**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?**

El decimoquinto dia del estudio fue el 5 de Febrero.



El 23 de marzo se registraron 6368 contagios.

**b) ¿Qué posición ocupa el 3 de marzo?**



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?**

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 cuanto ascendió la cifra? Puedes explorar la función which.max()**





El día con mas 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?**

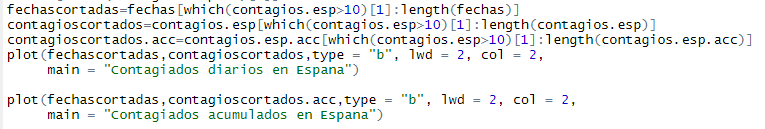
Se han registrado 3 días con más de 8000 contagios.

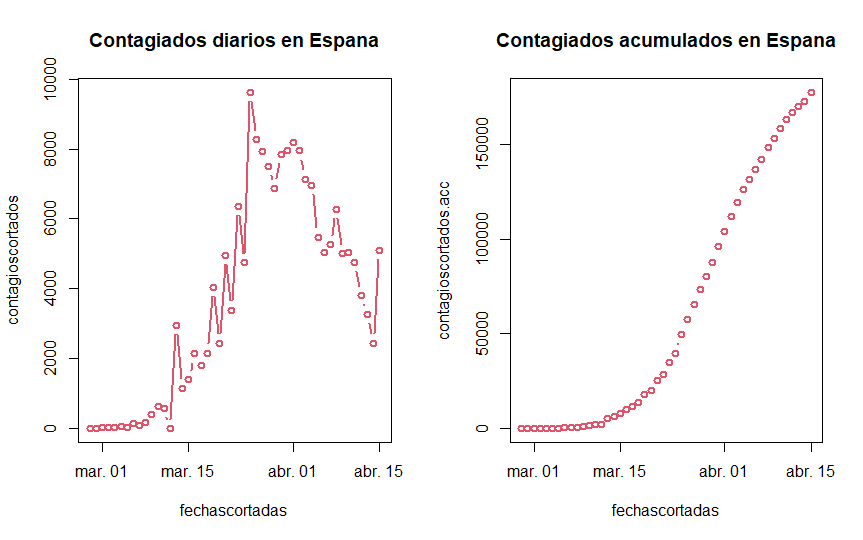
**f) ¿Cuántos datos hay en la tabla coronavirus referentes a España?**

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

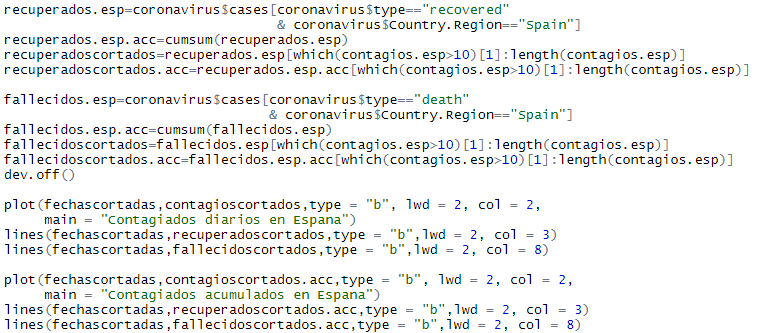
**3. Los gráficos tienen que ser informativos y en ambos casos hay una gran superficie en blanco (y, por tanto, no informativa). Ajusta las figuras para que resulten más informativas: puedes pintar los datos desde una fecha concreta o cambiar la escala de los gráficos.**

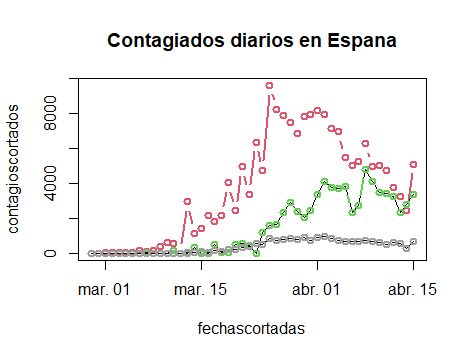
Recortaremos los datos para que solo se tengan en cuenta las fechas a partir del momento en el que hubo más de 10 contagios, el 28 de febrero.

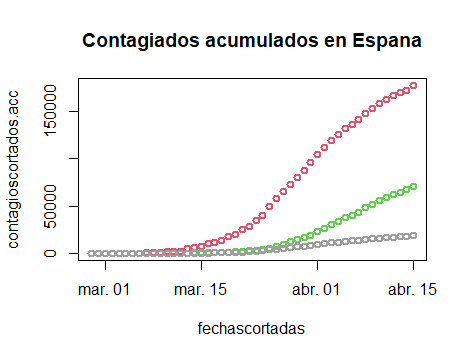


Dejando así ambas gráficas:

**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() .**





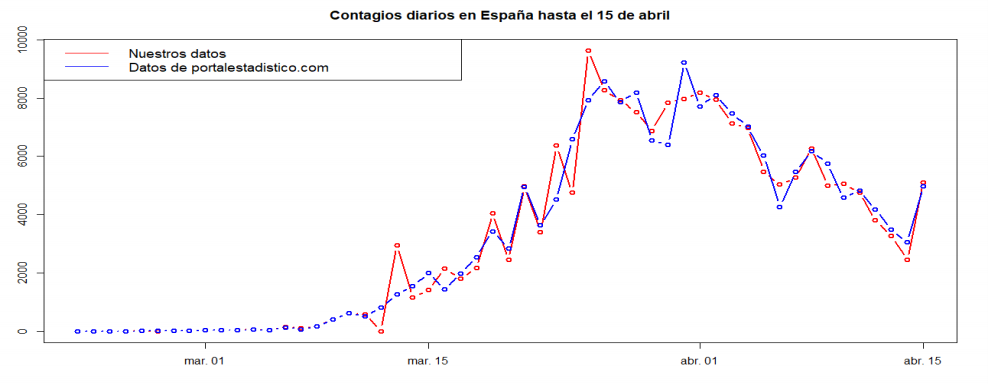


Representando en rojo los contagiados, en verde los recuperados y en gris los fallecidos.

**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?**

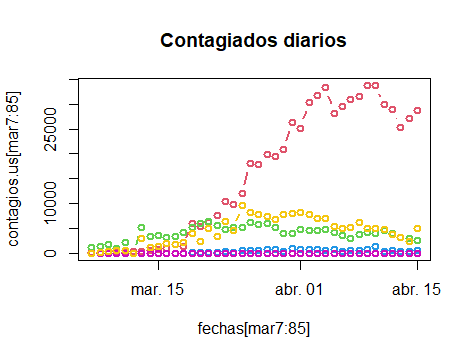
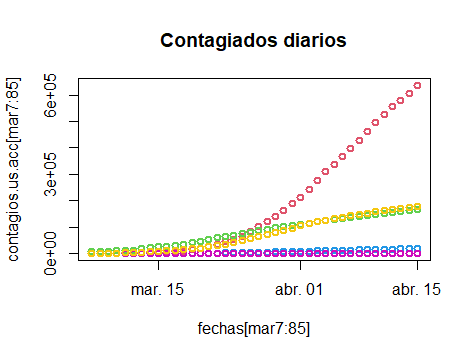
Es razonable que los contagios suban y posteriormente bajen. Las muertes representan aparentemente un porcentaje de los contagiados en cada fecha, y los picos de contagiados se manifiestan como picos de recuperados unos días después. Sin embargo, hay un gran pico al final como último dato que habría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

**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.

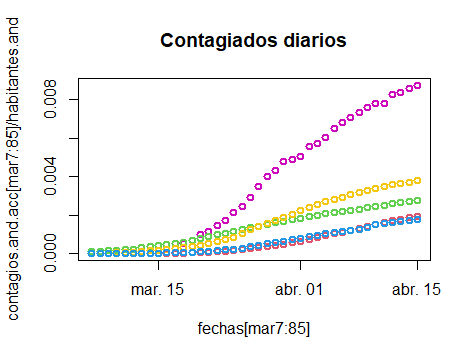
**7. También es interesante la comparativa entre países. Dibuja las curvas de contagiados diarios y acumulados para España, Italia (“Italy”), Estados Unidos (“US”), Andorra (“Andorra”) y otro país de tu elección a partir del 7 de marzo. Nota: Ten cuidado con la forma de pintar las gráficas para que no 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.

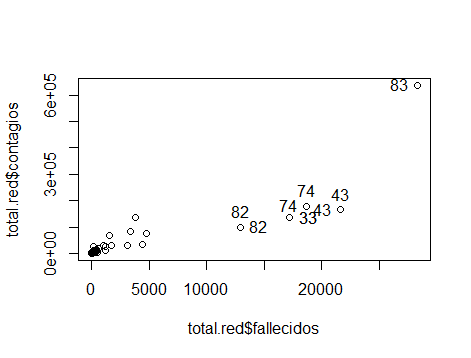
**9. Hasta ahora hemos considerado solo frecuencias absolutas y puede que esto no sea muy justo. Repite la gráfica de los contagios acumulados con los contagios/habitante. ¿Mantienes las mismas conclusiones? ¿Se te ocurre alguna explicación para el comportamiento de los datos de Andorra?**

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 Km2 ) en comparación con por ejemplo la de Estados unidos (33 habitantes por Km2 ), pero habría que explorar con más detalle el porqué de este resultado.

**Ejercicio 2**

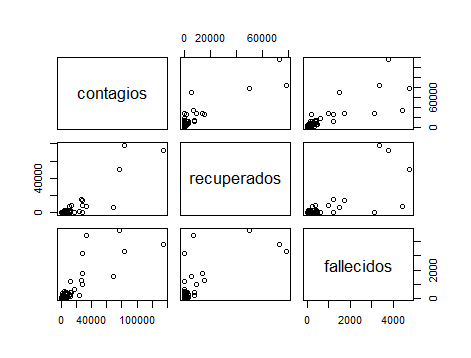
**1. Dibujar el diagrama de dispersión de los fallecidos frente a los contagiados. ¿Hay algún valor atípico? Si es así identifícalo.**



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

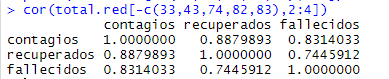
**2. Eliminar los posibles atípicos y dibujar el diagrama de pares (o matriz de diagramas de dispersión) con todas las variables numéricas. Indicar si hay relaciones y de qué tipo entre las variables, o si se produce algún efecto extraño.**





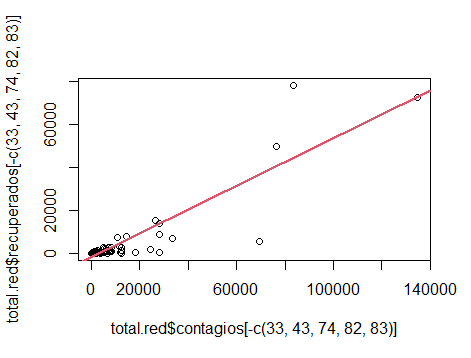
Se puede intuir en todos los casos una relación lineal, pero cuando los valores se vuelven grandes, aparece una gran variabilidad y los datos se aparecen en forma de cono.

**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.**



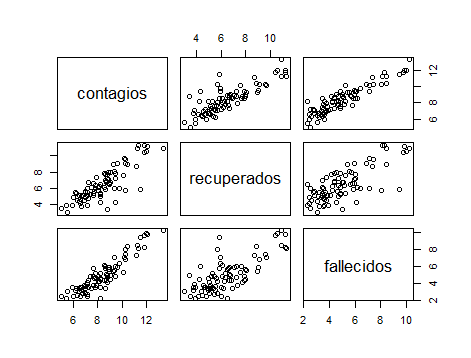
Podemos observar que hay una correlación positiva, tal y como podíamos intuir por la pendiente aparente de los diagramas anteriores, pero no se ajusta totalmente a un modelo lineal. El que más puede aproximarse, tal y como podemos ver en las covarianzas, es el de contagios frente a recuperados:

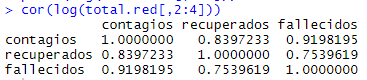




Podemos ver que la recta se ajusta a los datos, pero sigue habiendo datos muy alejados de ella, especialmente a medida que nos alejamos del origen de coordenadas.

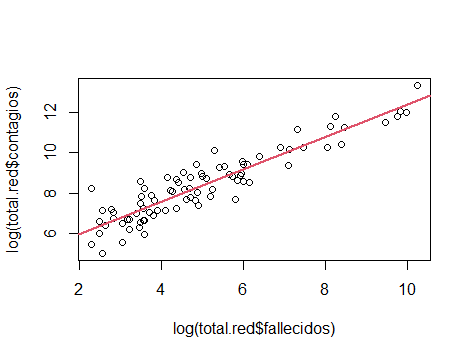
**4. Posiblemente te hayan aparecido nubes de puntos “en forma de cono” con muy poca variabilidad al inicio y mucha al final. Esto es efecto de la heterocedasticidad (las varianzas no son iguales en los distintos puntos). Para eliminar este efecto y “linealizar” los datos se puede aplicar una transformación 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?**





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 million?**

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:



Obtenemos que habrá aproximadamente 22.200 fallecidos.

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



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