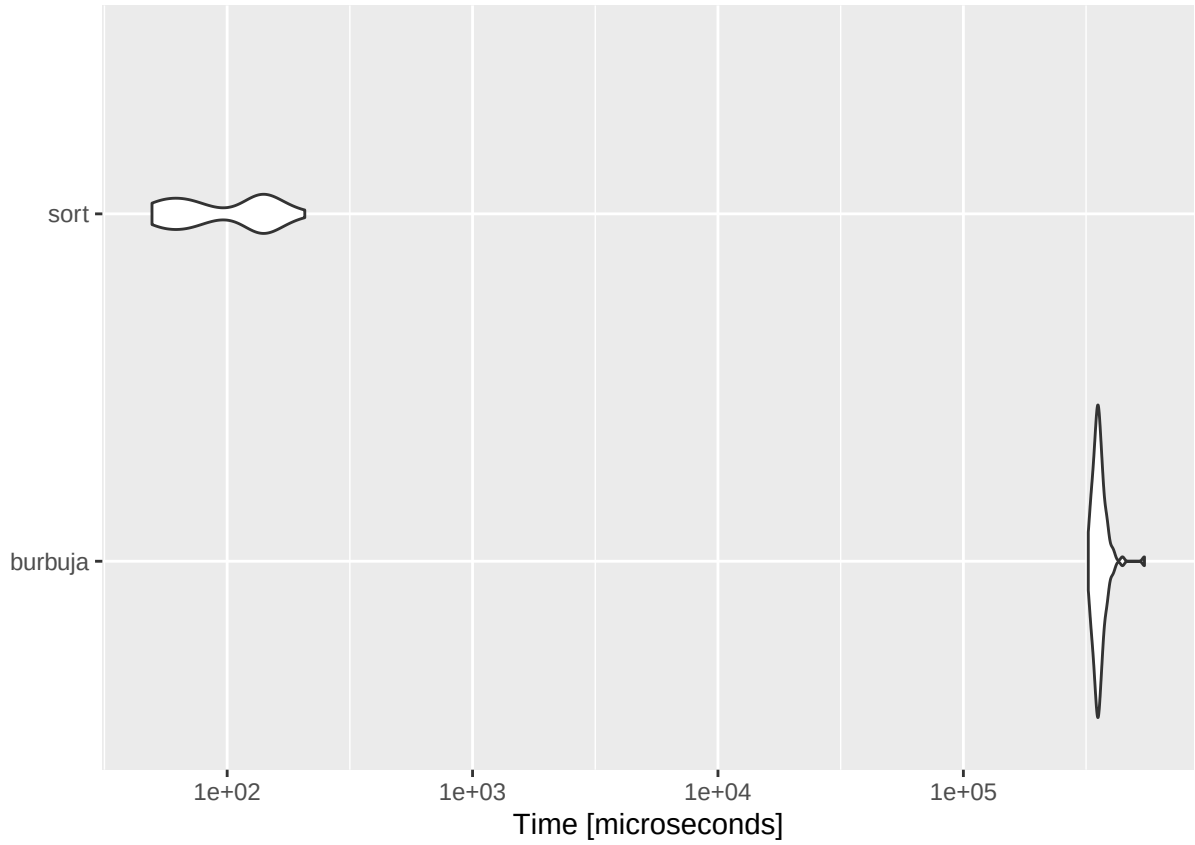
Con microbenchmark

```
library(microbenchmark)
x<-sample(1:2000,2000)

burbuja <- function(x){
  n<-length(x)
  for(j in 1:(n-1)){
    for(i in 1:(n-j)){
      if(x[i]>x[i+1]){
        temp<-x[i]
        x[i]<-x[i+1]
        x[i+1]<-temp
      }
    }
  }
  return(x)
}

mbm <- microbenchmark("burbuja" = { res<-burbuja(x) },
                      "sort" = {
                        sort(x)
                      })

mbm
```

```
## Unit: microseconds
##     expr      min       lq       mean   median        uq      max neval
##  burbuja 322890.1 343741.0 358946.495 354873.6 367203.95 547002.1   100
##     sort     49.4     64.2    104.408     98.9    138.15    207.2   100
```

```
library(ggplot2)
autoplot(mbm)
```

```
## Coordinate system already present. Adding new coordinate system, which will replace the e
```



Caso de estudio : Progresión del Covid

A continuación se utilizarán las herramientas aprendidas durante el curso para realizar el ordenamiento y presentación de datos oficiales de casos de covid en distintos países. Se analizarán los países Brasil y Perú. La información obtenida se puede encontrar en las siguientes páginas web: https://covid19.minsa.gob.pe/sala_situacional.asp y https://covid.saude.gov.br/

Progresión de casos en Brasil

Con los datos obtenidos se realizará un gráfico lineal.

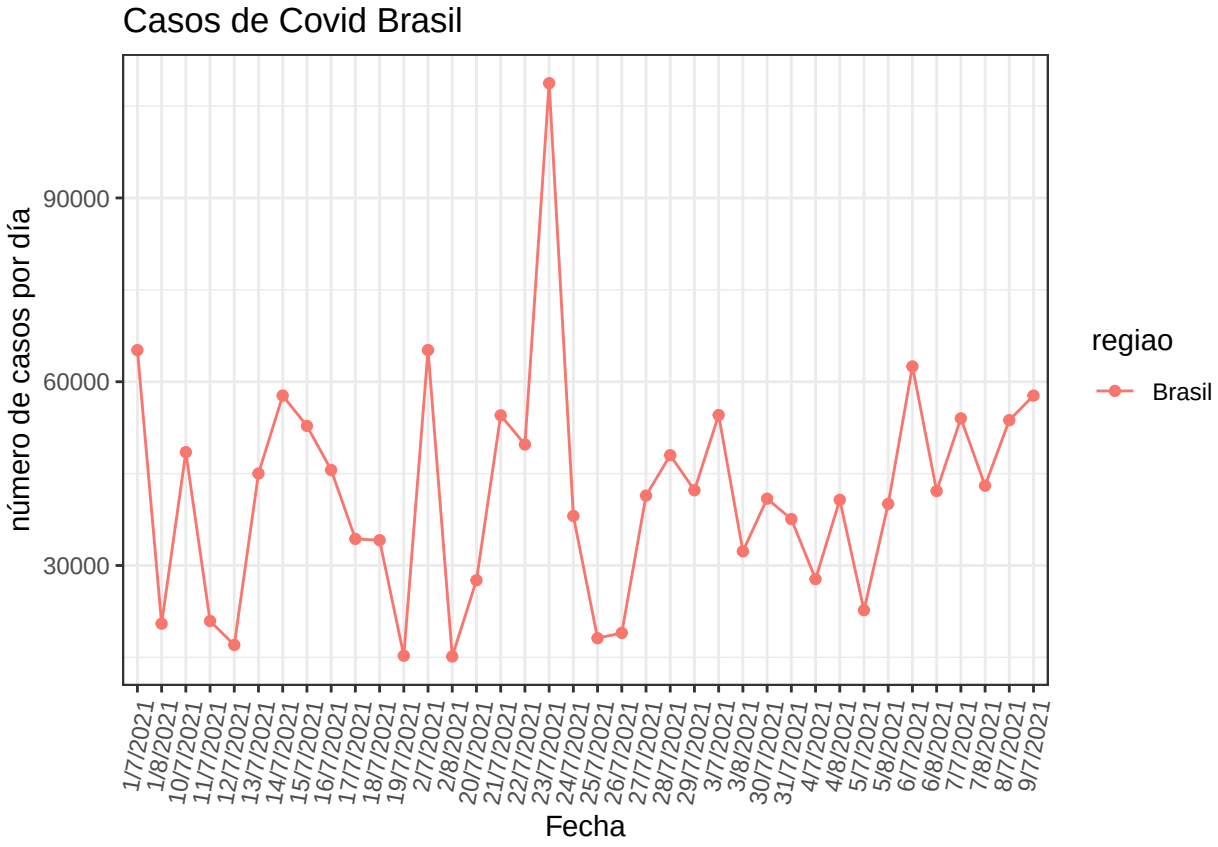
Para iniciar serán necesarios dos paquetes: install packages (dplyr) install packages (ggplot)

En primer lugar tomaremos los datos del archivo csv, luego seleccionaremos solo los datos que queremos analizar y por último utilizando la herramienta ggplot realizaremos un gráfico lineal.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)

#seleccionamos el directorio donde guardamos el archivo
setwd(dir = "C:/Users/Gabriel/Documents/R/ProgramacionenR")
datos <- read.table("Brasiljulio.csv", sep = ";", header = TRUE, dec = ".")

datossel <- datos %>% select(regiao,data,casosNovos,casosAcumulado)

datossel %>% ggplot(aes(x=data, y= casosNovos, group=regiao,color=regiao,fill=regiao)) +
            geom_point() + geom_line() +
            ggtitle("Casos de Covid Brasil") +
            theme_bw() +
            xlab("Fecha") +
            ylab("número de casos por día") +
           # scale_x_date(date_breaks = "10 days") +
            theme(axis.text.x = element_text(angle= 80, hjust = 1))
```



Casos por región en Perú

Para el siguiente análisis se seguirá el mismo procedimiento, pero esta vez se realizará un gráfico de barras donde se indicarán la cantidad de casos por región durante el mes de Julio.

```
library(dplyr)
library(ggplot2)

#seleccionamos el directorio donde guardamos el archivo
setwd(dir = "C:/Users/Gabriel/Documents/R/ProgramacionenR")
datos <- read.table("Peru.csv", sep = ";", header = TRUE, dec = ".")
#datos


#x <- datos %>% select(REGION, Confirmado)
```

```
ggplot(data=datos, aes(x=REGION, y=Confirmado, fill=Pais)) + geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle= 90, hjust = 1)) +
  ggtitle("Cantidad de casos por Región")
```