SOLUCIONES LAB1: Introducción a los datos

Objetivos

Se introducen las herramientas básicas para trabajar con conjuntos de datos y para producir resúmenes básicos, tanto numéricos como gráficos, reforzando los conceptos vistos en clase. Es especialmente importante la interpretación de los resultados obtenidos.

Los objetivos específicos de esta sesión son,

- Manejar conjuntos de datos en R
- Resumir los datos numéricamente
- Resumir los datos gráficamente

Se van a manejar dos conjuntos de datos distintos y estructuramos la sesión teniendo en cuenta cada uno de ellos.

Dataset1: cdc.dat

En Estados Unidos se lleva a cabo una encuesta telefónica anual conocida como *the Behavioral Risk Factor Surveillance System* (*BRFSS*) cuyo objetivo es identificar factores de riesgo para la salud de población adulta y describir nuevas tendencias en este campo. El cuestionario contiene preguntas acerca de la dieta, ejercicio, hábitos poco saludables como el tabaco, así como de la cobertura sanitaria. El *dataset* cdc.dat contiene las respuestas de 20,000 adultos americanos en la encuesta realizada en el año 2000 con la siguiente información:

Variable	Descripción
case	Identificador de individuo, número entero del 1 al 20,000
genhlth	Estado de salud general. Puede tomar los valores `excellent`, `very good`, `good`, `fair`, `poor`
exerany	Hizo ejercicio en el último mes, con valores 0 para No y 1 para Sí
hlthplan	Cobertura sanitaria, con valores 0 para No y 1 para Sí
smoke100	Ha fumado más de 100 cigarrillos en su vida, con valores 0 para No y 1 para Sí
height	Altura en pulgadas
weight	Peso en libras
wtdesire	Peso que el encuestado quisiera tener (en libras)
age	Edad en años
gender	Sexo del individuo, que puede tomar los valores `m` para varones (*male*) o `f` para mujeres (*female*)

1. Cargar el conjunto de datos. Crear un *data frame* llamado cdc.

```
# establezco el directorio de trabajo
```

```
workingDir <- "D:/Ingenieria Biomedica/Curso 2019-
2020/Material/Laboratorios"
setwd(workingDir)
# cargamos los datos
cdc <- read.table(file.path(workingDir,"datos/cdc.dat"),
header=TRUE, sep = "\t")
```

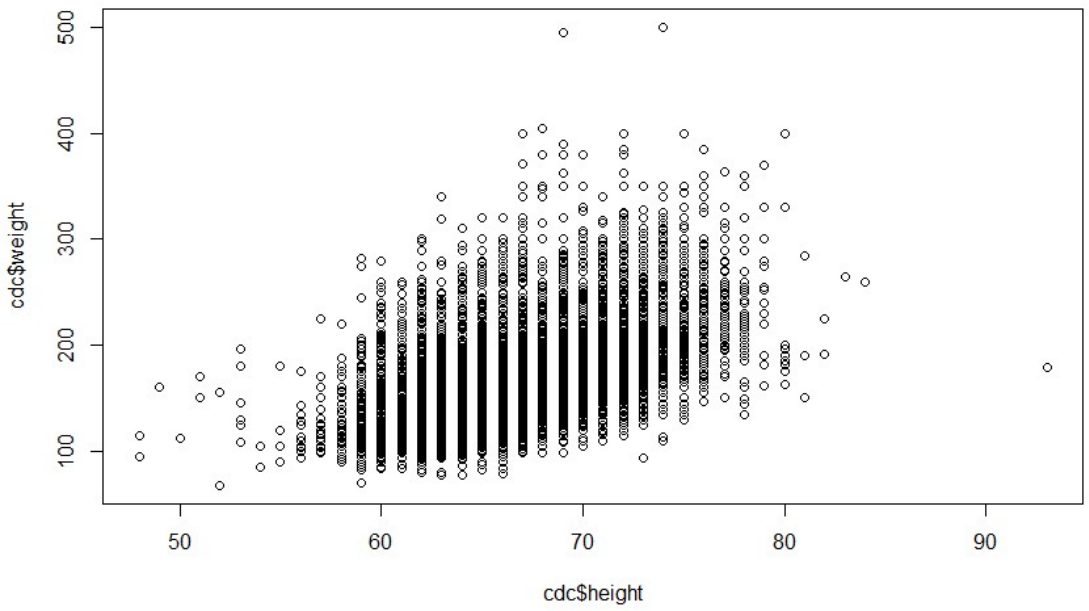
2. Podemos utilizar el comando `View(cdc)` para echar un vistazo a los datos. Aparecerán en una nueva pestaña, en el panel superior izquierdo, en forma de tabla. En la pestaña *Environment* del panel superior derecho, al hacer clic en la flecha azul junto al nombre del conjunto de datos puede verse un resumen de las 9 variables, incluido el tipo de dato, contenidas en el *data frame*.

   Cada fila de este *data frame* representa un individuo y cada columna una variable. Cada variable es la respuesta a un ítem del cuestionario. Cada variable es un vector de enteros, salvo el estado de salud y el sexo que son factores.

3. El operador `$` en R se utiliza para acceder a las columnas de un *data frame* utilizando su nombre, por ejemplo con el comando `cdc$height` hacemos referencia a la variable `height` del *data frame* `cdc`. Construir el diagrama de dispersión para la altura y el peso utilizando la función `plot()` con el comando,

   ```
   plot(cdc$weight ~ cdc$height)
   ```

   ¿Dirías que estas dos variables están asociadas?



La tendencia en la nube de puntos muestra que la altura y el peso están asociados positivamente, es decir que el peso tiende a aumentar a medida que aumenta la altura. Podemos calcular el coeficiente de correlación lineal con el comando,

```
cor(cdc$weight , cdc$height)
```

```
[1] 0.5553222
```

obtenemos un coeficiente de Pearson de 0.555.

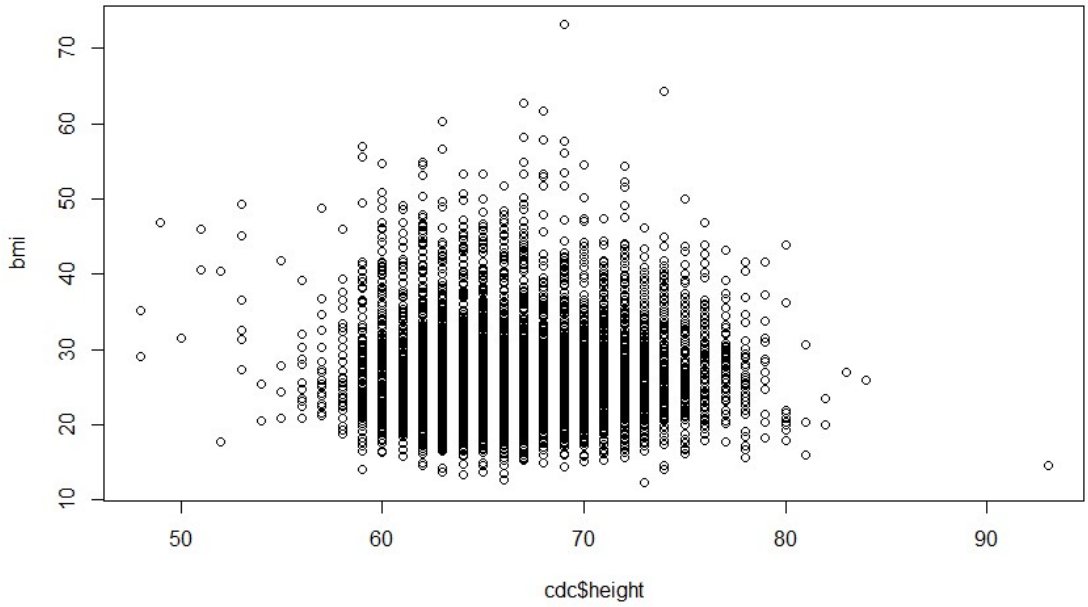
4. Una pulgada equivale a 0.0254 metros. Crea una nueva variable `height.m` que recoja el peso en metros. Análogamente, sabiendo que 1 libra equivale a 0.454 kilogramos, crea la variable `weight.kg` con el peso en kilogramos. Una vez creadas estas nuevas variables añadirlas al data frame cdc con el comando,

```
cdc<-data.frame(cdc, height.m = height.m, weight.kg = weight.kg)
# crear height.m
height.m <- cdc$height * 0.0254
# crear weight.kg
weight.kg <- cdc$weight * 0.454
# podemos almacenarlas en el data frame cdc
cdc<-data.frame(cdc, height.m = height.m, weight.kg = weight.kg)
head(cdc)
```

	case	genhlth	exerany	hlthplan	smoke100	height	weight	wtdesire	age	gender	height.m	weight.kg
1	1	good	0	1	0	70	175	175	77	m	1.7780	79.450
2	2	good	0	1	1	64	125	115	33	f	1.6256	56.750
3	3	good	1	1	1	60	105	105	49	f	1.5240	47.670
4	4	good	1	1	0	66	132	124	42	f	1.6764	59.928
5	5	very good	0	1	0	61	150	130	55	f	1.5494	68.100
6	6	very good	1	1	0	64	114	114	55	f	1.6256	51.756

5. El índice de masa corporal (BMI) se calcula como el peso en kilogramos dividido entre la altura (en metros) al cuadrado. Crea una nueva variable `bmi` que almacene el índice de masa corporal de cada encuestado. Construye un diagrama de dispersión de la altura (en pulgadas) y este índice, ¿te parece que existe asociación entre estas dos variables?

```
# crear la variable bmi
bmi <- (weight.kg)/(height.m^2)
# diagrama de dispersión: bmi vs altura
plot(cdc$height, bmi)
```



Estas dos variables no parecen estar relacionadas, es el plot nulo.

Un índice de masa corporal mayor o igual a 30 es considerado sobrepeso. ¿Por qué crees que es más apropiado elegir este índice como indicador de sobrepeso que simplemente el peso?

El índice de masa corporal es más útil puesto que no depende de la altura. Puesto que es lógico pensar que a mayor altura mayor peso, si utilizamos directamente el peso no estamos teniendo en cuenta que aquellos individuos más altos pesarán más no porque tengan sobrepeso, sino porque simplemente son más altos.

6. Los corchetes pueden utilizarse para acceder a un subconjunto de los datos. Por ejemplo, para mostrar la variable `weight`, que ocupa la columna 7, del encuestado 567 podemos utilizar,

```
cdc[567, 7]
```

```
[1] 160
```

Para ver el peso de los 10 primeros encuestados,

```
cdc[1:10, 7]
```

```
[1] 175 125 105 132 150 114 194 170 150 180
```

Si no especificamos el índice de la columna se mostrarán las filas completas de los 10 primeros encuestados,

```
cdc[1:10, ]
```

```
   case   genhlth exerany hlthplan smoke100 height weight wtdesire age gender height.m weight.kg
1     1      good       0        1        0     70    175      175  77      m   1.7780    79.450
2     2      good       0        1        1     64    125      115  33      f   1.6256    56.750
3     3      good       1        1        1     60    105      105  49      f   1.5240    47.670
4     4      good       1        1        0     66    132      124  42      f   1.6764    59.928
5     5 very good       0        1        0     61    150      130  55      f   1.5494    68.100
```

6	6	very good	1	1	0	64	114	114	55	f	1.6256	51.756
7	7	very good	1	1	0	71	194	185	31	m	1.8034	88.076
8	8	very good	0	1	0	67	170	160	45	m	1.7018	77.180
9	9	good	0	1	1	65	150	130	27	f	1.6510	68.100
10	10	good	1	1	0	70	180	170	44	m	1.7780	81.720

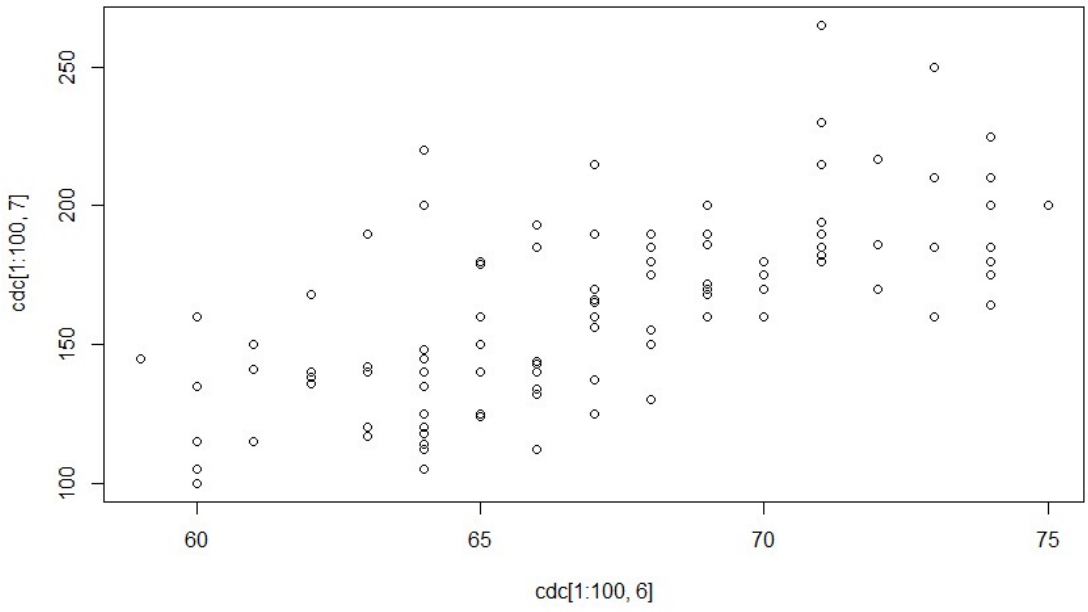
Si no especificamos el índice de la fila, se accederá al peso de los 20,000 encuestados.

```
cdc[ ,7]
```

7. Utilizando corchetes construir un diagrama de dispersión del peso y la altura para los 100 primeros encuestados.

   Hay varias formas de hacerlo:

```
# 1. Crear un nuevo data frame sólo con las 100 primeras filas
cdc.100 <- cdc[1:100, ]
plot(cdc.100$height, cdc.100$weight)
# 2. Crear las dos variables
cdc.100.weight <- cdc[1:100, 7]
cdc.100.height <- cdc[1:100, 6]
plot(cdc.100.height, cdc.100.weight)
# 3. No crear ninguna estructura nueva
plot(cdc[1:100, 6], cdc[1:100, 7])
```



8. El comando `summary(x)` da como resultado un resumen numérico de la variable x. Este resumen numérico está formado por 6 medidas de posición relevantes: mínimo y máximo los 3 quartiles y la media. Da un resumen de la variable `weight.kg`. ¿Te llama la atención el valor de alguna de las 6 medidas obtenidas? ¿Podrías deducir algo acerca de la forma de la distribución de la variable `weight.kg`?

```
# calcular el resumen numérico. 5 medidas: min, Q1, Q2, Q3 y max
summary(cdc$weight.kg)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.87   63.56   74.91   77.04   86.26  227.00
```

El valor del máximo es bastante mayor que las otras posiciones, lo cual sugiere que la distribución de los datos será asimétrica a la derecha. Esto también lo podemos observar comparando la media y la mediana: la media es mayor que la mediana 77.04>74.91, lo que indica asimetría a la derecha. Podríamos haber obtenido la media con la función `mean()`,

```
mean(cdc$weight.kg)
```

```
[1] 77.03606
```

La mediana con la función `median()`,

```
median(cdc$weight.kg)
```

```
[1] 74.91
```

Y los cuartiles con `quantile()`,

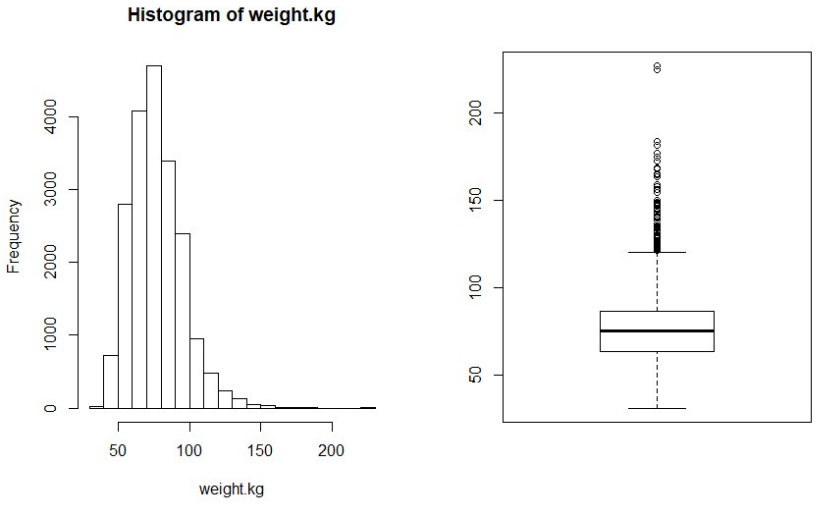
```
quantile(cdc$weight.kg, probs = c(0.25,0.5,0.75))
```

```
  25%   50%   75%
63.56 74.91 86.26
```

9. Construye un histograma y un diagrama de cajas de la distribución de la variable `weight.kg`. ¿Qué puedes decir acerca de la forma de la distribución?

   En R los gráficos se muestran en la ventana de gráficos. Podemos mostrar varios gráficos en la misma ventana utilizando el comando `par(mfrow = c(nrow, ncol))` que dividirá la ventana de gráficos en `nrow` filas y `ncol` columnas.

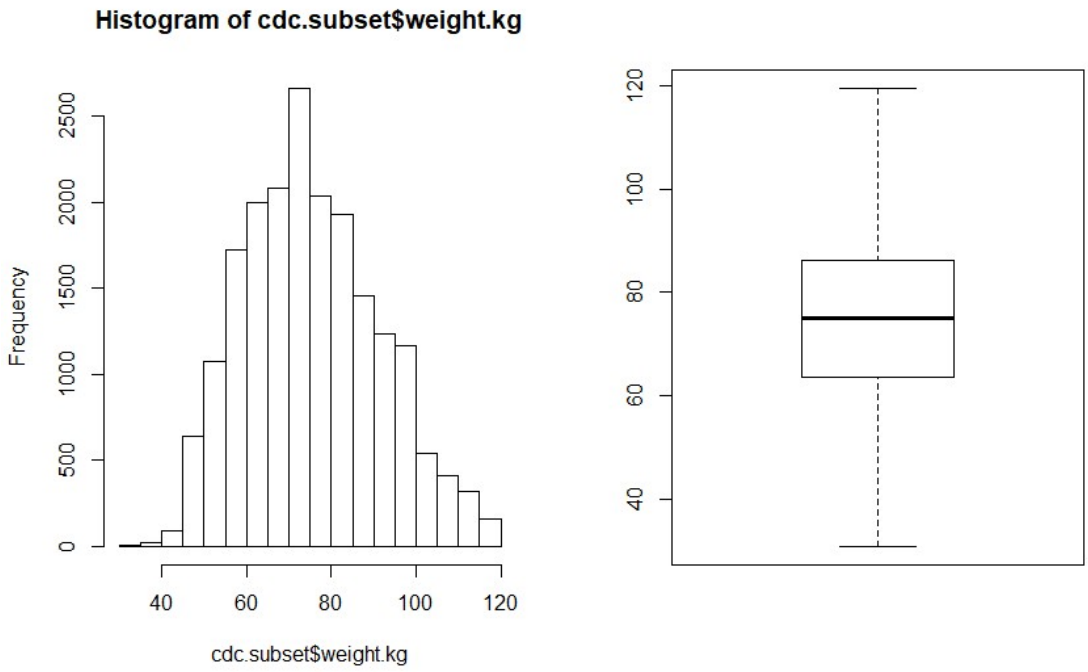
```
# mostramos los dos gráficos a la vez
par(mfrow = c(1,2)) # 1 fila y 2 columnas
# Histograma
hist(cdc$weight.kg)
# Diagrama de cajas
boxplot(cdc$weight.kg)
```



Estos gráficos confirman la asimetría a la derecha. Vemos además que hay 2 individuos con pesos bastante extremos que podrían considerarse outliers.

10. Para una variable x, el comando x<a se utiliza para seleccionar el conjunto de observaciones de x que son menores que a. Utiliza la función `subset()` para construir un nuevo *data frame*, `cdc.subset`, que solo contenga los individuos que pesen menos de 120 kg. Utilizando este nuevo conjunto de datos, rehacer el histograma y en diagrama de cajas del punto anterior, y calcular de nuevo el resumen numérico para la variable `weight.kg`. ¿Cómo es ahora la forma de la distribución de esta variable?

```
# crear el nuevo conjunto de datos
cdc.subset <- subset(cdc, cdc$ weight.kg < 120)
# crear el histograma y el diagrama de cajas
par(mfrow = c(1,2))
hist(cdc.subset$weight.kg)
boxplot(cdc.subset$weight.kg)
```




```
# el resumen numérico
summary(cdc.subset$weight.kg)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.87   63.56   74.91   75.65   86.26  119.40
```

La apariencia tanto del histograma como del diagrama de cajas podría corresponder a una distribución simétrica. Cuando comparamos la media y la mediana, observamos que aunque la media está ligeramente desplazada hacia la derecha, apenas hay diferencia entre ellas, menos de 1 kg.

11. El comando `cdc$weight.kg < 120` devuelve un vector de TRUE/FALSE de la misma longitud que `cdc$weight.kg`. Cuando sumamos las componentes de este vector obtenemos el número de individuos que cumplen esa condición.

```
sum(cdc$weight.kg < 120)
```

```
[1] 19523
```

Utiliza la función `nrow()` para determinar cuántos de entre los 20000 encuestados pesan más de 120 kg.

```
# directamente podemos restar el n de filas de cdc del de cdc.subset s

nrow(cdc) - nrow(cdc.subset)
```

```
[1] 477
```

Efectivamente son los 20000 encuestados menos los 19523 que pesan menos de 120 kg.

Podríamos obtener lo mismo con los comandos,

```
sum(!(cdc$weight.kg < 120)) # los que NO (!) cumplen la condición
sum(cdc$weight.kg >= 120)   # los que cumplen la condición >=
```

12. Utiliza el siguiente código para extraer una muestra aleatoria del 10% de los datos originales. La muestra se obtiene sin reemplazamiento, es decir que cada encuestado no puede elegirse más de una vez. Utilizamos el comando `set.seed()` para establecer una semilla del generador de números aleatorios y así poder obtener la misma muestra siempre que queramos, esto es, que la extracción de la muestra sea reproducible. El comando `round(a,n)` redondea el valor a a n decimales.

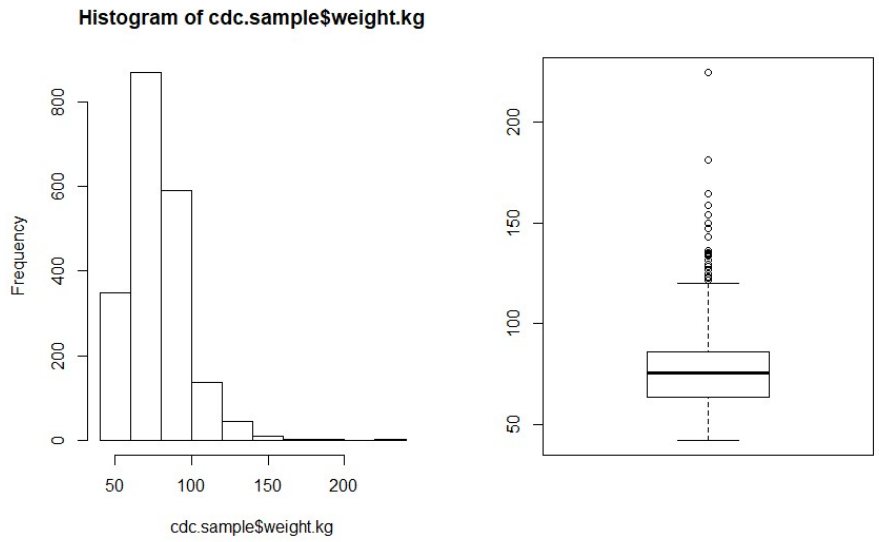
```
set.seed(1050)
sample.n <- round(0.1 * nrow(cdc), 0)
sample.seq <- sample(1 : nrow(cdc), size = sample.n, replace = FALSE)
cdc.sample <- cdc [sample.seq, ]
```

Con este nuevo conjunto de datos, obtener el resumen numérico de la variable `weight.kg`, así como su histograma y el diagrama de cajas. ¿La distribución de esta variable es similar a la obtenida con todos los datos o a la que se obtiene al eliminar los más extremos?

```
# resumen numérico
summary(cdc.sample$weight.kg)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  42.22   63.56   75.59   77.45   86.26  224.73
```

```
# histograma y diagrama de cajas
par(mfrow = c(1,2))
hist(cdc.sample$weight.kg)
boxplot(cdc.sample$weight.kg)
```



La distribución es más parecida a la obtenida con todas las observaciones, se observa asimetría a la derecha, cosa que era esperable, puesto que se trata de una muestra aleatoria.

13. En este conjunto de datos la variable `gender` es un *factor*, que puede tomar dos valores 'm' o 'f' para hombres y mujeres respectivamente. Utiliza la función `summary()` con esta variable. ¿Qué resumen numérico obtienes?

```
summary(cdc$gender)
```

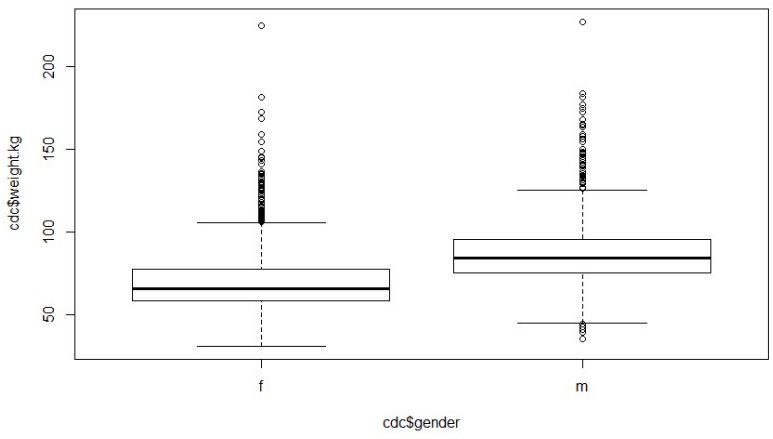
```
    f     m
10431  9569
```

Se obtiene la tabla de frecuencias ya que un factor almacena información de una variable cualitativa.

14. Construir el diagrama de cajas múltiples para la variable `weight.kg` agrupada según el sexo del encuestado. Utiliza el comando,

```
boxplot(cdc$weight.kg ~ cdc$gender)
```

¿Dirías que hay relación entre el peso del encuestado y su género?



Se observa que el peso en el grupo de varones (m) es mayor que en el de las mujeres (f), lo que sugiere que entras dos variables hay relación.

15. Con el comando `table(x,y)` obtenemos la tabla de contingencia en la que la variable `x` se coloca por filas y la variable `y` por columnas. Si omitimos la variable `y`, obtenemos la tabla de frecuencias de `x`. Utiliza esta función para construir la tabla de frecuencias de la variable estado de salud, así como la tabla de contingencia para dicha variable y el hábito de fumar (columna). Utiliza el comando `addmargins()` para añadir a esta tabla (pasarla como argumento) las marginales de fila y columna.

```
# tabla de frecuencias del estado de salud
table(cdc$genhlth)
```

```
excellent      fair      good      poor very good
     4657      2019      5675       677      6972
```

```
# tabla de contingencia
table(cdc$genhlth, cdc$smoke100)
```

```
               0    1
  excellent 2879 1778
  fair       911 1108
  good      2782 2893
  poor       229  448
  very good 3758 3214
```

```
# añadiendo las marginales
addmargins(table(cdc$ genhlth, cdc$smoke100))
```

```
                0     1   Sum
  excellent  2879  1778  4657
  fair        911  1108  2019
  good       2782  2893  5675
  poor        229   448   677
  very good  3758  3214  6972
  Sum       10559  9441 20000
```

Dataset2: nhanes.dat

La encuesta *The National Health and Nutrition Examination Survey* (NHANES) es una encuesta realizada anualmente por el *US National Center for Health Statistics* (NCHS), cuyo objetivo es evaluar el estado de salud y la nutrición de población americana, tanto adulta como infantil. En la siguiente sección utilizamos el conjunto de datos `nhnes.dat`, que contiene la información de 10000 estadounidenses. Las variables a utilizar en este ejercicio, junto con su descripción se resume en la siguiente tabla,

Variable	Descripción

ID	Identificador de individuo.
Age	Edad en años. La edad de los individuos con 80 años o más se registra como 80.
Education	Nivel educativo más alto del participante mayor de 20 años. Los valores que puede tomar son `8th Grade`, `9 - 11th Grade`, `High School`, `Some College` o `College Grade.`
Poverty	Ratio de pobreza. Se calcula como el cociente de los ingresos familiares y el umbral de ingresos para que una familia sea considerada "pobre". Valores < 1 indican ingresos por debajo del nivel de pobreza, y >1 por encima de este umbral.
Weight	Peso en kilogramos.
Height	Altura en centímetros.
Diabetes	`Yes`, si el participante tiene diabetes diagnosticada y `No` en otro caso.
PhysActive	`Yes`, si el participante practica deporte de intensidad moderada/alta y `No` en otro caso. Se recoge para mayores de 12 años.

Cuestión 1

a) Describe la distribución de la edad para los participantes en este estudio. Da resúmenes numéricos y gráficos y trata de resumir las principales características de la distribución con palabras.

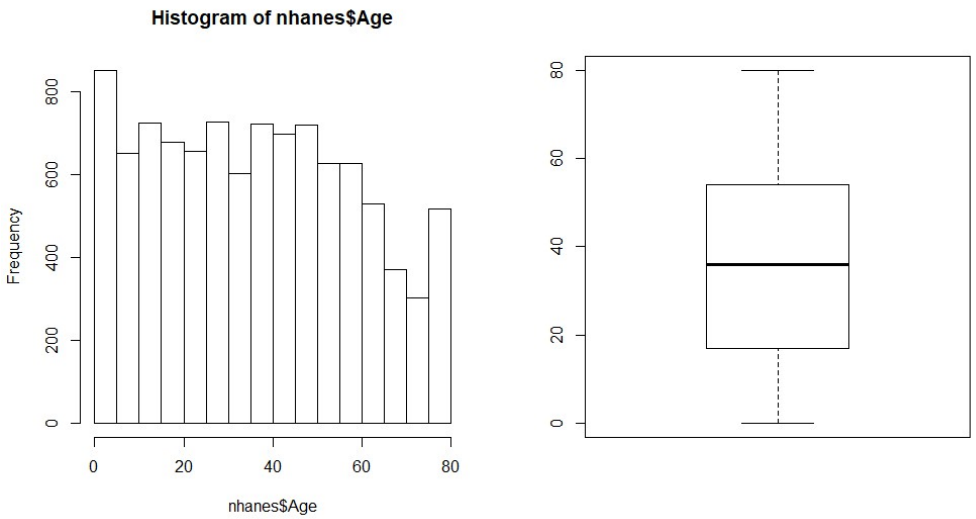
```
# Lo primero es leer los datos
# establezco el directorio de trabajo
workingDir <- "D:/Ingenieria Biomedica/Curso 2019-
2020/Material/Laboratorios"
setwd(workingDir)
# cargamos los datos y creamos un data frame
nhanes <- read.table(file.path(workingDir,"datos/nhanes.dat"),
header=TRUE, sep = "\t")
# resumen numérico
summary(nhanes$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   17.00   36.00   36.74   54.00   80.00
```

```
# no tenemos una medida de dispersión, calculamos la desviación típica
sd(nhanes$Age)
```

```
[1] 22.39757
```

```
# gráficos: histograma y diagrama de cajas
par(mfrow = c(1,2))
hist(nhanes$Age)
boxplot(nhanes$Age)
```



Observamos que la distribución de la edad es bastante simétrica. El 50% de los participantes son menores de 50 años y el 50% central de estas edades se encuentra entre 37 y 63 años. Aunque el valor máximo de la edad es 80, hay que tener en cuenta que no se han registrado valores mayores que ese valor, lo que significa que podría haber algún sujeto mayor de 80 años.

b) Describe ahora la distribución de la altura en pulgadas. Tened en cuenta que 1 centímetro es aproximadamente 0.39 pulgadas.

```
# creamos la nueva variable
altura.pul <- 0.39*nhanes$Height
# resumen numérico
summary(altura.pul)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  32.60   61.15   64.74   63.13   68.06   78.16     353
```

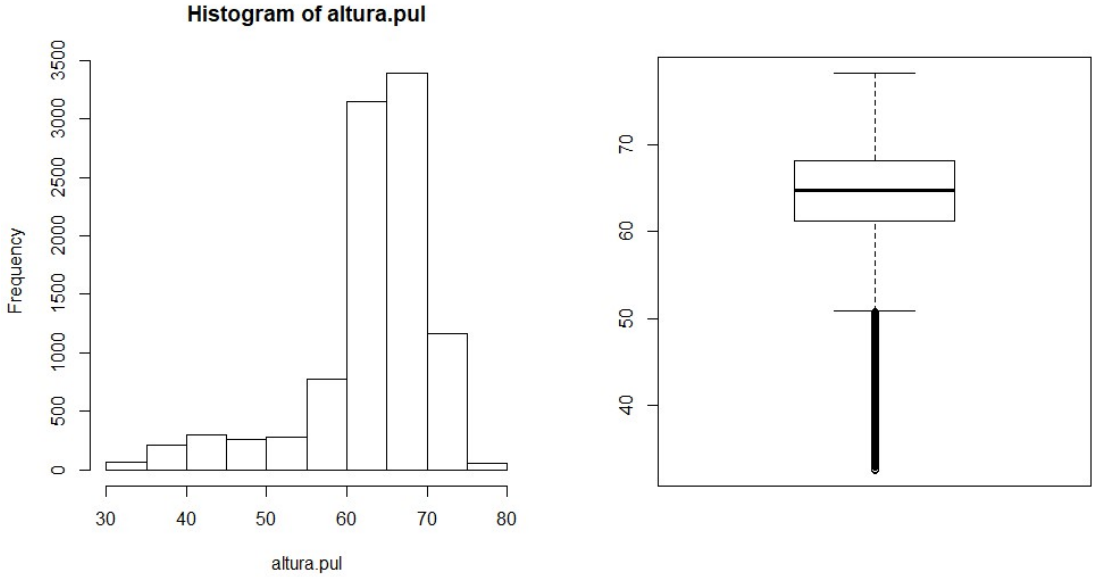
```
# no tenemos una medida de dispersión, calculamos la desviación típica
sd(altura.pul)
```

```
[1] NA
```

```
# como la variable tiene NA's necesitamos el argumento na.rm=TRUE
sd(altura.pul,na.rm=TRUE)
```

```
[1] 7.872761
```

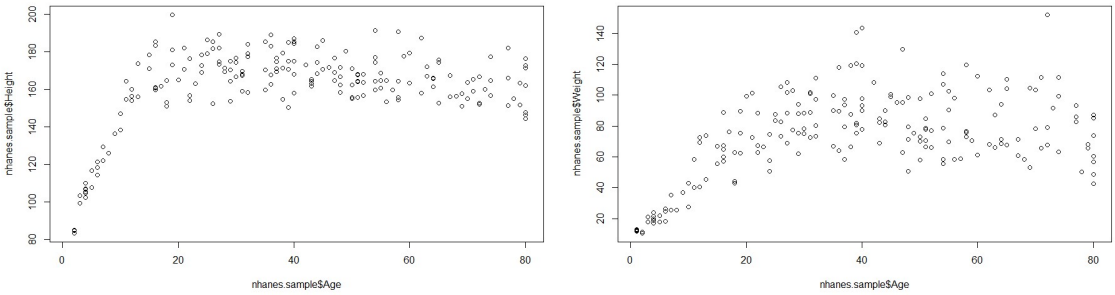
```
# gráficos: histograma y diagrama de cajas
par(mfrow = c(1,2))
hist(altura.pul)
boxplot(altura.pul)
```



La distribución de la altura en este conjunto de datos es bastante asimétrica a la izquierda, hay muchos más participantes altos que bajos. La mediana es de 64.74 pulgadas y la media de 63.13. Gráficamente se ve muy claro, tanto en el histograma como en el diagrama de cajas, donde observamos bastantes observaciones representadas por puntos para valores bajos de la altura. A diferencia de los que ocurría con la edad, en este caso hay 353 individuos missing's, es decir para los que no se ha conseguido medir la altura.

c) Estableciendo la semilla del generador de números aleatorios en 5011, extrae una muestra aleatoria sin reemplazamiento de tamaño 200. En este nuevo conjunto de datos, `nhanes.sample`, ¿a qué edad, aproximadamente, un adulto alcanza su altura final? Utiliza un diagrama de dispersión para establecer aproximadamente este valor. ¿Es posible determinar a qué edad se alcanzará el peso final? ¿Por qué?

```
# extraemos la muestra aleatoria
set.seed(5010)
nhanes.sample <- nhanes [sample(1 : nrow(nhanes), size = 200, replace
= FALSE), ]
# hacemos el diagrama altura vs edad y peso vs edad
par(mfrow = c(1,2))
plot(nhanes.sample$Age, nhanes.sample$Height)
plot(nhanes.sample$Age, nhanes.sample$Weight)
```



En el diagrama de dispersión Altura vs Edad observamos que, hasta los 20 años, la altura aumenta muy rápidamente, para posteriormente estabilizarse. Sin embargo, para el peso no está tan claro, ya que el peso varía más en la edad adulta.

Cuestión 2

El nivel educativo es una variable cualitativa ordinal, podemos ordenar los niveles de esta variable utilizando el comando,

```
Education.ord <- factor(nhanes$Education, order = TRUE, levels = c("8th Grade", "9 - 11th Grade", "High School", "Some College", "College Grad"))
```

a) Utilizando esta nueva variable, ¿qué proporción de americanos con al menos 25 años son graduados universitarios (*college graduates*)?

```
# seleccionar los individuos de al menos 25 años
Education.ord.25 <- Education.ord[nhanes$Age >= 25]
# tabla de frecuencias absolutas del nivel educativo
table(Education.ord.25) # o summary(Education.ord.25)
```

```
Education.ord.25
     8th Grade 9 - 11th Grade    High School   Some College   College Grad
           435            814           1345           1951           2016
```

```
# para calcular el porcentaje necesitamos el tamaño muestral (n)
total.25 <- length(Education.ord.25) # o sum(nhanes$Age >= 25)
# cálculo
2016/total.25
```

```
[1] 0.3068026
```

```
# Podríamos haber obtenido la tabla de frecuencias relativas completa
table(Education.ord.25)/total.25
```

```
Education.ord.25
     8th Grade 9 - 11th Grade    High School   Some College   College Grad
    0.06619997     0.12387764     0.20468726     0.29691067     0.30680262
```

Un 30.7% de americanos con al menos 25 años son graduados universitarios.

b) ¿Qué proporción de americanos con al menos 25 años no tienen una titulación universitaria (nivel educativo < *Some College*)? *Pista*: utiliza la función `cumsum()` para obtener frecuencias acumuladas.

```
# tabla de frecuencias absolutas acumuladas del nivel educativo
cumsum(table(Education.ord.25))
```

```
     8th Grade 9 - 11th Grade    High School   Some College   College Grad
           435           1249           2594           4545           6561
```

```
# cálculo
2594/total.25
```

```
[1] 0.3947649
```

```
# tabla de frecuencias relativas acumuladas completa
cumsum(table(Education.ord.25))/total.25
```

```
     8th Grade 9 - 11th Grade    High School   Some College   College Grad
    0.06619997     0.19007761     0.39476488     0.69167554     0.99847816
```

Un 39.5% de americanos con al menos 25 años no tienen titulación universitaria.

c) ¿Qué proporción de americanos con al menos 25 años y sin titulación universitaria tienen educación secundaria?

```
# Ahora el total no son todos los mayores de 25 años
# hay que eliminar a los que tienen titulación universitaria
total.25.noUn <- sum(nhanes$Age >= 25 & (nhanes$Education!= "Some
College" & nhanes$Education!= "College Grad"))
total.25.noUn
```

```
[1] NA
```

```
# Hay NA's, hay que no tenerles en cuenta
total.25.noUn <- sum (nhanes$Age >= 25 & (nhanes$Education!= "Some
College" & nhanes$Education!= "College Grad"), na.rm=TRUE)
total.25.noUn
```

```
[1] 2594
```

```
# cálculo
1345/total.25.noUn
```

```
[1] 0.5185042
```

Un 51.9% de americanos con al menos 25 años y sin titulación universitaria tienen educación secundaria.

Cuestión 3

a) Calcula la mediana y el rango intercuartílico (función `IQR()`) de la distribución de la variable `Poverty`.

```
# es necesario utilizar el argumento na.rm = TRUE para que el
resultado no sea NA
median(nhanes$Poverty, na.rm = TRUE)
```

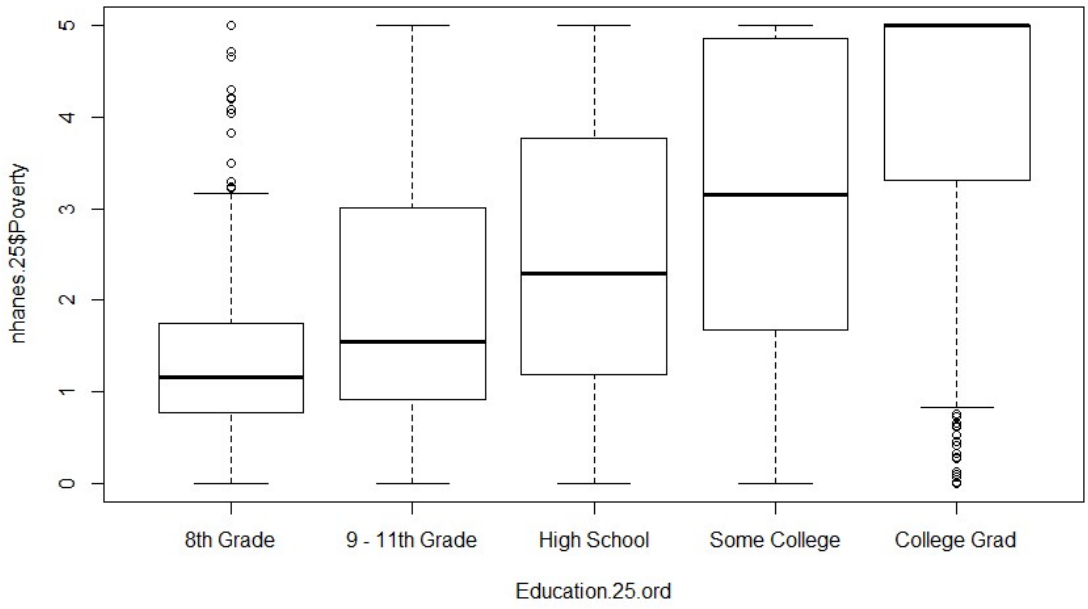
```
[1] 2.7
```

```
IQR(nhnes$Poverty, na.rm = TRUE)
[1] 3.47
```

La mediana es 2.7 y el rango intercuartílico 3.47. La mediana indica que el 50% de los participantes tienen un ratio de pobreza por encima de 2.7. Teniendo en cuenta lo que significa esta variable, esto indica que los encuestados tienen un poder adquisitivo 2.7 veces mayor que el umbral de pobreza establecido.

b) Compara gráficamente la distribución de `Poverty` para cada grupo de nivel educativo. Considera sólo los individuos cuya edad sea mayor o igual a 25 años. ¿Dirías que existe una relación entre estas dos variables?

```
# seleccionar los individuos de al menos 25 años
nhanes.25 <- nhanes[nhanes$Age >= 25, ]
# creamos el factor nivel educativo con sus valores ordenados
Education.25.ord<- factor(nhanes.25$Education, order = TRUE, levels =
c("8th Grade", "9 - 11th Grade", "High School", "Some College",
"College Grad"))
# diagrama de cajas múltiple
boxplot(nhanes.25$Poverty ~ Education.25.ord)
```



La mediana de la variable `poverty` incrementa a medida que aumenta el nivel de estudios. Mientras que participantes en el primer nivel (8th grade) tienen una mediana en `poverty` en torno a 1.1, los participantes con el máximo nivel (`college grad`) tienen un ratio en torno a 5. En el gráfico también vemos que hay individuos con `8th grade` que tienen un ratio muy

alto (outliers en el primer nivel) y algunos de los máximo nivel con ratios de pobreza muy cercanos a 0.

Cuestión 4

a) Construye una tabla de contingencia con la variable `PhysActive` por filas y la variable `Diabetes` por columnas. Entre los participantes que no practican actividad física, ¿cuál es la proporción que tienen diabetes? ¿Y entre los que sí practican ejercicio? *Pista*: las funciones `rowSums()` y `colSums()` calculan la suma por filas y por columnas en una matriz pasada como argumento, respectivamente.

```
# creamos la tabla
tbl.fis.diab <- table(PhysActive=nhanes$PhysActive, Diabetes=
nhanes$Diabetes)
# con marginales
addmargins(tbl.fis.diab)
```

```
          Diabetes
PhysActive   No  Yes  Sum
       No  3203  472 3675
       Yes 4361  285 4646
       Sum 7564  757 8321
```

```
# para contestar a las preguntas que nos hacen tenemos que construir
# la tabla condicionada por filas
tbl.cond.act <- tbl.fis.diab /rowSums(tbl.fis.diab)
tbl.cond.act
```

```
          Diabetes
PhysActive         No        Yes
       No  0.87156463 0.12843537
       Yes 0.93865691 0.06134309
```

Entre los participantes que no tienen actividad física prácticamente el 13% tienen diabetes, mientras que entre los que sí la tienen sólo hay un 6%

b) En este contexto, una medida de la asociación entre dos variables cualitativas que se puede calcular es el **riesgo relativo**, definido como el cociente de las proporciones de diabéticos en cada uno de los dos grupos de actividad. Se interpreta como cuantas veces es más frecuente la diabetes en el grupo que se pone en el numerador respecto al que se pone en el denominador. El valor de 1 indica que no hay relación, mientras que valores mayores que 1 indicarán que la condición del numerador es un riesgo de diabetes, y menores que 1 que protege de esta enfermedad. La no actividad física, ¿es un riesgo para la diabetes? ¿es su causa?

```
# calculamos el riesgo relativo
rr.diabetes <- tbl.cond.act[1,2] / tbl.cond.act [2,2]
rr.diabetes
```

```
[1] 2.093722
```

El riesgo relativo es 2.09, es decir que es 2 veces más frecuente encontrar diabetes entre los participantes que no tienen actividad física que entre los que sí que la tienen. Es importante tener en cuenta que, a pesar de ello, no podemos decir que la no actividad física sea la causa de la diabetes.