David Fontaneda Arenas
Elisa Seisdedos Castro

Proyecto de análisis de datos con R

Estadística aplicada - Primer Curso del Grado en Bioquímica (2019/20)

Ejercicio 1

1. Hemos utilizado las funciones unique() y cumsum() para extraer las fechas del data frame y calcular los acumulados respectivamente. Explica qué hace cada función. Si no puedes deducirlo del contexto siempre puedes utilizar la ayuda de R.

La función unique() nos elimina los datos duplicados de un vector o data frame. En este caso, por ejemplo, de la lista de todas las fechas para cada país, obtenemos una lista con todas las fechas, apareciendo una sola vez cada una.

La función cumsum() devuelve un vector con la suma cumulativa de los elementos que pasemos como argumento. Por ejemplo, si los contagios diarios son (3,4,5), obtendremos el vector (3,7,12).

2. Antes de centrarnos en los gráficos vamos a hacer un poco de "gimnasia" con los datos.
a) ¿Qué fecha era el decimoquinto día del estudio? ¿Cuántos contagios se registraron el 23 de marzo?

```
> fechas[15]
[1] "2020-02-05"
```

El decimoquinto día del estudio fue el 5 de Febrero.

```
> contagios.esp[which(fechas == "2020-03-23")]
[1] 6368
```

El 23 de marzo se registraron 6368 contagios.

b) ¿Qué posición ocupa el 3 de marzo?

```
> which(fechas == "2020-03-03")
[1] 42
```

El 3 de marzo fue el día 42 del estudio.

c) ¿Cuántos contagios se registraron entre el 23 de marzo y el 27 de marzo?

```
> sum(contagios.esp[which(fechas == "2020-03-23"):which(fechas == "2020-03-27")])
[1] 36951
```

Entre el 23 de marzo y el 27 de marzo de registraron 36951 contagios.

d) ¿Cuál fue el día con más contagios registrados y a cuánto ascendió la cifra? Puedes explorar la función which.max()

```
> fechas[which.max(contagios.esp)]
[1] "2020-03-25"
> contagios.esp[which.max(contagios.esp)]
[1] 9630
```

El día con más contagios fue el 25 de marzo, en el que se dieron 9630 contagios.

e) ¿Cuántos días se han registrado más de 8000 contagios en la serie temporal?

```
> length(which(contagios.esp >=8000))
[1] 3
```

Se han registrado 3 días con más de 8000 contagios.

f) ¿Cuántos datos hay en la tabla coronavirus referentes a España?

```
> length(coronavirus$cases[coronavirus$Country.Region=="Spain"])
[1] 255
```

Hay 255 datos referentes a España, dado que hay 85 días de datos, y para cada día se dan los infectados, recuperados y fallecidos.

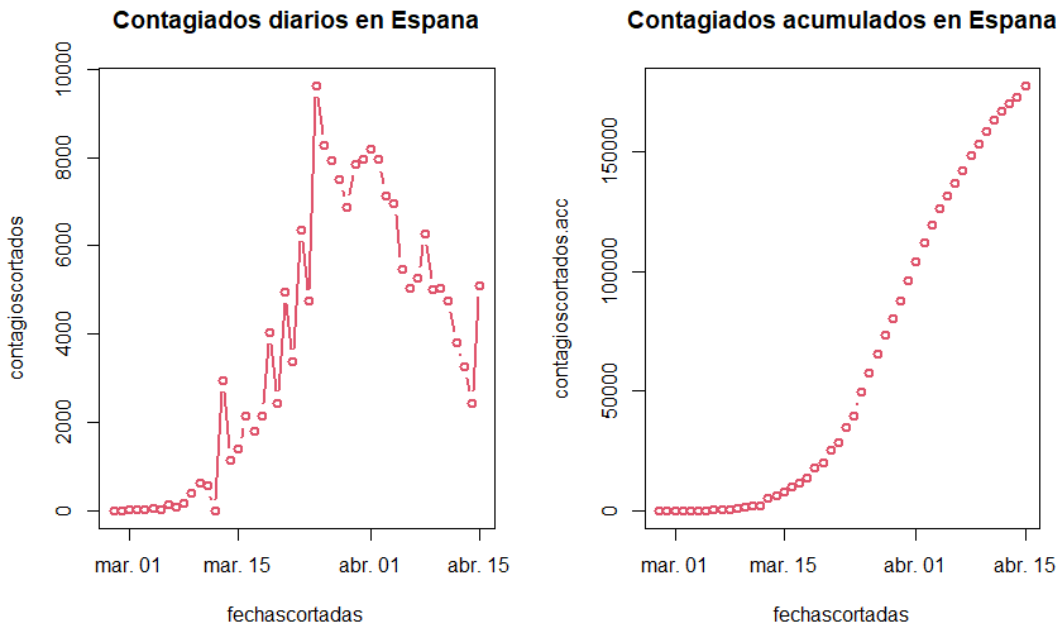
3. Los gráficos tienen que ser informativos y en ambos casos hay una gran superficie en blanco (y, por tanto, no informativa). Ajusta las figuras para que resulten más informativas: puedes pintar los datos desde una fecha concreta o cambiar la escala de los gráficos.

Recortaremos los datos para que solo se tengan en cuenta las fechas a partir del momento en el que hubo más de 10 contagios, el 28 de febrero.

```
fechascortadas=fechas[which(contagios.esp>10)[1]:length(fechas)]
contagioscortados=contagios.esp[which(contagios.esp>10)[1]:length(contagios.esp)]
contagioscortados.acc=contagios.esp.acc[which(contagios.esp>10)[1]:length(contagios.esp.acc)]
plot(fechascortadas,contagioscortados,type = "b", lwd = 2, col = 2,
     main = "Contagiados diarios en Espana")

plot(fechascortadas,contagioscortados.acc,type = "b", lwd = 2, col = 2,
     main = "Contagiados acumulados en Espana")
```

Dejando así ambas gráficas:



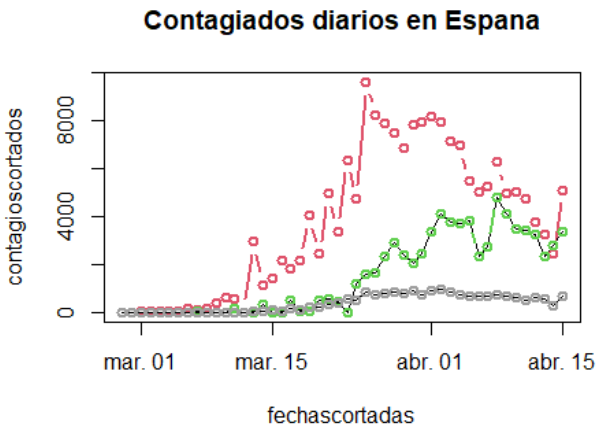
4. Añade a las gráficas del apartado anterior los fallecidos y los recuperados (tanto diarios como acumulados). Recuerda que una vez que está hecho un gráfico con plot() puedes añadir líneas con la función lines() .

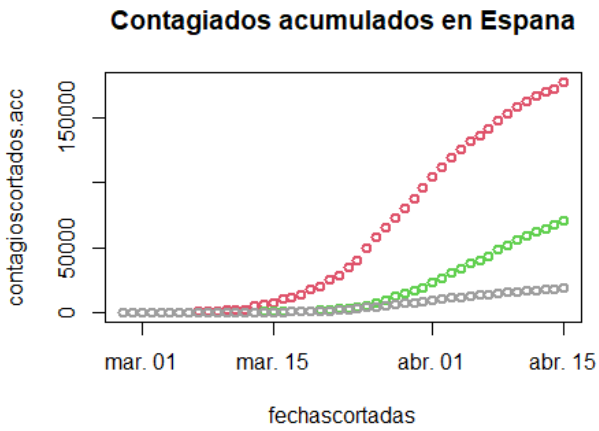
```
recuperados.esp=coronavirus$cases[coronavirus$type=="recovered"
                                  & coronavirus$Country.Region=="Spain"]
recuperados.esp.acc=cumsum(recuperados.esp)
recuperadoscortados=recuperados.esp[which(contagios.esp>10)[1]:length(contagios.esp)]
recuperadoscortados.acc=recuperados.esp.acc[which(contagios.esp>10)[1]:length(contagios.esp)]

fallecidos.esp=coronavirus$cases[coronavirus$type=="death"
                                 & coronavirus$Country.Region=="Spain"]
fallecidos.esp.acc=cumsum(fallecidos.esp)
fallecidoscortados=fallecidos.esp[which(contagios.esp>10)[1]:length(contagios.esp)]
fallecidoscortados.acc=fallecidos.esp.acc[which(contagios.esp>10)[1]:length(contagios.esp)]
dev.off()

plot(fechascortadas,contagioscortados,type = "b", lwd = 2, col = 2,
     main = "Contagiados diarios en Espana")
lines(fechascortadas,recuperadoscortados,type = "b",lwd = 2, col = 3)
lines(fechascortadas,fallecidoscortados,type = "b",lwd = 2, col = 8)

plot(fechascortadas,contagioscortados.acc,type = "b", lwd = 2, col = 2,
     main = "Contagiados acumulados en Espana")
lines(fechascortadas,recuperadoscortados.acc,type = "b",lwd = 2, col = 3)
lines(fechascortadas,fallecidoscortados.acc,type = "b",lwd = 2, col = 8)
```





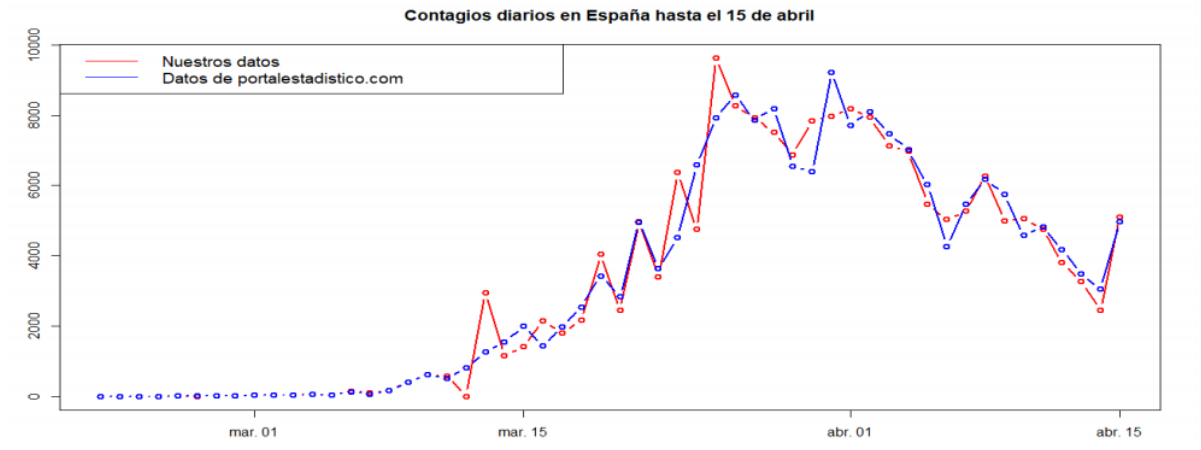
Representando en rojo los contagiados, en verde los recuperados y en gris los fallecidos.

5. Analiza los gráficos, ¿encuentras alguna información de interés? ¿es razonable lo que ocurre o hay algo que te llame la atención? Aparte, ¿sabrías decir cuál de los gráficos elementales que estudiamos en el tema 1 esta "encubierto" en estos gráficos?

Es razonable que los contagios suban y posteriormente bajen. Las muertes representan aparentemente un porcentaje de los contagiados en cada fecha, y los picos de contagiados se manifiestan como picos de recuperados unos días después. Sin embargo, hay un gran pico al final como último dato que habría que estudiar.

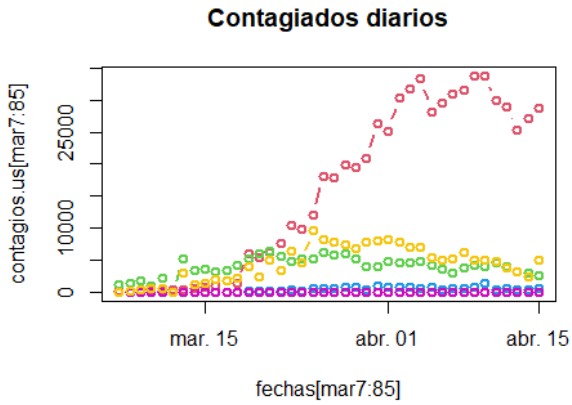
Son gráficos de líneas: representan el número de infectados en una serie temporal, mediante puntos conectados con líneas.

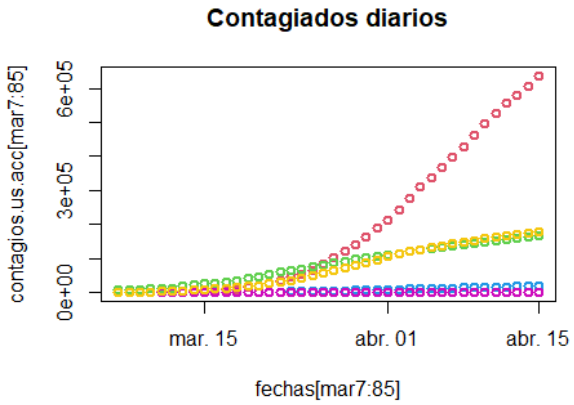
6. En la siguiente figura se comparan nuestros datos de contagio (en rojo) con los de www.portalestadistico.com (en azul), que se supone que están extraídos de los datos oficiales. ¿Encuentras diferencias significativas? ¿Sabrías plantear alguna hipótesis sobre las diferencias entre las curvas? El total de contagiados es de 177644 y 177517 personas respectivamente.



A pesar de las diferencias, las tendencias son prácticamente idénticas y reflejan la misma información. Dado que el total es prácticamente el mismo y solo varía el día al que fue asignado el positivo, es posible que nuestros datos reflejen la fecha en la que el test dio positivo, mientras que los datos de portalestadístico.com reflejen el día en el que se extrajeron las muestras del paciente, que puede haber sido del mismo día o de días anteriores.

7. También es interesante la comparativa entre países. Dibuja las curvas de contagiados diarios y acumulados para España, Italia ("Italy"), Estados Unidos ("US"), Andorra ("Andorra") y otro país de tu elección a partir del 7 de marzo. Nota: Ten cuidado con la forma de pintar las gráficas para que no se corten valores y con el país que elijas ya que hay países como China, Reino Unido, Canadá o Australia que tienen datos por provincias, por lo que si los quieres usar tendrás que añadir una condición más a la selección. Puedes ver cuántos datos hay de cada país con la tabla de frecuencias table(coronavirus$Country.Region) .



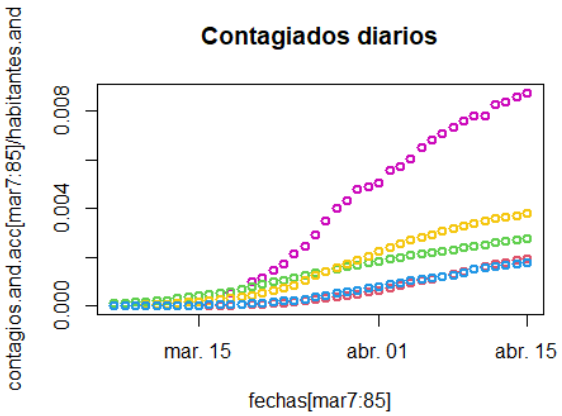


Rojo: US; Azul: Portugal; Morado: Andorra; Amarillo: España; Verde: Italia

8. Analiza los gráficos, ¿qué conclusiones sacas de la comparativa de países?

En estas gráficas podemos ver el crecimiento de los infectados. Podemos ver que, en el caso de Estados unidos, los casos han crecido a una velocidad muy alta, mientras que en el resto de países ha tendido a estabilizarse. Como se ve en el gráfico de los casos acumulados, Andorra y Portugal se han estabilizado en un número de casos diarios bajo. En España e Italia los casos diarios son algo mayores, y en Estados unidos los casos diarios se han estabilizado en un número mucho más alto.

9. Hasta ahora hemos considerado solo frecuencias absolutas y puede que esto no sea muy justo. Repite la gráfica de los contagios acumulados con los contagios/habitante. ¿Mantienes las mismas conclusiones? ¿Se te ocurre alguna explicación para el comportamiento de los datos de Andorra?

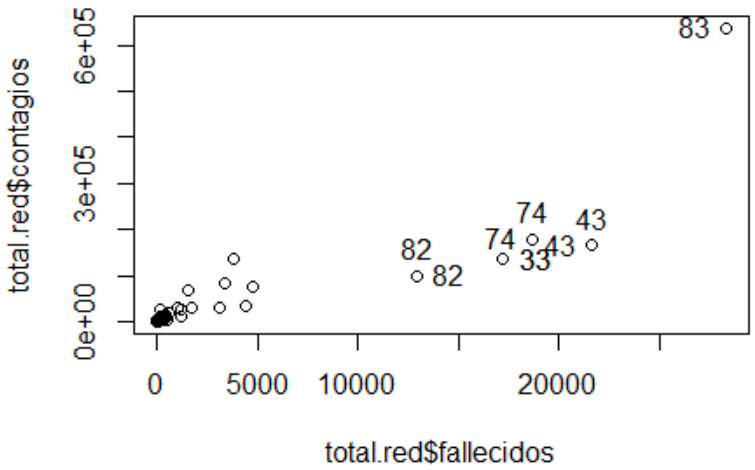


Rojo: US; Azul: Portugal; Morado: Andorra; Amarillo: España; Verde: Italia

Al normalizar vemos una información totalmente distinta. España e Italia tienen una tendencia similar, con más casos por habitante que Portugal o Estados unidos. En Andorra podemos observar que los contagios por habitante son muy altos, probablemente esto se pueda explicar teniendo en cuenta, por ejemplo, que es un país con una gran densidad poblacional (162 habitantes por Km^2) en comparación con por ejemplo la de Estados unidos (33 habitantes por Km^2), pero habría que explorar con más detalle el porqué de este resultado.

Ejercicio 2

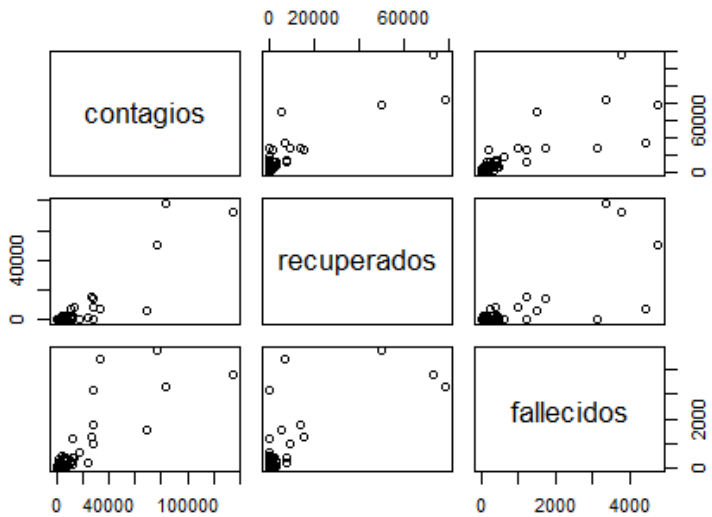
1. Dibujar el diagrama de dispersión de los fallecidos frente a los contagiados. ¿Hay algún valor atípico? Si es así identifícalo.



Eliminamos los valores atípicos que se alejan demasiado del punto en el que se encuentran la mayoría de los datos.

2. Eliminar los posibles atípicos y dibujar el diagrama de pares (o matriz de diagramas de dispersión) con todas las variables numéricas. Indicar si hay relaciones y de qué tipo entre las variables, o si se produce algún efecto extraño.

```
> pairs(total.red[-c(33,43,74,82,83),2:4])
```



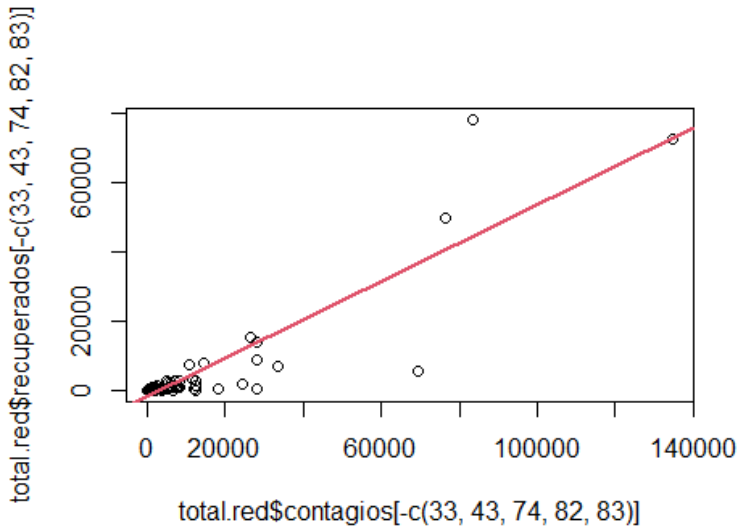
Se puede intuir en todos los casos una relación lineal, pero cuando los valores se vuelven grandes, aparece una gran variabilidad y los datos se aparecen en forma de cono.

3. Calcula la matriz de correlaciones de todas las variables numéricas. Relaciona los resultados con el diagrama anterior. ¿Crees que estos datos son adecuados para ajustar un modelo de regresión lineal? Si en alguno de los casos te parece razonable, dibuja el diagrama de dispersión correspondiente, calcula la recta de regresión y píntala en el diagrama.

```
> cor(total.red[-c(33,43,74,82,83),2:4])
            contagios recuperados fallecidos
contagios   1.0000000   0.8879893  0.8314033
recuperados 0.8879893   1.0000000  0.7445912
fallecidos  0.8314033   0.7445912  1.0000000
```

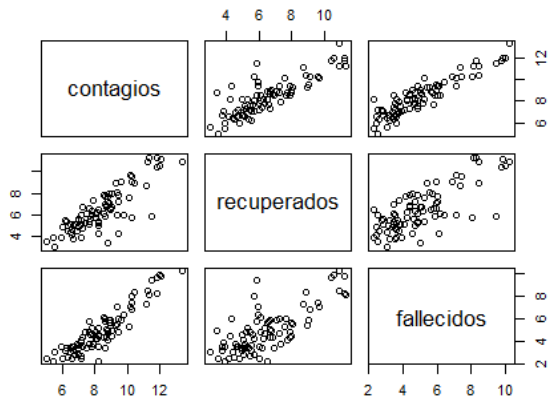
Podemos observar que hay una correlación positiva, tal y como podíamos intuir por la pendiente aparente de los diagramas anteriores, pero no se ajusta totalmente a un modelo lineal. El que más puede aproximarse, tal y como podemos ver en las covarianzas, es el de contagios frente a recuperados:

```
> plot(total.red$contagios[-c(33,43,74,82,83)],total.red$recuperados[-c(33,43,74,82,83)])
> recta= lm(total.red$recuperados[-c(33,43,74,82,83)] ~ total.red$contagios[-c(33,43,74,82,83)])
> abline(recta, col=2, lwd = 2)
```



Podemos ver que la recta se ajusta a los datos, pero sigue habiendo datos muy alejados de ella, especialmente a medida que nos alejamos del origen de coordenadas.

4. Posiblemente te hayan aparecido nubes de puntos "en forma de cono" con muy poca variabilidad al inicio y mucha al final. Esto es efecto de la heterocedasticidad (las varianzas no son iguales en los distintos puntos). Para eliminar este efecto y "linealizar" los datos se puede aplicar una transformación de doble logaritmo, es decir, reemplazar las variables originales por sus logaritmos. Prueba a hacer esta transformación (con todos los datos, sin excluir los atípicos) y a recalcular el diagrama de pares y la matriz de correlaciones con los logaritmos de las variables. Comenta los resultados. ¿Qué ha ocurrido con el/los atípicos?

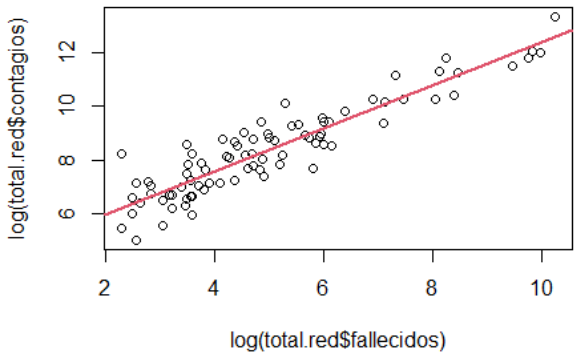


```
> cor(log(total.red[,2:4]))
            contagios recuperados fallecidos
contagios   1.0000000   0.8397233  0.9198195
recuperados 0.8397233   1.0000000  0.7539619
fallecidos  0.9198195   0.7539619  1.0000000
```

Los datos ahora se parecen mucho más a una recta. Si observamos la covarianza de los contagios con los fallecidos, podemos observar que es muy alta, y en el diagrama de dispersión correspondiente podemos ver que efectivamente la línea es muy definida.

Los atípicos han desaparecido casi por completo en la mayoría de gráficos.

5. Calcula la recta de regresión del logaritmo del número de fallecidos en función del logaritmo del número de contagiados (los datos transformados). ¿Es bueno el ajuste? Dando por bueno el ajuste, ¿cuántos fallecidos cabría esperar en un país con medio millón de contagiados? ¿Y con un millón?



Podemos ver que los datos se ajustan correctamente a la recta, por lo que podremos usar este ajuste para predecir datos.

En un país con 500.000 contagiados, usando la ecuación explicita de la recta:

```
> contagios.recta=500000
> fallecidos.recta=exp(recta$coefficients[2]*log(contagios.recta)+recta$coefficients[1])
> as.numeric(fallecidos.recta)
[1] 22226.6
```

Obtenemos que habrá aproximadamente 22.200 fallecidos.

En un país con 1.000.000 de contagiados, usando el mismo método que antes:

```
> contagios.recta=1000000
> fallecidos.recta=exp(recta$coefficients[2]*log(contagios.recta)+recta$coefficients[1])
> as.numeric(fallecidos.recta)
[1] 46186.44
```

Obtenemos que habrá aproximadamente 46.200 fallecidos, poco más del doble que en el ejemplo anterior.

Ejercicio 3

1. Cargad los datos de manera adecuada. Hay que tener en cuenta que las funciones de lectura de ficheros usuales no funcionan bien con los datos de Excel.

a. ¿Cómo interpretarías los valores de la variable PCR?

Los valores de esta variable son "NEG", "POS" y "NO REALIZADA", es decir, hay tres posibilidades: que la prueba dé negativo, que dé positivo o que no se realice esa prueba al paciente. En el caso de nuestros datos, 8 pruebas han dado negativo, 8 han dado positivo y para 9 pacientes no se ha realizado PCR.

b. En esta fase es muy común retocar la base de datos eliminando variables irrelevantes o redundantes, creando variables sintéticas, reduciendo la dimensión, etc. En este caso no vamos a cambiar nada, pero si tuvieras que eliminar dos variables ¿cuáles serían?

Si tuviéramos que eliminar dos variables, serían "IGM" e "IGG", los resultados de los test rápidos, porque aparecen englobados en la variable "TEST", la información es algo redundante. En cambio, el resto de variables son necesarias para el diagnóstico de covid-19.

2. Describir el tipo (cuantitativo, cualitativo, discreto, continuo,. . .) y posibles valores de las variables *LEUCOCITOS*, *PLAQUETAS*, *DIMERO D* y *COVID*.

LEUCOCITOS es una variable cuantitativa continua, ya que se trata de una característica cuantificable en la que los valores se agrupan en intervalos. Sus posibles valores, según la muestra, se encuentran entre 3.10 y 11 (miles/mm^3).

PLAQUETAS es también una variable cuantitativa continua, con valores entre 100 y 365 (miles/mm^3).

DIMERO D es una variable cuantitativa continua, cuyos valores se encuentran entre 0.0 y 0.8 (µg/ml).

COVID es una variable cualitativa, ya que los valores no son números. Sus posibles valores son "sí" y "no".

3. Elegir y calcular los estadísticos descriptivos adecuados para cada variable (de las anteriores) según su tipo.

Para las variables cuantitativas, es adecuado calcular media, mediana, varianza y desviación típica (como se calculan en R, se obtienen la cuasi-varianza y la cuasi-desviación típica).

DIMERO D

```
> mean(analisisCOVID$DIMERO_D)
[1] 0.3684
> median(analisisCOVID$DIMERO_D)
[1] 0.4
> var(analisisCOVID$DIMERO_D)
[1] 0.047464
> sd(analisisCOVID$DIMERO_D)
[1] 0.2178623
```

Media = 0.3684. Mediana = 0.4. Cuasi-varianza = 0.047464. Cuasi-desviación típica = 0.2178623

PLAQUETAS

```
> mean(analisisCOVID$PLAQUETAS)
[1] 209.08
> median(analisisCOVID$PLAQUETAS)
[1] 190
> var(analisisCOVID$PLAQUETAS)
[1] 7430.743
> sd(analisisCOVID$PLAQUETAS)
[1] 86.20176
```

Media = 209.08. Mediana = 190. Cuasi-varianza = 7430.743. Cuasi-desviación típica = 86.20176

LEUCOCITOS

```
> mean(analisisCOVID$LEUCOCITOS)
[1] 5.4476
> median(analisisCOVID$LEUCOCITOS)
[1] 5
> var(analisisCOVID$LEUCOCITOS)
[1] 3.237161
> sd(analisisCOVID$LEUCOCITOS)
[1] 1.799211
```

Media = 5.4476. Mediana = 5. Cuasi-varianza = 3.237161. Cuasi-desviación típica = 1.799211

COVID: distribución de Bernoulli $\begin{cases} 1 \text{ si "SI"} \\ 0 \text{ si "NO"} \end{cases}$ p = probabilidad de éxito (SI) = $\frac{11}{25}$ = 0.44

4. Entre las variables *LEUCOCITOS* y *PLAQUETAS*, ¿cuál presenta una mayor dispersión/variabilidad?

La variable *PLAQUETAS* presenta mayor dispersión, ya que su desviación típica es mucho mayor que la desviación típica de la variable *LEUCOCITOS*, lo que quiere decir que los valores se alejan más de la media.

5. Representar gráficamente (de manera adecuada según el tipo) las variables *LINFOCITOS*, *PLAQUETAS* y *NEUMONIA*.

Plaquetas: histograma

```
> a=sort(analisisCOVID$PLAQUETAS)
> hist(a, main="PLAQUETAS",
+         col=c("red", "orange", "green", "skyblue", "blue","grey" ))
```

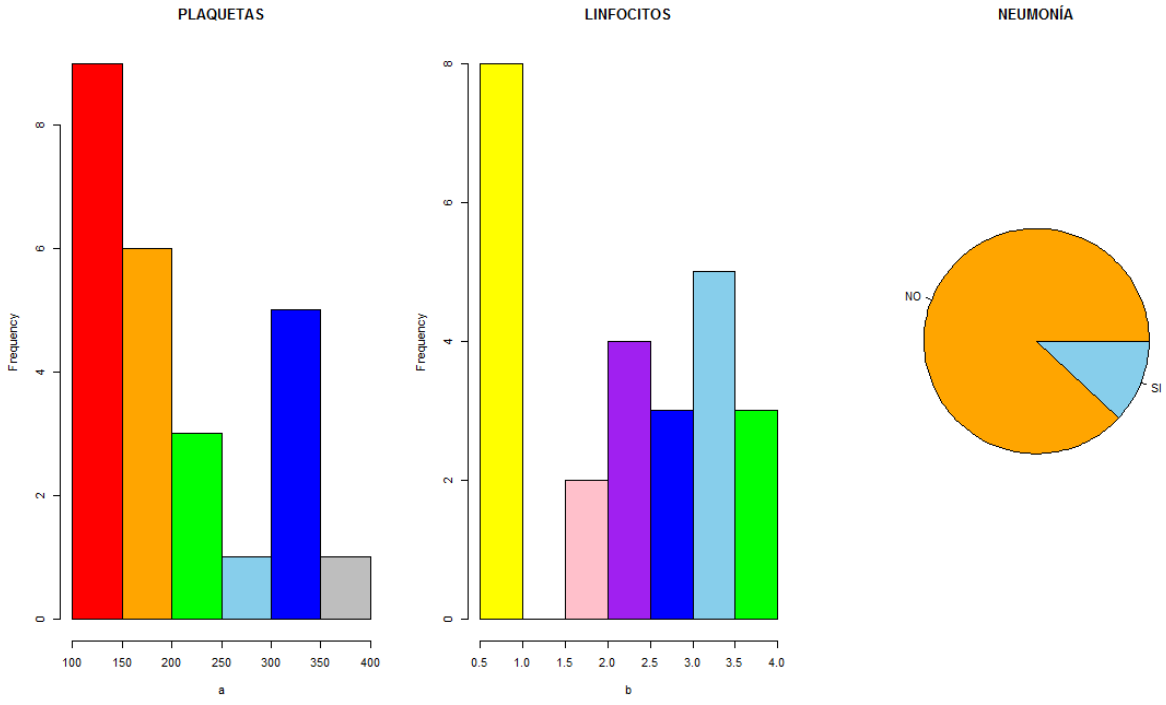
Linfocitos: histograma

```
> b=sort(analisisCOVID$LINFOCITOS)
> hist(b, main="LINFOCITOS",
+      col=c("yellow", "yellow", "pink", "purple", "blue", "skyblue", "green"))
```

Neumonía: gráfico circular

```
> #Neumonía: gráfico circular
> analisisCOVID$NEUMONIA
 [1] NO NO SI NO NO NO NO NO NO SI NO NO NO NO NO NO NO NO NO NO NO NO NO SI
Levels: NO SI
> table(analisisCOVID$NEUMONIA)

NO SI
22  3
> pie(table(analisisCOVID$NEUMONIA),
+     main="NEUMONÍA", col=c("orange", "skyblue"))
> #media y mediana de linfocitos
> mean(analisisCOVID$LINFOCITOS)
```



Interpretar los resultados y sacar conclusiones sobre las distribuciones de las variables. ¿Qué relación hay entre la media, la mediana y la forma de las distribuciones de las variables numéricas?

En cuanto a la variable *PLAQUETAS*, se observa que la distribución es asimétrica por la derecha, ya que hay una mayor dispersión hacia ese lado. Esto se confirma mediante la media y la mediana: como son distintas, existe asimetría, y como la media es mayor que la mediana, es asimetría por la derecha.

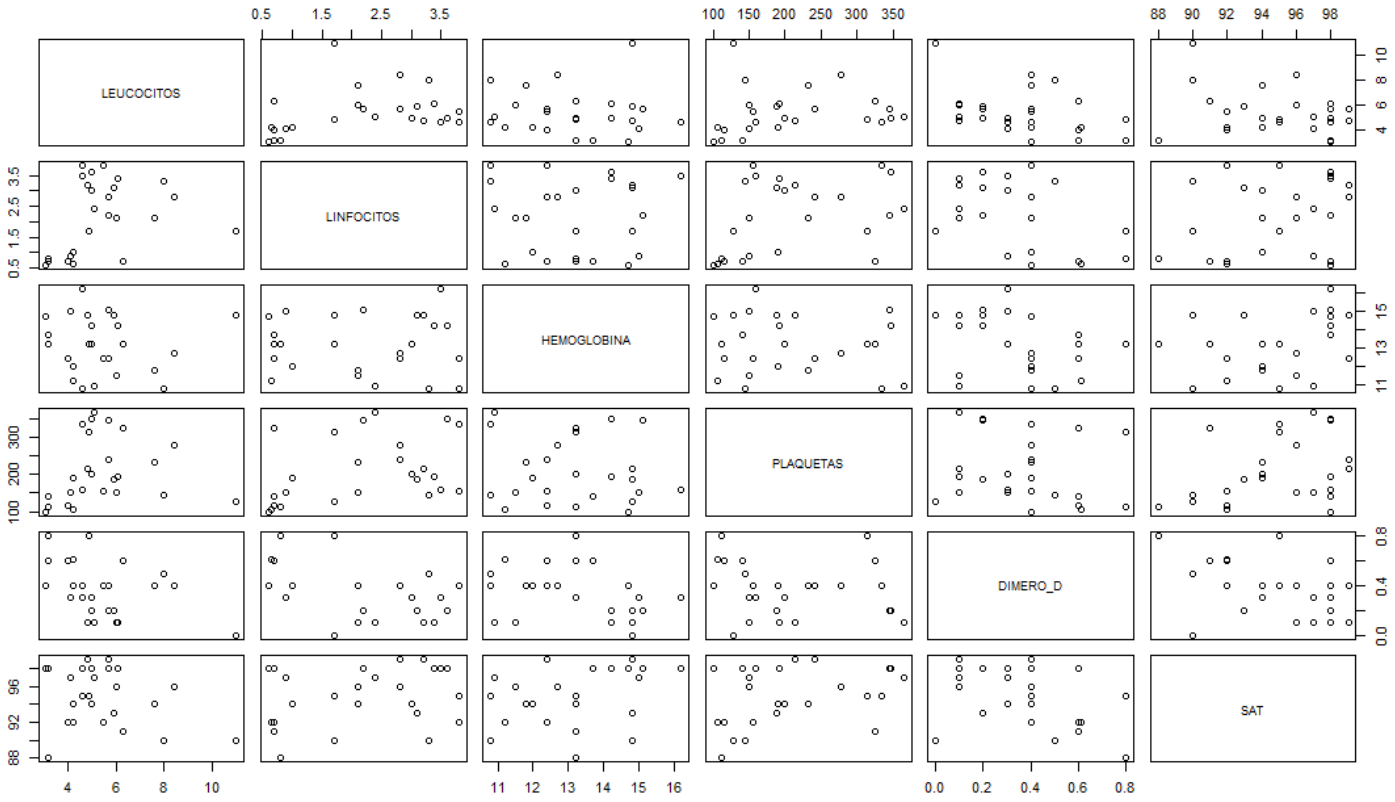
En cuanto a la variable *LINFOCITOS*, si se eliminan los valores contenidos en el intervalo (0.5, 1), que son valores atípicos, se observa cierta simetría en el histograma, que recuerda al de una distribución normal. Al estudiar la media (2.1812) y la mediana (2.2), se ve que son prácticamente iguales, lo que indica simetría.

En cuanto a la variable *NEUMONÍA*, no es una variable numérica, por lo que no se pueden estudiar su media y su mediana, pero viendo su representación gráfica, se puede afirmar que, en la muestra utilizada, hay una mayor proporción del valor "no".

Ejercicio 4

1. Dibujar el diagrama de pares (o matriz de diagramas de dispersión) con todas las variables numéricas. Indica si hay relaciones y de qué tipo entre los distintos pares.

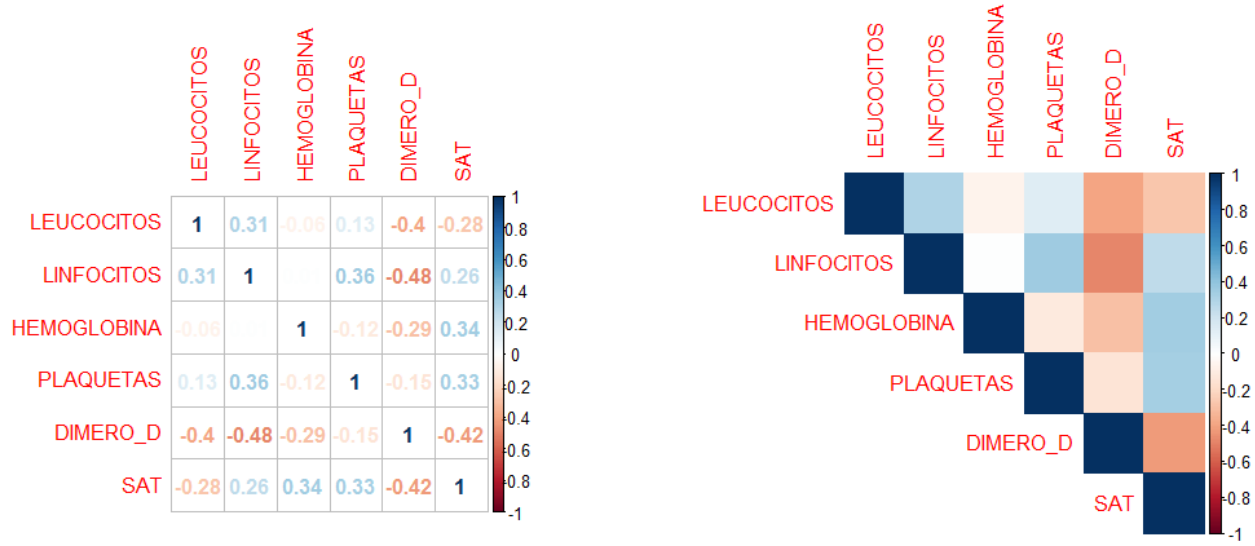
```
> variable_numerica=analisisCOVID[, c(5, 6, 7, 8, 9, 11)]
> view(variable_numerica)
> pairs(variable_numerica)
```



No se observa mucha relación entre las variables numéricas, aunque se puede intuir una relación lineal débil entre algunos pares de variables, como *LEUCOCITOS* y *LINFOCITOS*, *LINFOCITOS* y *PLAQUETAS* o *PLAQUETAS* y *SAT*.

2. Calcular la matriz de correlaciones para todas las variables numéricas. Relaciona los resultados con el diagrama anterior.

```
> correlaciones<-cor(variable_numerica)
> corrplot(correlaciones, method="number")
> corrplot(correlaciones, method="color", type="upper")
```



El coeficiente de correlación mide el grado de relación lineal entre dos variables: si vale 1 o -1, las valores están perfectamente alineados, y si vale 0, las variables son incorreladas.

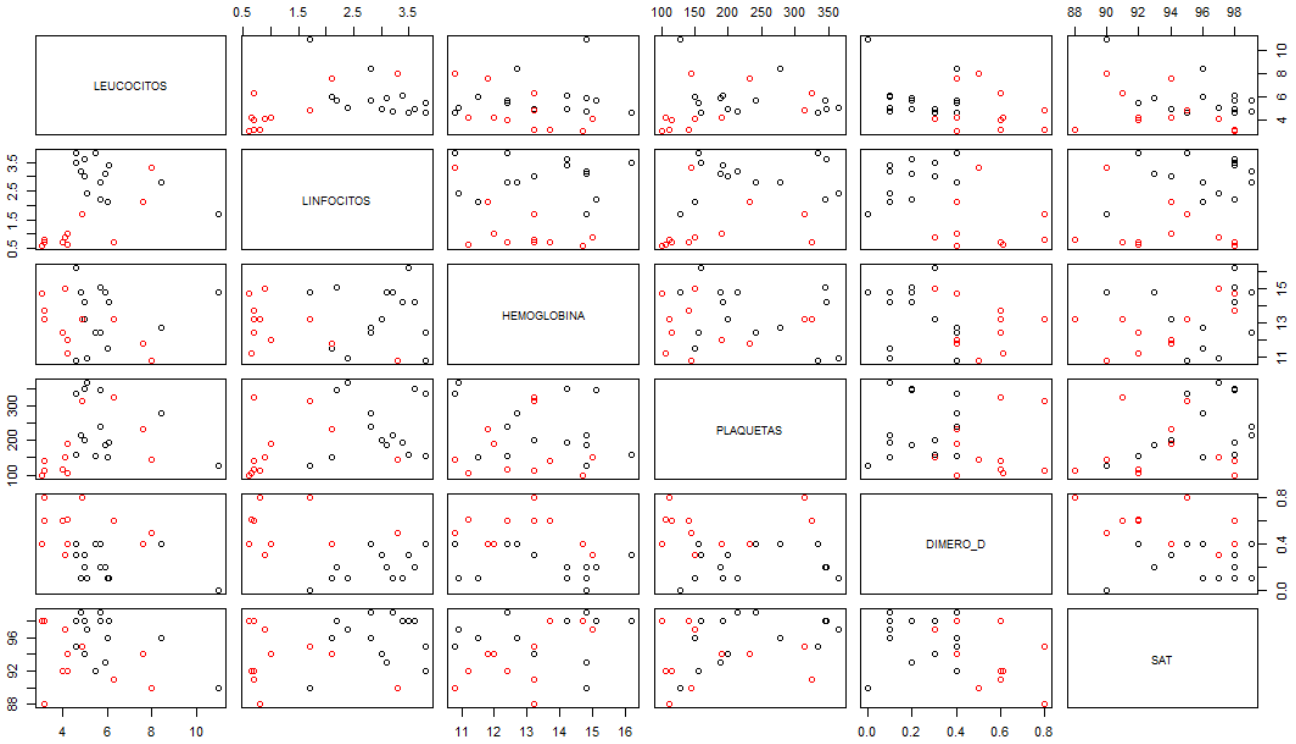
Como hemos mencionado en el apartado anterior, no se observa una relación lineal fuerte entre ningún par de variables, ya que el coeficiente de correlación más alto (en valor absoluto) no llega a 0.5.

3. Muchas veces es interesante segmentar o agrupar la información. En este caso, parece claro que puede ser útil ver si hay diferencias entre los que tienen la enfermedad y los que no. Para incluir esta información:

a. Colorea el diagrama de dispersión utilizando *COVID* como variable de agrupación.

Rojo = *COVID* positivo, negro = *COVID* negativo

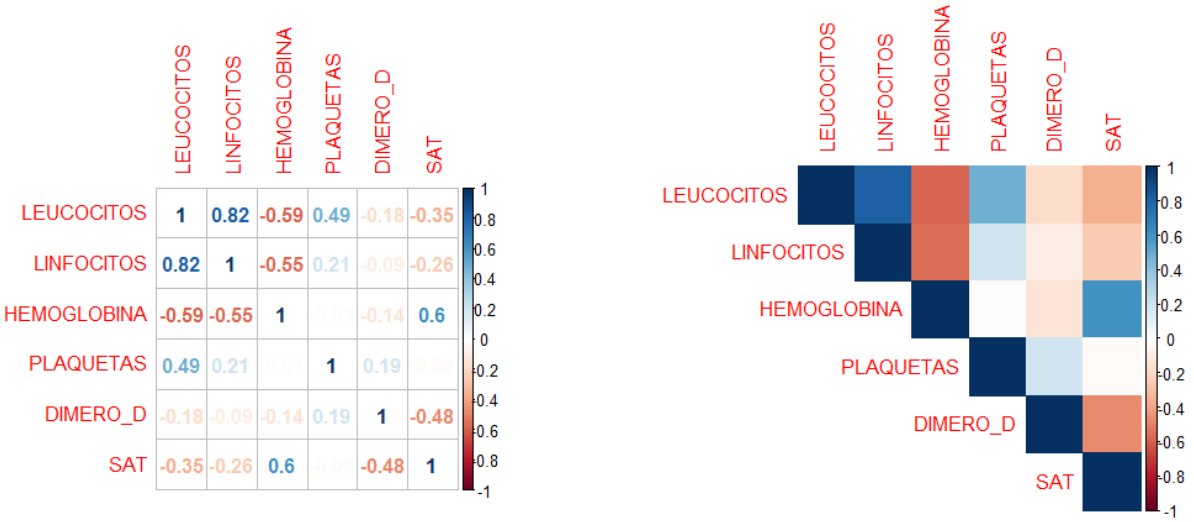
```
> pairs(variable_numerica, col= analisisCOVID$COVID)
```



b. Calcula las matrices de correlaciones solo con los pacientes con *COVID* positivo y después con los pacientes con *COVID* negativo.

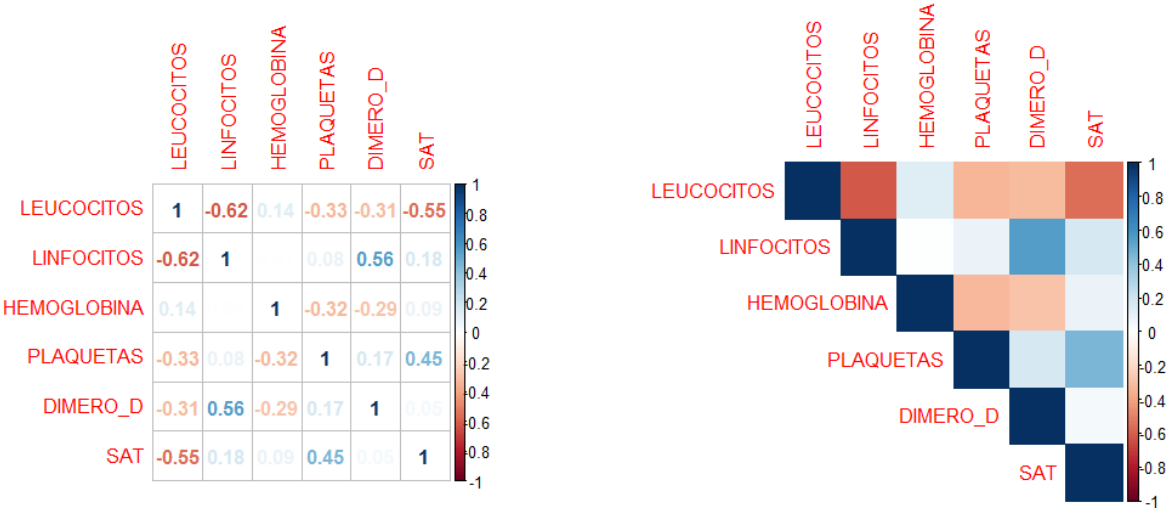
COVID positivo

```
> v.numerica_COVID=analisisCOVID[, c(5, 6, 7, 8, 9, 11, 13)]
> positivo=v.numerica_COVID[v.numerica_COVID$COVID=="SI",]
> positivo6=positivo[, c(1, 2, 3, 4, 5, 6)]
> Covid_positivo<-cor(positivo6)
> corrplot(Covid_positivo, method="number")
> corrplot(Covid_positivo, method="color", type="upper")
```



COVID negativo

```
> negativo=v.numerica_COVID[v.numerica_COVID$COVID=="NO",]
> negativo6=negativo[, c(1, 2, 3, 4, 5, 6)]
> Covid_negativo<-cor(negativo6)
> corrplot(Covid_negativo, method="number")
> corrplot(Covid_negativo, method="color", type="upper")
```



c. ¿Mantienes las mismas conclusiones? ¿Ha aparecido información nueva?

Al calcular las matrices de correlación solo con *COVID* positivo o solo con *COVID* negativo se observa que la relación lineal entre algunas variables se fortalece. Esto se debe a que, como hay dos subgrupos, cuando se recogen todos los datos en un mismo diagrama, la relación positiva anula la negativa y viceversa, y los coeficientes de correlación aparecen inferiores a lo que realmente son cuando se estudian los subgrupos por separado.

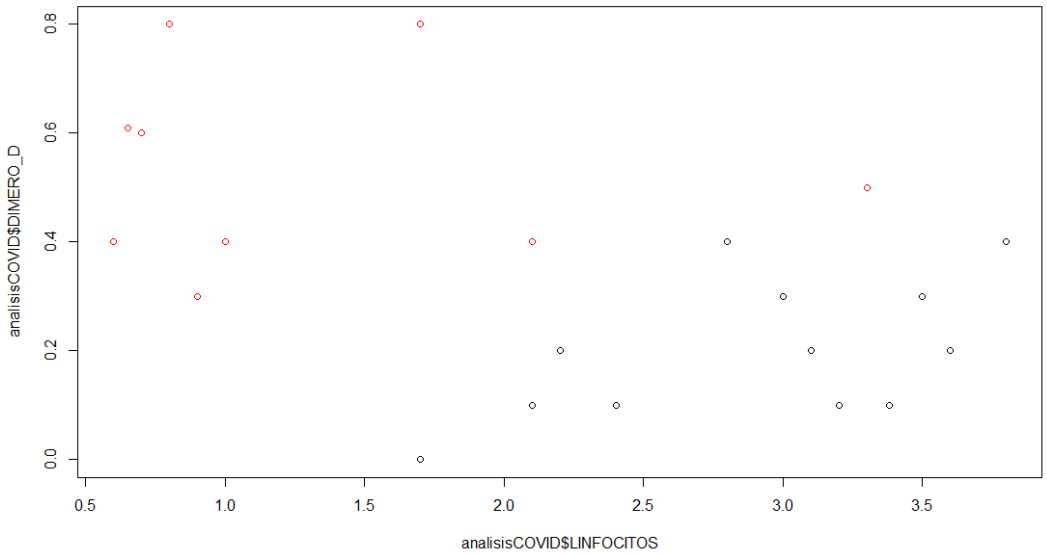
Un ejemplo es la relación entre *LEUCOCITOS* y *LINFOCITOS*: en la matriz de correlaciones para todas las variables numéricas vemos que el valor del coeficiente de correlación es 0.31. Sin embargo, en la matriz de correlaciones con pacientes con *COVID* positivo, el coeficiente vale 0.82, y en la matriz con pacientes con *COVID* negativo su valor es de -0.62; es decir, indican una relación lineal más fuerte entre ambas variables.

4. Un problema muy relevante es la identificación de factores que influyen en una determinada variable. En este caso nos podríamos preguntar qué indicadores de los análisis tienen una mayor relación con el virus, es decir, cuáles nos permitirían determinar con mayor precisión si una persona tiene coronavirus (o lo ha pasado) o no (sin contar los test).

a. Elige las dos variables numéricas que creas que predicen mejor la variable *COVID* (que separan mejor los puntos de cada color) y pinta su diagrama de dispersión.

Hemos seleccionado las variables *LINFOCITOS* y *DÍMERO D*, ya que los valores para *COVID* positivo y los valores para *COVID* negativo están bastante bien separados.

```
> plot(analisisCOVID$LINFOCITOS, analisisCOVID$DIMERO_D,
+      col=analisisCOVID$COVID)
```



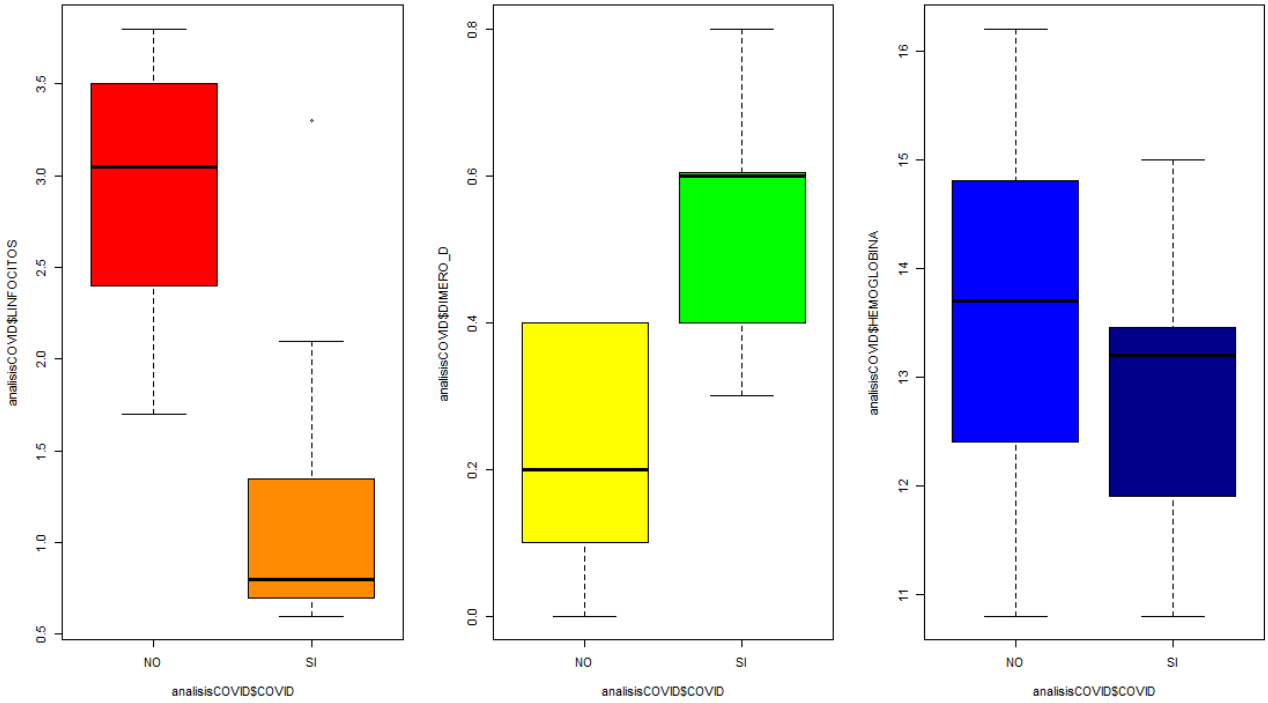
b. ¿Estas dos variables están muy correladas entre sí? ¿Te parece razonable que las dos variables que mejor separan estén así de relacionadas entre sí?

Si observamos la matriz de correlaciones sin distinguir entre *COVID* positivo y *COVID* negativo, ambas variables no están muy correladas, su coeficiente de correlación es -0.48. En cambio, si se estudia la correlación con COVID negativo, el coeficiente de correlación es 0.56, más cercano a 1, por lo que la correlación es más fuerte.

Esto es razonable, ya que si las variables permiten separar los datos con *COVID* positivo de los datos con *COVID* negativo, no pueden estar muy correladas entre sí (la relación lineal sería más fuerte y los puntos no se podrían separar tan bien, al estar más juntos entre ellos). Una vez dentro de cada subgrupo, ambas variables pueden estar más o menos correladas.

c. Para esas dos variables y una tercera cualquiera, dibuja los diagramas de cajas agrupando por la variable *COVID*. Interpreta los resultados, ¿confirmas tu decisión anterior?

```
> #diagrama LINFOCITOS
> boxplot(analisisCOVID$LINFOCITOS ~ analisisCOVID$COVID, analisisCOVID,
+         col = c("red", "darkorange"))
> #diagrama DIMERO_D
> boxplot(analisisCOVID$DIMERO_D ~ analisisCOVID$COVID, analisisCOVID,
+         col = c("yellow", "green"))
> #diagrama HEMOGLOBINA
> boxplot(analisisCOVID$HEMOGLOBINA ~ analisisCOVID$COVID, analisisCOVID,
+         col = c("blue", "darkblue"))
```



Los diagramas de cajas son de las variables *LINFOCITOS*, *DIMERO D* y *HEMOGLOBINA*, respectivamente. Viendo estos diagramas, confirmamos que las dos primeras variables no están muy correladas entre sí.

Además, para la variable *LINFOCITOS*, observamos que valores por debajo de 1.5 se corresponden con *COVID* positivo, mientras que para la variable DIMERO D, son valores superiores a 0.4 los que se corresponden con *COVID* positivo. En cambio, para la variable *HEMOGLOBINA*, no se pueden sacar valores que se correspondan con *COVID* positivo de forma clara.

d. Otra forma de ver qué variable está más relacionada con una variable objetivo (en este caso tener el virus o no) es medir la dependencia estadística. Nosotros sabemos que la correlación mide la dependencia (lineal) entre variables. Aunque en general la correlación no se puede calcular con variables discretas, las variables binarias (como *COVID*) son una excepción. Calcula las correlaciones entre todas las variables numéricas y la variable objetivo *COVID*. ¿Se confirma tu decisión o cambiarías algo?

```
> v.numerica_COVID$COVID=as.numeric(v.numerica_COVID$COVID)
> cor(v.numerica_COVID$LEUCOCITOS, v.numerica_COVID$COVID)
[1] -0.3260849
```

Correlación entre *LEUCOCITOS* y *COVID*: -0.3260849

```
> cor(v.numerica_COVID$LINFOCITOS, v.numerica_COVID$COVID)
[1] -0.7735681
```

Correlación entre *LINFOCITOS* y *COVID*: -0.7735681

```
> cor(v.numerica_COVID$HEMOGLOBINA, v.numerica_COVID$COVID)
[1] -0.1942156
```

Correlación entre *HEMOGLOBINA* y *COVID*: -0.1942156

```
> cor(v.numerica_COVID$PLAQUETAS, v.numerica_COVID$COVID)
[1] -0.3528976
```

Correlación entre *PLAQUETAS* y *COVID*: -0.3528976

```
> cor(v.numerica_COVID$DIMERO_D, v.numerica_COVID$COVID)
[1] 0.7390023
```

Correlación entre *DIMERO D* y *COVID*: 0.7390023

```
> cor(v.numerica_COVID$SAT, v.numerica_COVID$COVID)
[1] -0.3781568
```

Correlación entre *SAT* y *COVID*: -0.3781568

Observamos, efectivamente, que las correlaciones más fuertes son entre *LINFOCITOS* y *COVID* y entre *DIMERO D* y *COVID*. Los signos indican que *COVID* positivo se corresponde con menor cantidad de linfocitos y mayor de dímero D.

5. La aplicación directa de este tipo de razonamientos conduce al problema de clasificación. En este caso el problema de clasificación consistiría en asignar la clase más probable (sano o enfermo) para una nueva observación (nuevo paciente) en función de la información que tenemos (análisis) y del conocimiento previo (conjunto de datos bien clasificado). En esta asignatura no vamos a estudiar métodos para resolver el problema de clasificación, pero es algo que todos hacemos de manera natural todos los días. Por ejemplo, reciclar, determinar si un alimento está en mal estado o si un correo es spam. Por último, esto también se puede ver como un caso particular del problema de regresión en el que la variable respuesta (Y) es discreta.

Se han realizado análisis a dos nuevos pacientes, con la información obtenida en los apartados anteriores, ¿a cuál de las clases (*COVID* positivo y *COVID* negativo) es más probable que pertenezcan según sus resultados?

LEUCOCITOS	LINFOCITOS	HEMOGLOBINA	PLAQUETAS	DIMERO_D	NEUMONIA	SAT	FIEBRE
6,6	1,91	14,2	241	0,2	NO	97	SI
3,2	0,7	13,8	110	0,6	SI	93	SI

Algunos indicativos de que un paciente tiene covid son un número de linfocitos por encima de 1.5 y un nivel de dímero D por encima de 0.4.

El primer paciente nuevo tiene 1.91 miles de linfocitos por mm^3 y 0.2 µg de dímero D por ml, por lo que es probable que su resultado pertenezca a *COVID* negativo. En cambio, el segundo paciente tiene 0.7 miles de linfocitos por mm^3 y 0.6 µg de dímero D por ml, por lo que es probable que pertenezca a *COVID* positivo.

Ejercicio 5

1. Supongamos que el test de IgM y la PCR tienen que dar el mismo resultado, es decir, PCR indicaría el valor real al ser extremadamente fiable. Calcular la matriz de confusión con los datos de la clínica (utilizando solo los 16 para los que hay valores de la PCR) y calcular la sensibilidad y la especificidad de esta prueba IGM.

```
> pcr_ordenado<-arrange(analisisCOVID, analisisCOVID$PCR)
> view(pcr_ordenado)
> pcr=pcr_ordenado[c(1:8, 18:25), ]
> view(pcr)
> pcr$PCR<-factor(pcr$PCR)
> pcr$PCR
 [1] NEG NEG NEG NEG NEG NEG NEG NEG POS POS POS POS POS POS POS POS
Levels: NEG POS
> matriz<-table(pcr$IGM, pcr$PCR, dnn= c("estimado", "realidad"))
> matriz
```

estimado \ realidad	NEG	POS
NEG	6	3
POS	2	5

La realidad se corresponde con los resultados de la PCR, mientras que los valores estimados son los correspondientes a la prueba IGM, ya que es esta la que se quiere estudiar.

Sensibilidad = porcentaje de positivos entre los enfermos = $\frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos negativos}} = \frac{\text{verdaderos positivos}}{\text{enfermos}}$

Sensibilidad = $\frac{5}{5+3} = 0.625$

Especificidad = porcentaje de negativos entre los sanos = $\frac{\text{verdaderos negativos}}{\text{verdaderos negativos} + \text{falsos positivos}} = \frac{\text{verdaderos negativos}}{\text{sanos}}$

Especificidad = $\frac{6}{6+2} = 0.75$

2. Calcular la probabilidad de falso negativo y falso positivo para la prueba IgM con los datos de las especificaciones del test y con los de la clínica. ¿Cuál de los dos errores crees que es más importante y, por tanto, es prioritario minimizar?

Clínica

Falso negativo = $\frac{3}{5+3}$ = 0.375

Falso positivo = $\frac{2}{6+2}$ = 0.25

Especificaciones del test

Falso negativo = $\frac{3}{17+3}$ = 0.15

Falso positivo = $\frac{2}{48+2}$ = 0.04

Es más importante el falso negativo, porque se le dice a una persona enferma que está sana y, por tanto, no se toman las medidas necesarias para curarla y evitar que se siga extendiendo el virus. Es prioritario minimizar los falsos negativos.

3. Tanto la sensibilidad como la especificidad se pueden ver como proporciones, pero en este caso no podemos calcular el intervalo de confianza asociado como hemos visto en clase y por tanto, no podemos utilizar la función `prop.test()` (¿por qué?). Pero sí que podemos calcular el intervalo con la siguiente fórmula: si p = $\frac{k}{n}$ (casos favorables entre el total).

$$IC_{1-\alpha}(p) = \left(\frac{k}{(n-k+1)F_{\alpha/2;2(n-k+1);2k} + k}, \frac{(k+1)F_{\alpha/2;2(k+1);2(n-k)}}{(n-k) + (k+1)F_{\alpha/2;2(k+1);2(n-k)}} \right)$$

Usa esta fórmula para calcular el intervalo de confianza al 95% para la sensibilidad obtenida con los datos de la clínica. ¿Se podría decir que la sensibilidad de la prueba realizada es significativamente distinta a la de las instrucciones?

No se puede usar `prop.test()` porque la muestra no es lo suficientemente grande como para aplicar el Teorema Central del Límite y calcularlo como si fuera una distribución normal, aproximadamente.

k = casos favorables = 5
n = total = 8
$F_{\alpha/2,\ 2(n-k+1),\ 2k} = F_{0.025,\ 8,\ 10} = 3.855$
$F_{\alpha/2,\ 2(k+1),\ 2(n-k)} = F_{0.025,\ 12,\ 6} = 5.366$

Clínica: $IC_{0.95}(p) = (0.2449, 0.9148)$
Instrucciones: $IC_{0.95}(p) = (0.621, 0.968)$

Como hay muy pocos datos, el intervalo es muy amplio, pero eso no significa que la sensibilidad sea distinta. No se puede descartar que la sensibilidad calculada y la de las instrucciones sean iguales, por el numero de datos de los que disponemos.

4. Parece que el dímero D es un buen indicador de Covid, alcanzando niveles más altos en los casos positivos. Comprueba si el nivel medio de dímero D es significativamente más alto en las personas con COVID positivo con una confianza del 99%. Supón que las distribuciones son normales.

Primero, comprobamos la homocedasticidad ($\sigma_x = \sigma_y$)

```
> SI=analisisCOVID[analisisCOVID$COVID=="SI",]
> X=SI$DIMERO_D
> NO=analisisCOVID[analisisCOVID$COVID=="NO",]
> Y=NO$DIMERO_D
> mean(X)-mean(Y)
[1] 0.3177922
> #comprobar homocedasticidad
> var.test(X, Y, conf.level = 0.99)
```

```
        F test to compare two variances

data:  X and Y
F = 1.404, num df = 10, denom df = 13, p-value = 0.5572
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
 0.2912871 7.8463969
sample estimates:
ratio of variances
          1.403986
```

$H_0 : \sigma_1 = \sigma_2 \qquad R = \left\{ \frac{s_1^2}{s_2^2} \notin (F_{n_1-1;n_2-1;1-\alpha/2}, F_{n_1-1;n_2-1;\alpha/2}) \right\}$

$\frac{\text{num df}}{\text{denom df}}$ pertenece al intervalo, el p-valor es mayor que α (α=0.01) y 1 pertenece al intervalo de confianza (0.2913, 7.8464). Por todo ello, no se puede rechazar la igualdad de varianzas, con una confianza del 99% y los datos disponibles. Por lo tanto, se supone homocedasticidad.

Suponemos que las distribuciones son normales, independientes y homocedásticas.

```
> #t.test
> t.test(X, Y,
+        alternative="greater",
+        mu = 0, paired=FALSE, var.equal=TRUE,
+        conf.level=0.99)

        Two Sample t-test

data:  X and Y
t = 5.2607, df = 23, p-value = 1.225e-05
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 0.1667777       Inf
sample estimates:
mean of x mean of y
0.5463636 0.2285714
```

Como p-valor<0.01, se puede rechazar H_0. Con los datos disponibles y un 99% de confianza, hay suficiente evidencia de que el nivel medio de dímero D es más alto en las personas con COVID positivo.

5. Comprueba para el mismo nivel de significación si existen diferencias significativas en el nivel medio de hemoglobina. Supón normalidad. En ambos casos explica las hipótesis que has considerado y el criterio para tomar la decisión.

Primero, comprobamos la homocedasticidad ($\sigma_h = \sigma_e$)

```
> H=SI$HEMOGLOBINA
> E=NO$HEMOGLOBINA
> mean(H)-mean(E)
[1] -0.5922078
> #comprobar homocedasticidad
> var.test(H, E, conf.level = 0.99)

        F test to compare two variances

data:  H and E
F = 0.63494, num df = 10, denom df = 13, p-value = 0.4772
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
 0.1317329 3.5484868
sample estimates:
ratio of variances
          0.6349446
```

$H_0 : \sigma_1 = \sigma_2 \qquad R = \left\{ \frac{s_1^2}{s_2^2} \notin (F_{n_1-1;n_2-1;1-\alpha/2}, F_{n_1-1;n_2-1;\alpha/2}) \right\}$

$\frac{\text{num df}}{\text{denom df}}$ pertenece al intervalo, el p-valor es mayor que α (α=0.01) y 1 pertenece al intervalo de confianza (0.1317, 3.5484). Por todo ello, no se puede rechazar la igualdad de varianzas, con una confianza del 99% y los datos disponibles. Por lo tanto, se supone homocedasticidad.

Suponemos que las distribuciones son normales, independientes y homocedásticas.

```
> #t.test
> t.test(H, E,
+         alternative="greater",
+         mu = 0, paired=FALSE, var.equal=TRUE,
+         conf.level=0.99)

        Two Sample t-test

data:  H and E
t = -0.94951, df = 23, p-value = 0.8239
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 -2.151378       Inf
sample estimates:
mean of x mean of y
 12.83636  13.42857
```

Como p-valor>0.01, no se puede rechazar H_0. Con los datos disponibles y un 99% de confianza, no hay suficiente evidencia de que el nivel medio de hemoglobina es más alto en las personas con COVID positivo.

Ejercicio opcional

Como ejercicio opcional, hemos realizado una comparativa de los contagios totales cada día en varios países. En concreto, hemos empleado los países de la gráfica de la primera parte del proyecto (Ejercicio 1.7) y Corea del Sur, porque tiene una evolución interesante.

```
make_barchart_race(total.acc_smoother,
                   Country.Region,
                   cases,
                   title="Evolución de los contagios totales\n de CoVid-19 en países seleccionados",
                   caption="",
                   nframes=1500,
                   fps=30,
                   end_pause=60)
anim_save("out.gif")
```

Esta comparativa está recogida en una animación que está incluida en el archivo .R junto con el código.