

Analítica de los Negocios

Explorando datos numéricos

Carlos Cardona Andrade

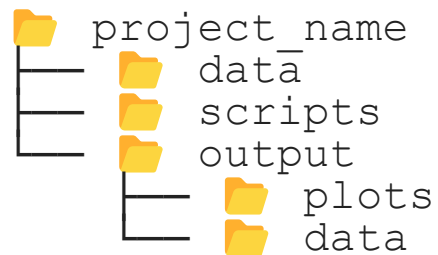
Plan para hoy

1. Estructura de un proyecto
2. Fundamentos sobre los datos
3. El Histograma
4. Medidas de tendencia central
5. Diagrama de caja
6. Medidas de dispersión
7. Medidas de relación entre dos variables

Estructura de un proyecto


Estructura de un proyecto

Un proyecto debería tener la siguiente estructura para facilitar la organización:




- Esta organización facilita el trabajo colaborativo y ayuda a retomar el proyecto más fácilmente
- Eviten usar **mayúsculas** o **espacios** en los nombres de archivos y carpetas. Ejemplos:
 - Incorrecto: `Archivo de datos.csv`
 - Correcto: `archivo_de_datos.csv`

El Directorio de Trabajo

- El directorio de trabajo es la carpeta en la que están trabajando actualmente
-  guarda y carga archivos desde esta carpeta por defecto
- Antes de comenzar, asegúrense de establecer tu directorio de trabajo con la siguiente función:

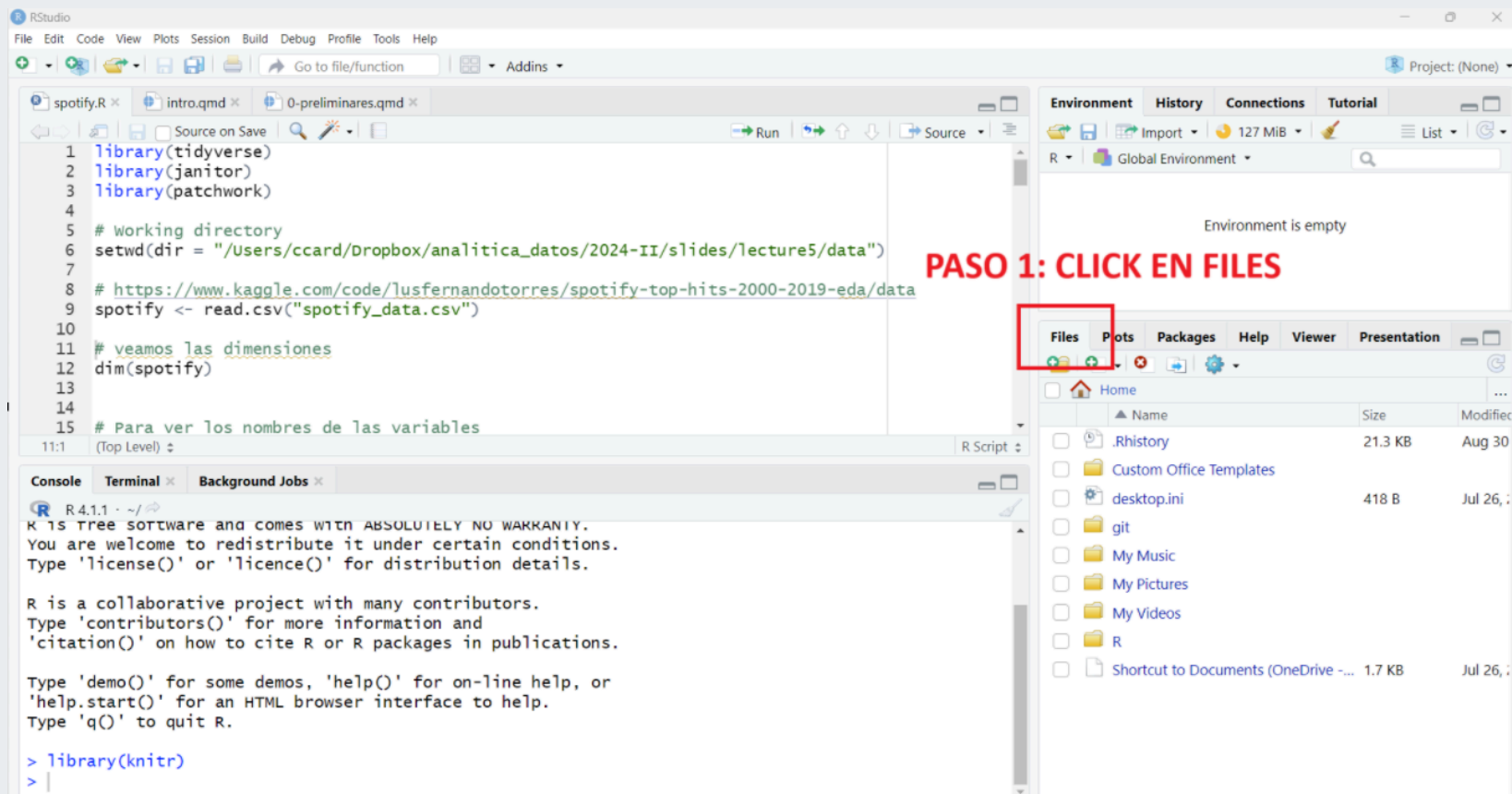
```
1 setwd("C:/Users/nombre_apellido/OneDrive/Documentos/analisis_de_datos/")
```

- Recuerden que en el explorador de archivos, las carpetas se dividen por "\", pero  requiere "/" como separador de carpetas

Establecer el Directorio de Trabajo Manualmente

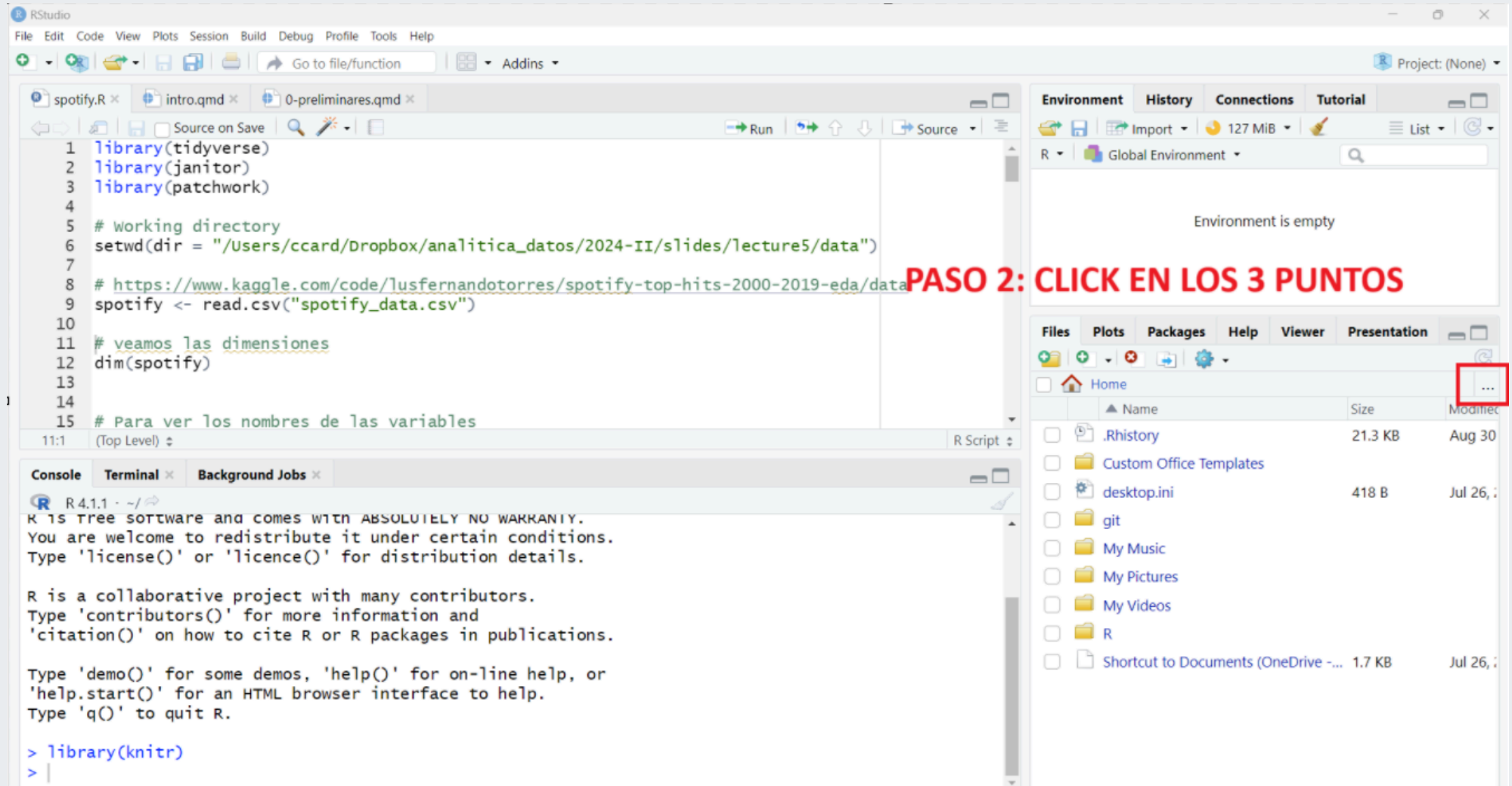
También puedes hacerlo manualmente desde el visualizador de archivos en R Studio:

Primero, van al panel [Files] en la parte inferior derecha.



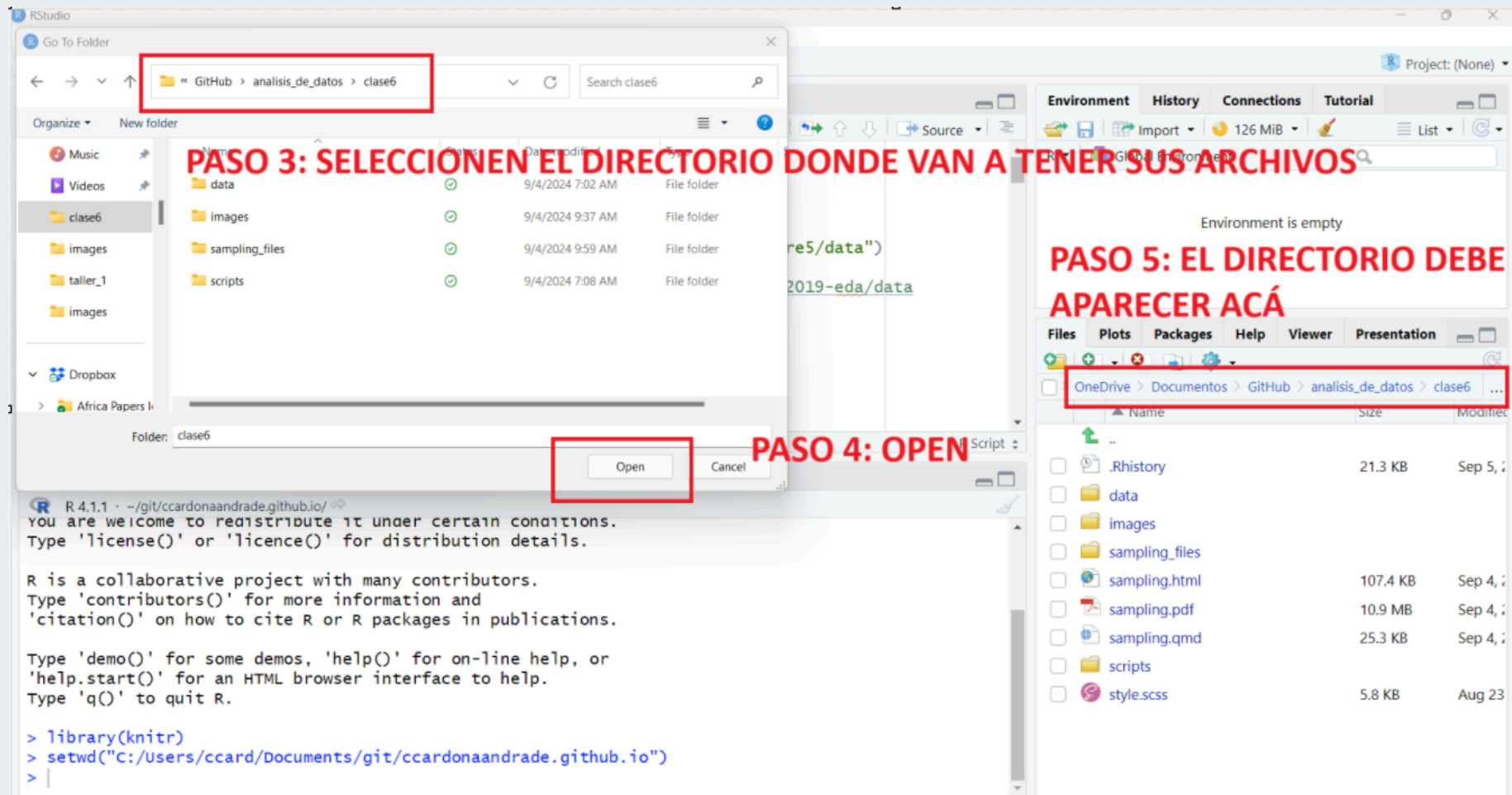
Establecer el Directorio de Trabajo Manualmente

Luego en la parte derecha, hagan clic en ... que es "Go to directory".



Establecer el Directorio de Trabajo Manualmente

Seleccionen la carpeta donde van a tener sus archivos. El directorio deberá aparecer en [Files].



Establecer el Directorio de Trabajo Manualmente

Seleccionen la tuerca en [FILES] y establezcan la carpeta como directorio de trabajo.

PASO 6: CLICK ACÁ Y SET AS WORKING DIRECTORY

EL DIRECTORIO DEBE APARECER EN LA CONSOLA

```
1 library(tidyverse)
2 library(janitor)
3 library(patchwork)
4
5 # Working directory
6 setwd(dir = "/Users/ccard/Dropbox/analitica_datos/2024-II/slides/lecture5/data")
7
8 # https://www.kaggle.com/code/lusfernandotorres/spotify-top-hits-2000-2019-eda/data
9 spotify <- read.csv("spotify_data.csv")
10
11 # veamos las dimensiones
12 dim(spotify)
13
14
15 # Para ver los nombres de las variables
```

R 4.1.1 · C:/Users/ccard/OneDrive/Documents/GitHub/analisis_de_datos/clase6/

type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(knitr)
> setwd("C:/Users/ccard/Documents/git/ccardonaandrade.github.io")
> setwd("C:/Users/ccard/OneDrive/Documents/GitHub/analisis_de_datos/clase6")
> |

- Copy...
- Copy To...
- Move...
- Copy Folder Path to Clipboard
- Open Selected in Source Pane
- Open Each File in New Columns
- Set As Working Directory
- Go To Working Directory
- Synchronize Working Directory
- Open New Terminal Here
- Show Folder in New Window
- Show Hidden Files

Ejercicio 1

1. Usando la plantilla con la que ya hemos trabajado anteriormente, establezcan el directorio de trabajo (la carpeta donde guardaron los datos de airbnb):

```
1 # Ejemplo:  
2 setwd("C:\\Users\\ccard\\Downloads")
```

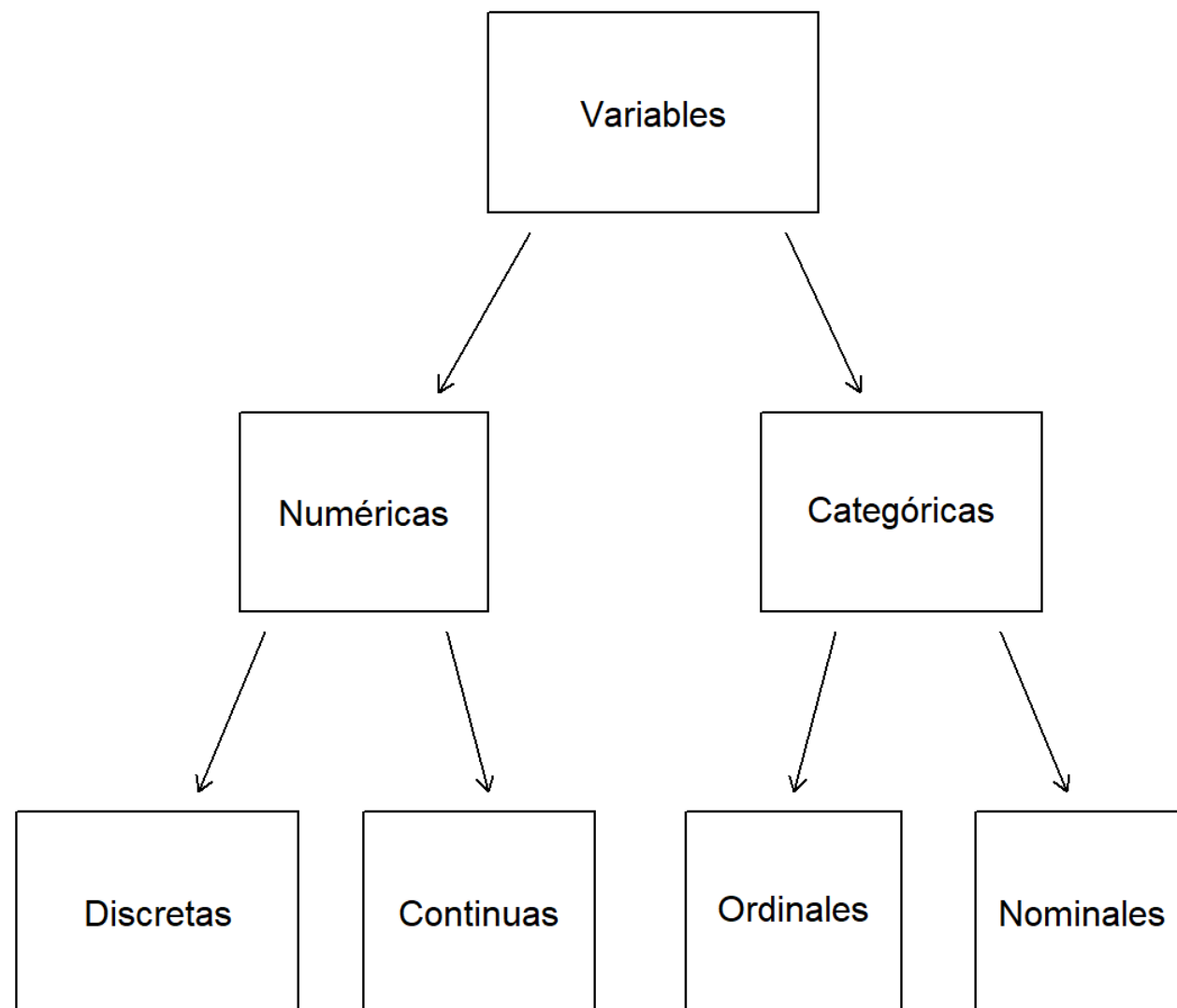
2. Carguen los datos `airbnb_ny_2019` de la siguiente manera usando el paquete `tidyverse`:

```
1 airbnb <- read.csv("airbnb_ny_2019.csv")
```

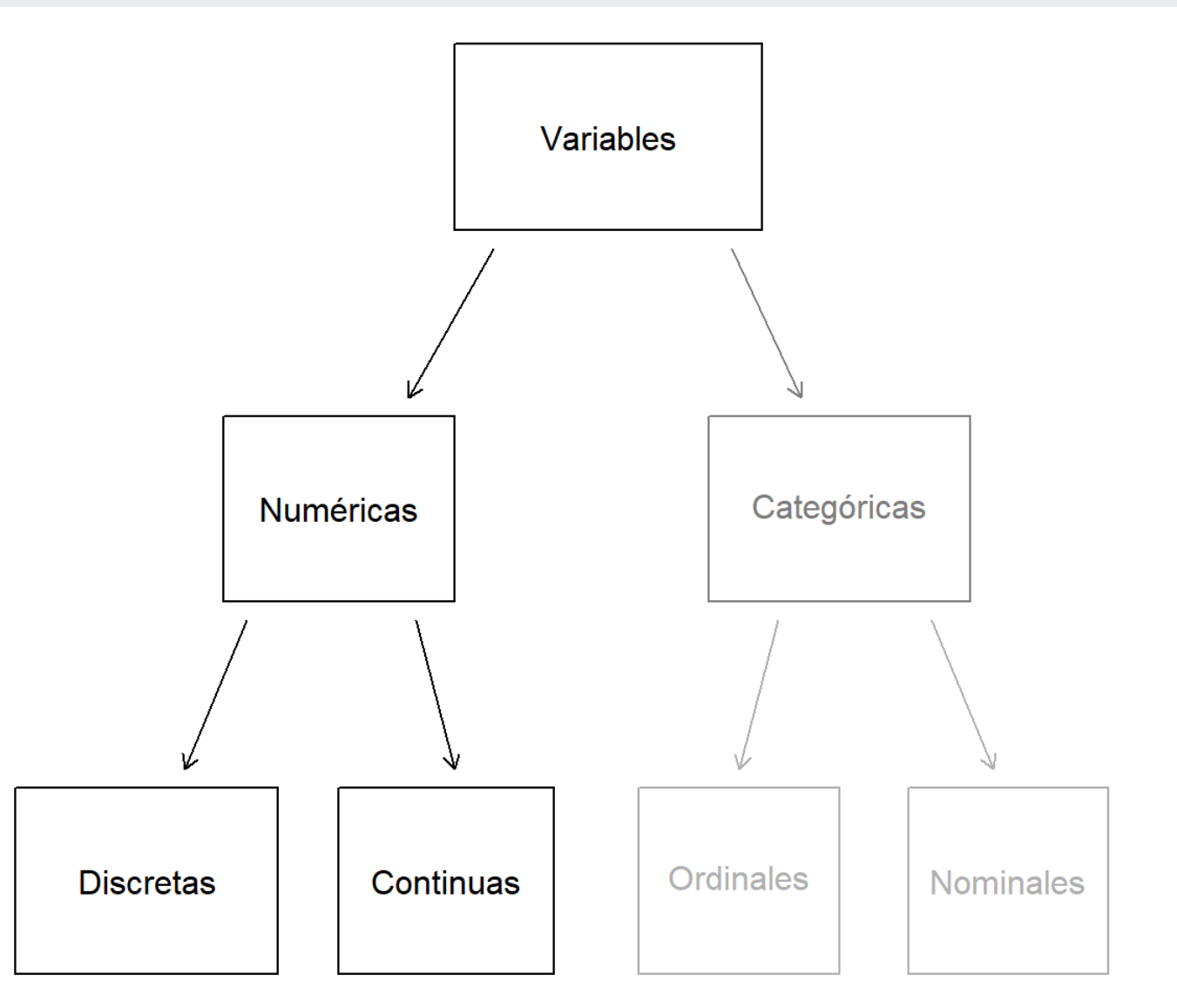
3. Exploren los datos usando la función `glimpse()`

Fundamentos sobre los datos

Tipos de variables



Variables numéricas



Variables numéricas

Una variable es **numérica** cuando puede tomar una amplia gama de valores numéricos y tiene sentido realizar operaciones aritméticas (suma, resta, promedio) con esos valores. De lo contrario, es **categorica**.

Pueden ser:

- **Discretas** si sus posibles valores forman un conjunto de valores separados, como 0, 1, 2, 3...
- **Continuas** si sus posibles valores forman un intervalo

Tipos de variables

```
# A tibble: 10 × 5
  spam   num_char line_breaks format number
<chr>   <dbl>      <int> <chr>   <fct>
1 no      21.7         551 html    small
2 no       7.01        183 html    big
3 yes      0.631         28 text    none
4 no       2.45         61 text    small
5 no      41.6        1088 html    small
6 no       0.057          5 text    small
7 no       0.809         17 text    small
8 no       5.23         88 html    small
9 no       9.28        242 html    small
10 no      17.2         578 html    small
```

- spam → categórica
- num_char → numérica
- line_breaks → numérica
- format → categórica
- number → categórica

El Histograma

¿Cómo hacer un histograma?

Paso 1: Dividir el rango de los valores en intervalos

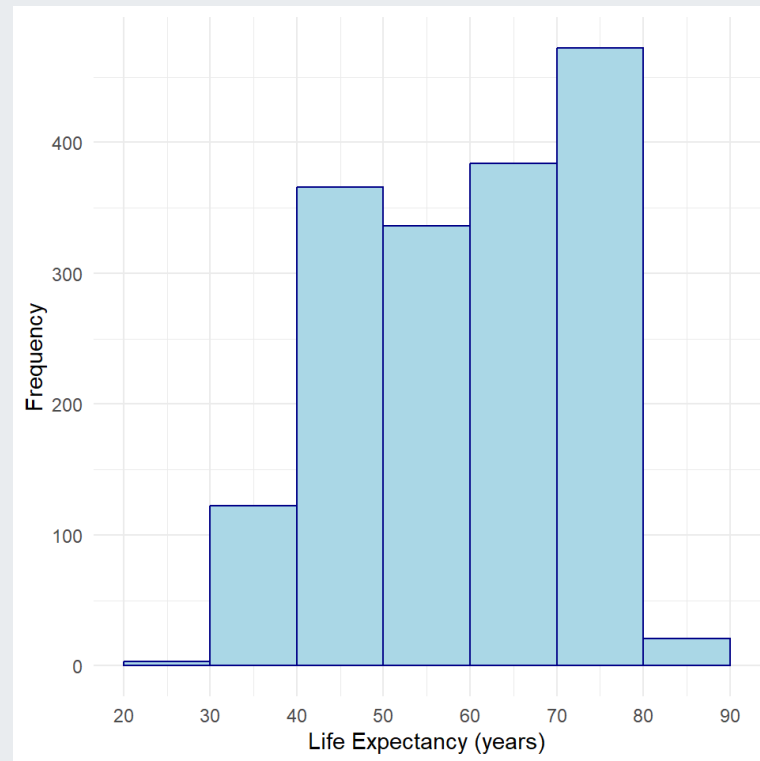
Paso 2: Contar el número de observaciones en cada intervalo

Life Expectancy	Frequency
(20 - 30]	3
(30 - 40]	122
(40 - 50]	366
(50 - 60]	336
(60 - 70]	384
(70 - 80]	472
(80 - 90]	21

¿Cómo hacer un histograma?

Paso 3: Dibujar el histograma

- No debe haber espacio entre las barras
- Nombrar el eje horizontal (con unidades!)

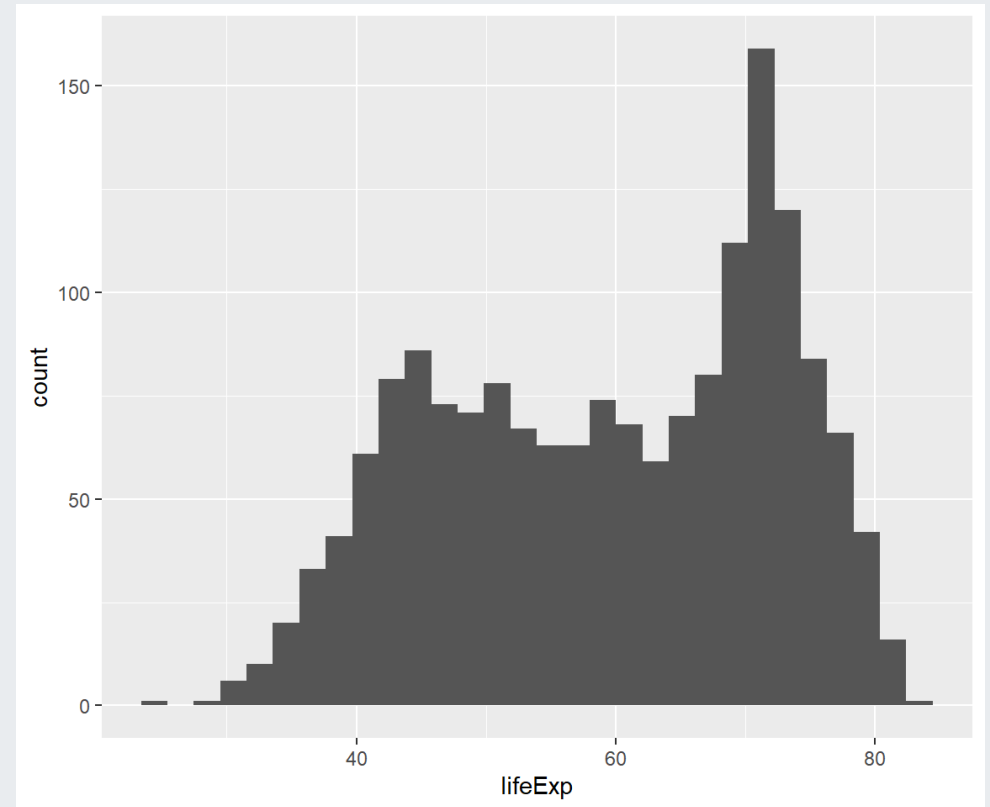


Histogramas

- Los histogramas proporcionan una visión de la **densidad de los datos**. Barras más altas indican regiones con más observaciones.
- Los histogramas son especialmente útiles para describir la **forma** de la distribución de los datos.
- La selección del **ancho de las barras** puede alterar la forma del histograma.

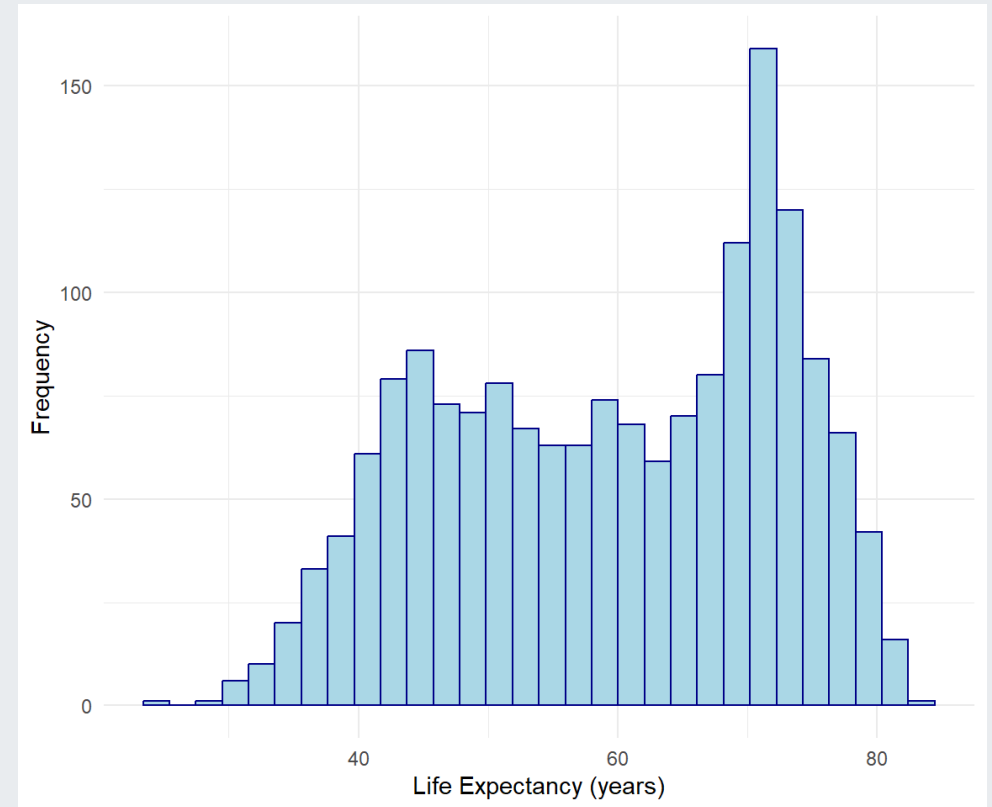
El histograma en R

```
1 ggplot(gapminder) +  
2   geom_histogram(aes(x=lifeExp))
```



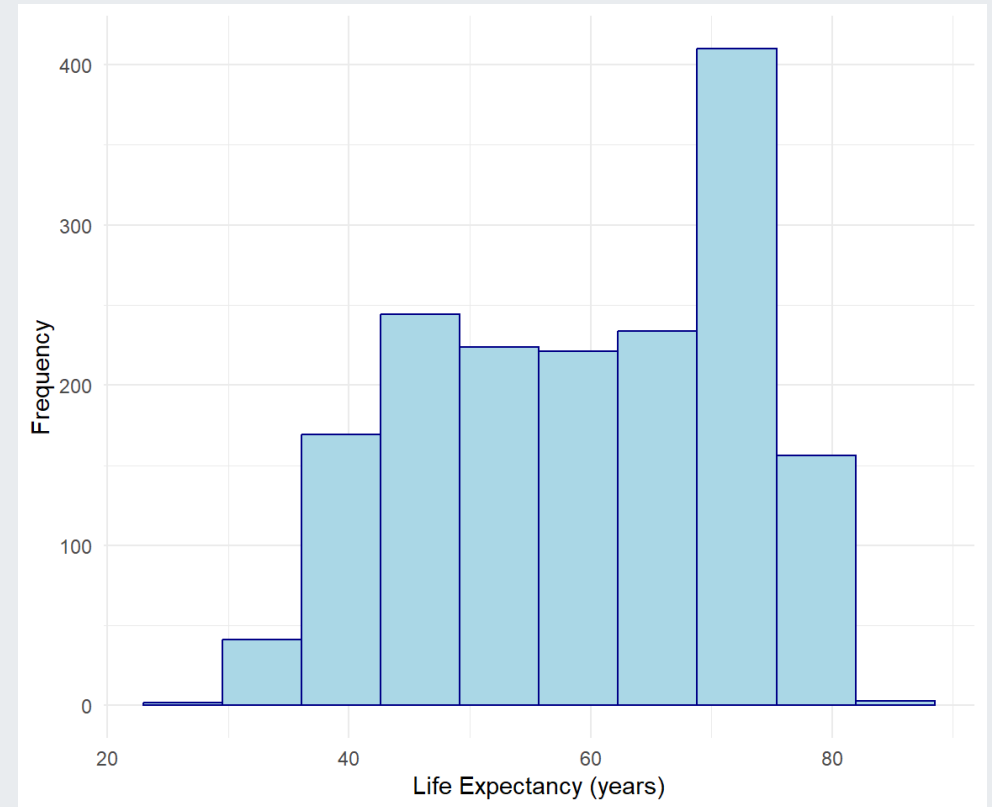
El histograma en R

```
1 ggplot(gapminder) +  
2   geom_histogram(aes(x=lifeExp),  
3                   fill = "lightblue",  
4                   color = "darkblue",  
5   labs(y = "Frequency",  
6         x = "Life Expectancy (years)",  
7   theme_minimal())
```



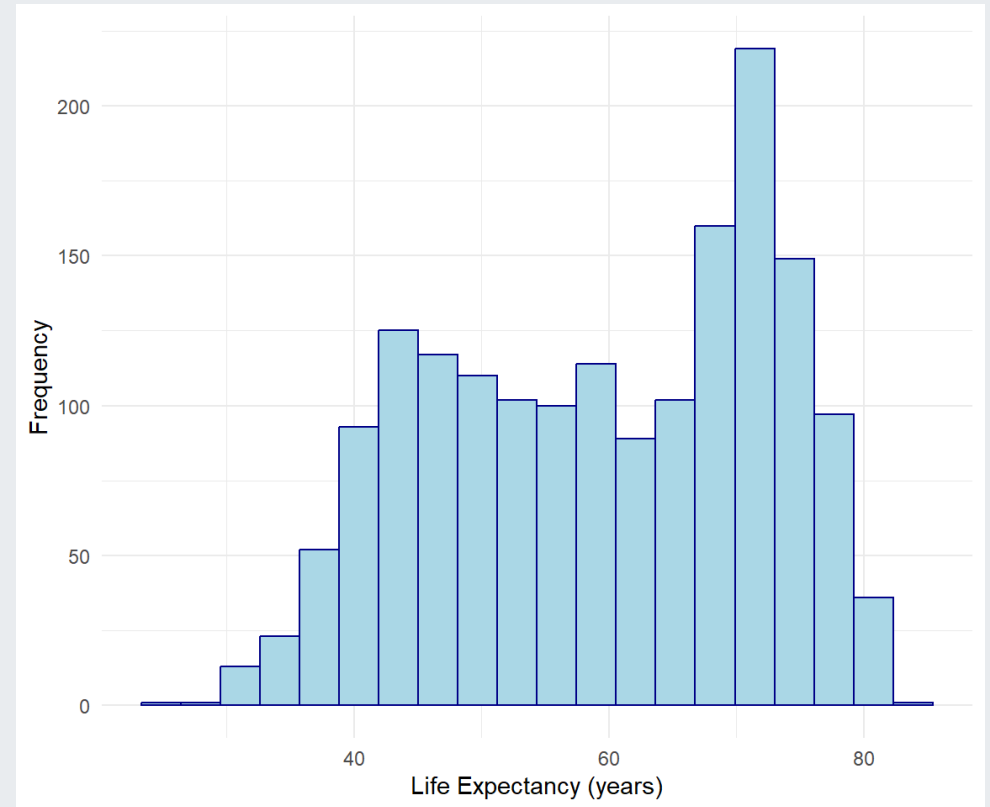
El histograma en R

```
1 ggplot(gapminder) +  
2   geom_histogram(aes(x=lifeExp),  
3                   fill = "lightblue",  
4                   color = "darkblue",  
5                   bins = 10) +  
6   labs(y = "Frequency",  
7        x = "Life Expectancy (years)",  
8   theme_minimal()
```



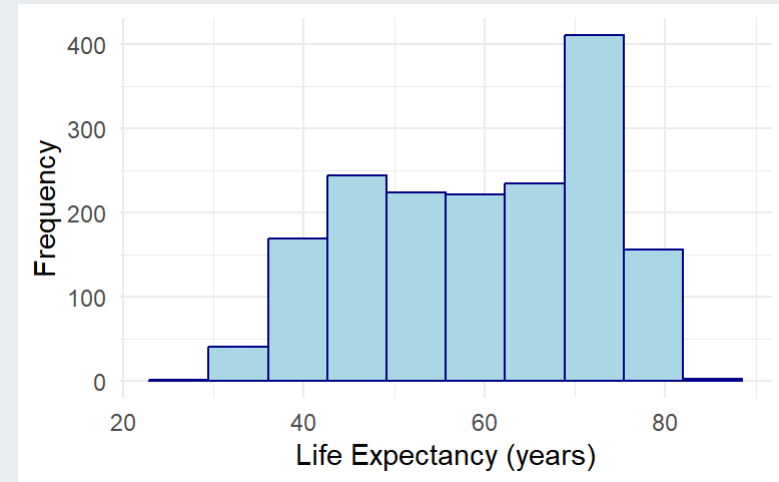
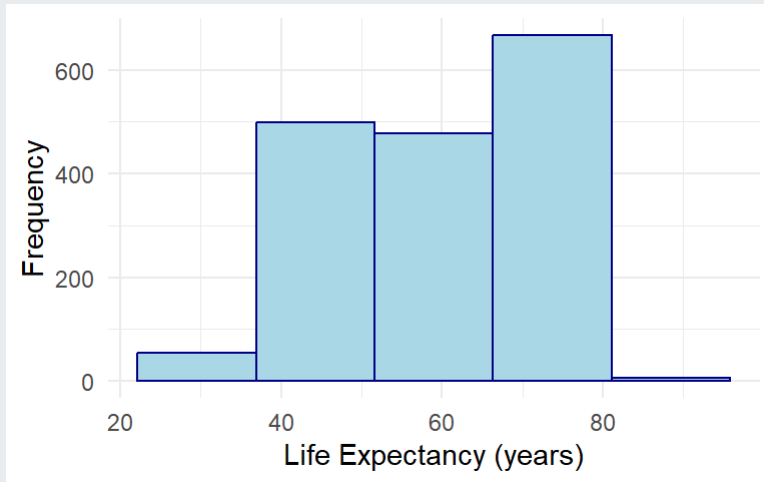
El histograma en R

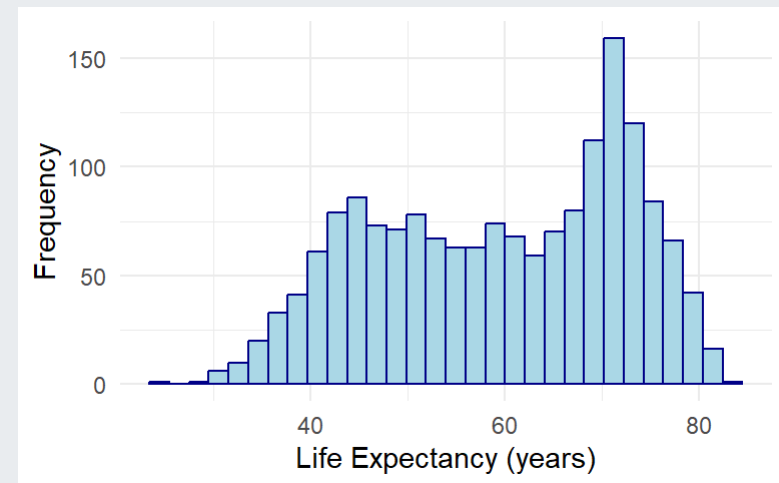
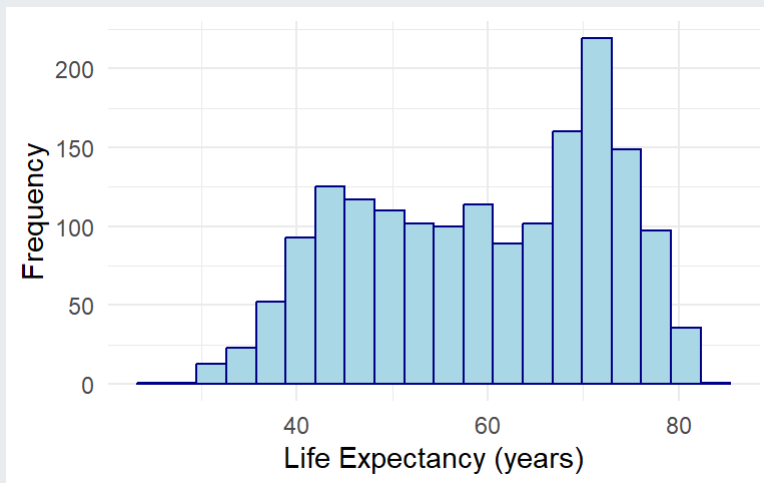
```
1 ggplot(gapminder) +  
2   geom_histogram(aes(x=lifeExp),  
3                   fill = "lightblue",  
4                   color = "darkblue",  
5                   bins = 20) +  
6   labs(y = "Frequency",  
7        x = "Life Expectancy (years)",  
8   theme_minimal()
```



Prueben diferentes anchos de barras (# bins)

¿Qué histograma revela mucho sobre los datos? ¿Cuál muy poco?





Selección del ancho de barras (# bins)

Es un proceso iterativo: prueba con diferentes números.

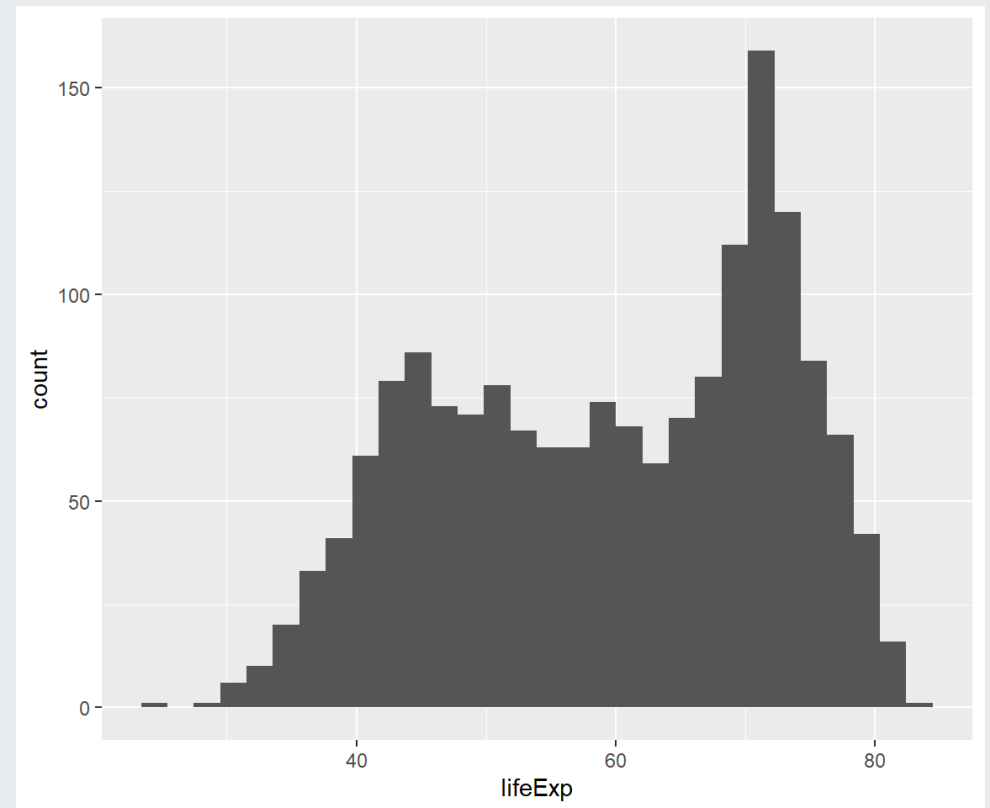
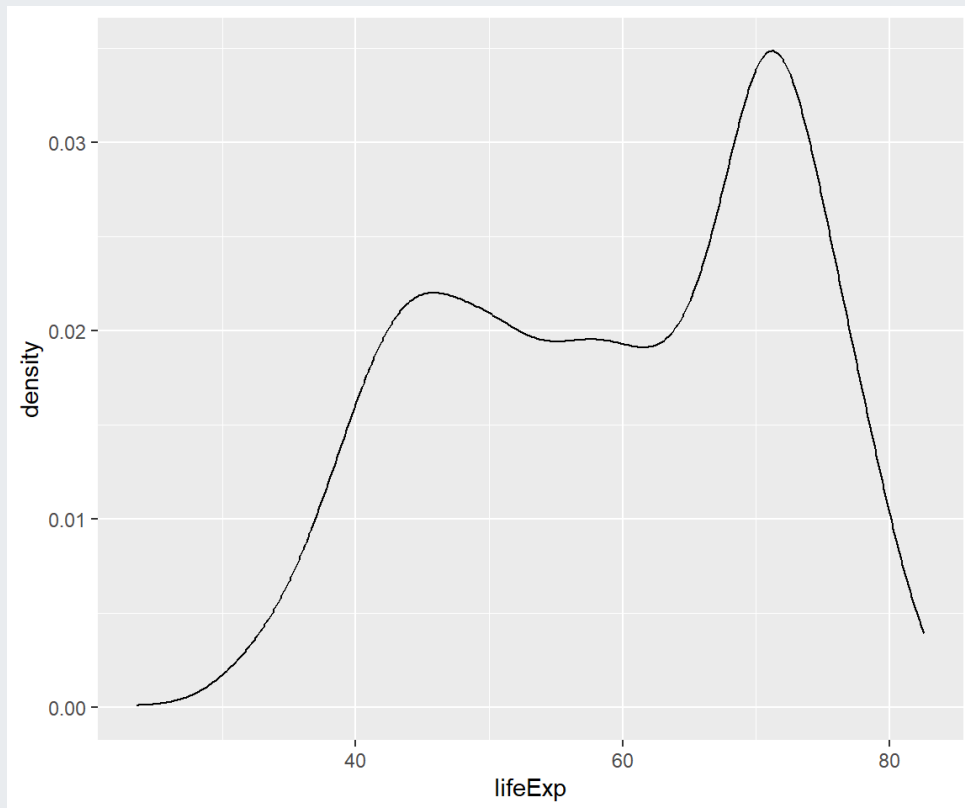
¿Qué ancho de banda deberían usar?

- No tan pocos como para que la mayoría de las barras tengan 0 o 1 observaciones.
- No tantos como para perder los detalles dentro de un barra.
- No hay un número de barras “perfecto” y único.

Regla general: **cuantas más observaciones haya, más barras se deben usar**

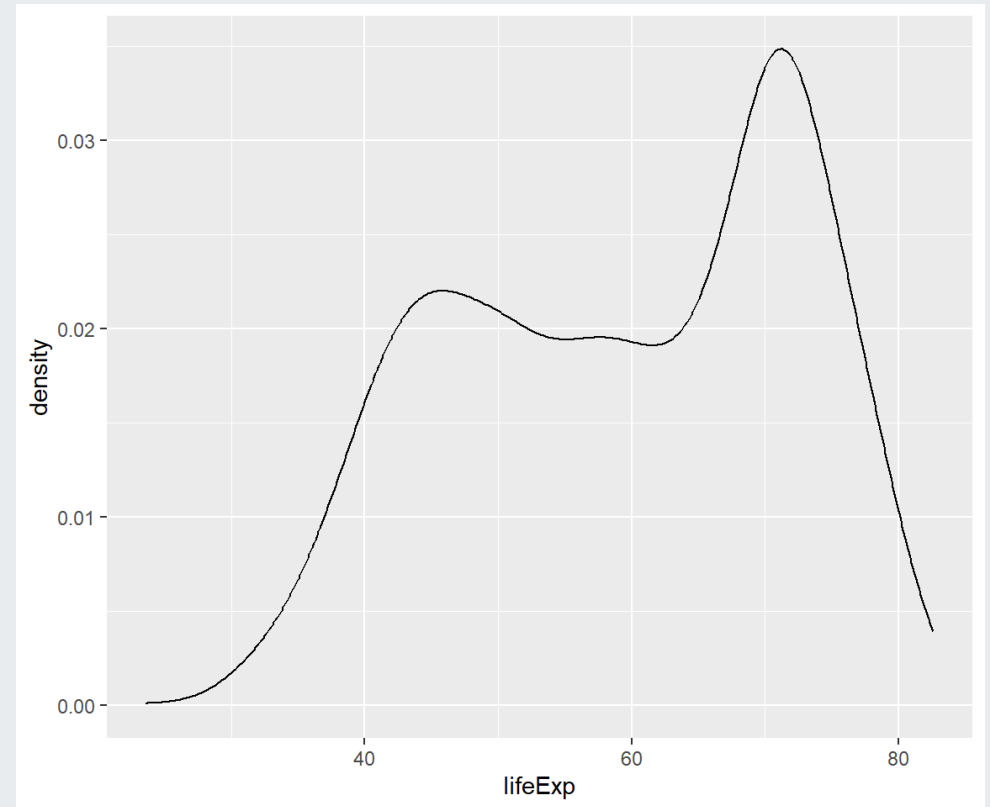
El gráfico de densidad

El gráfico de densidad es una versión suavizada del histograma. La forma, la escala y la dispersión de las observaciones son similares que en el histograma.



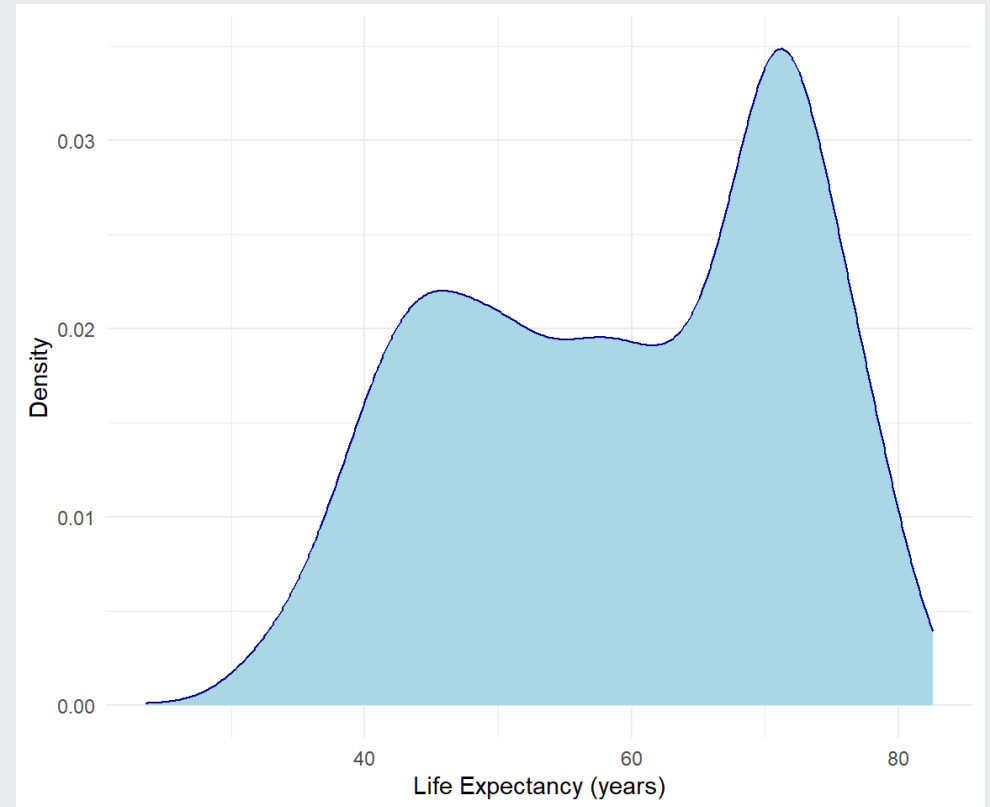
La gráfica de densidad en R

```
1 ggplot(gapminder) +  
2   geom_density(aes(x=lifeExp))
```



La gráfica de densidad en R

```
1 ggplot(gapminder) +  
2   geom_density(aes(x=lifeExp),  
3                 fill = "lightblue",  
4                 color = "darkblue")  
5 labs(y = "Density",  
6       x = "Life Expectancy (years)",  
7       theme_minimal())
```



¿Qué mirar en un histograma?

Centro

- ¿Dónde está el “medio” del histograma?
- Se representa usualmente con la **media** y la **mediana**.

Dispersión

- ¿Cuál es el rango de los datos?
- Se representa usualmente con la **desviación estándar** y el **rango intercuartílico** (se explicará pronto).

¿Qué mirar en un histograma?

Forma

- Simétrica o sesgada (asimétrica).
- Número de modas (picos).

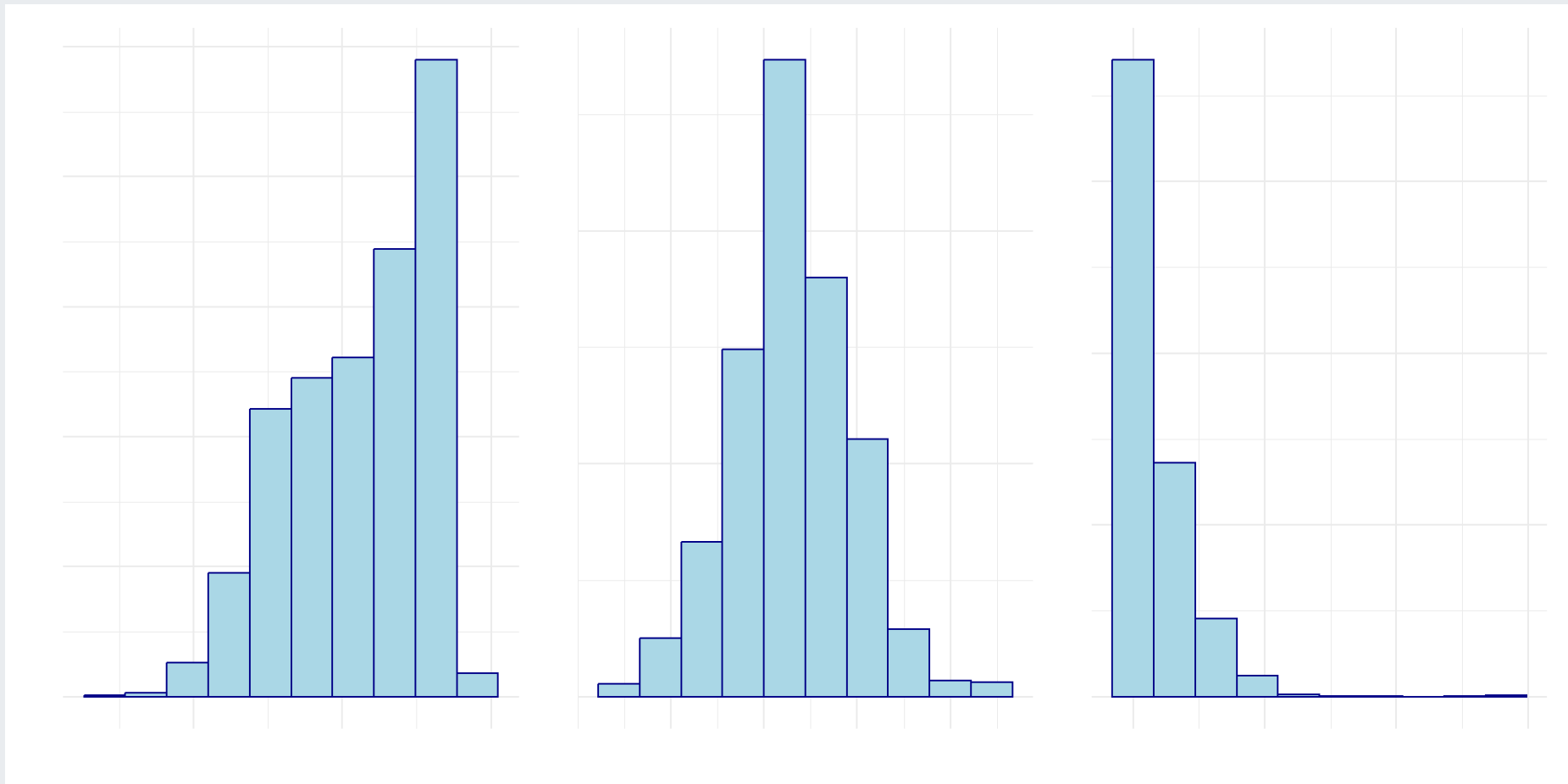
Valores atípicos (Outliers)

- ¿Hay observaciones que están fuera del patrón general?
- Pueden ser valores inusuales o errores. ¡Revísenlos!

Asimetría en los Histogramas

El sesgo mide qué tan asimétricos están distribuidos los datos

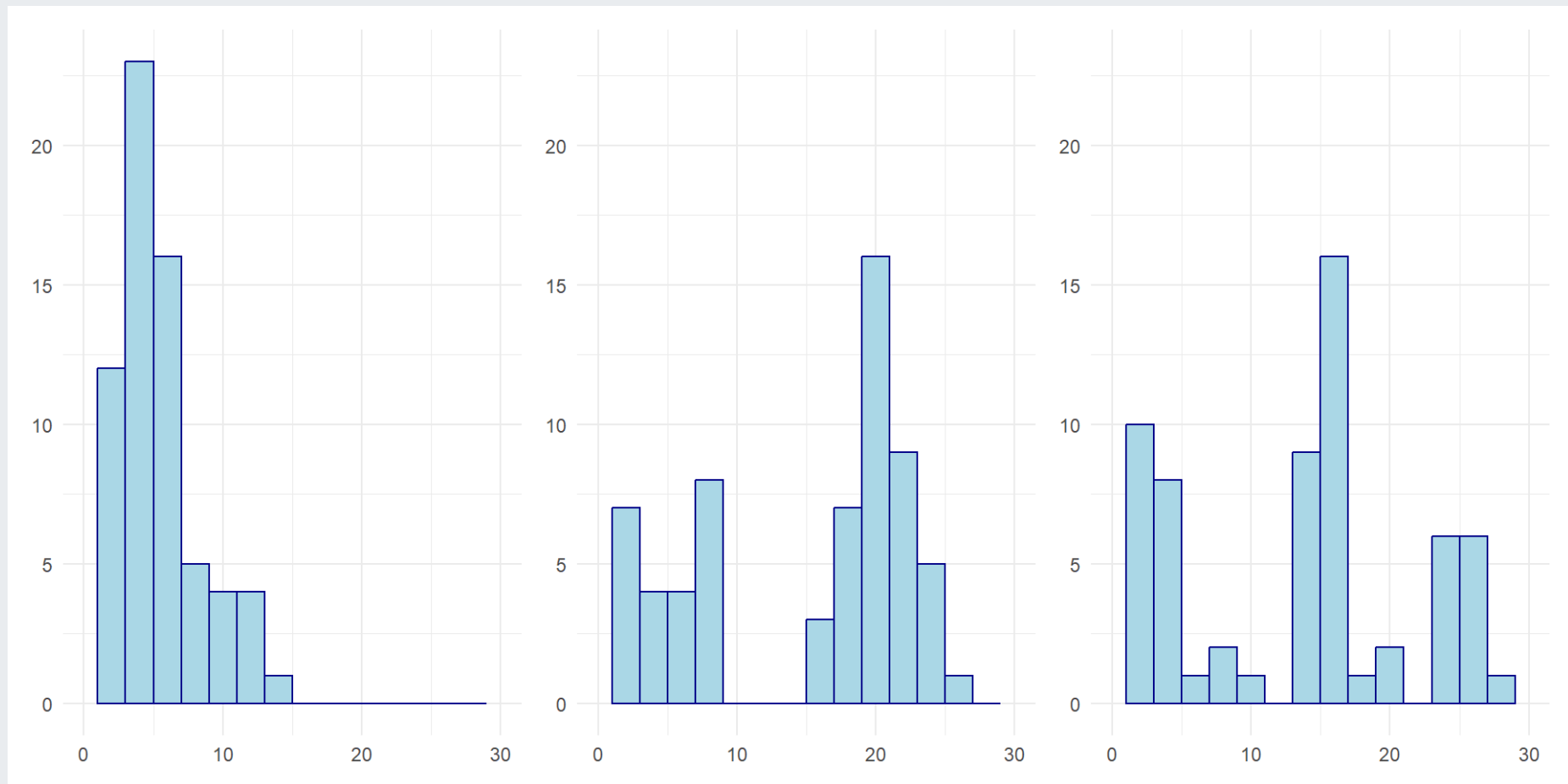
distribución sesgada a la izquierda (-), simétrica y sesgada a la derecha (+)



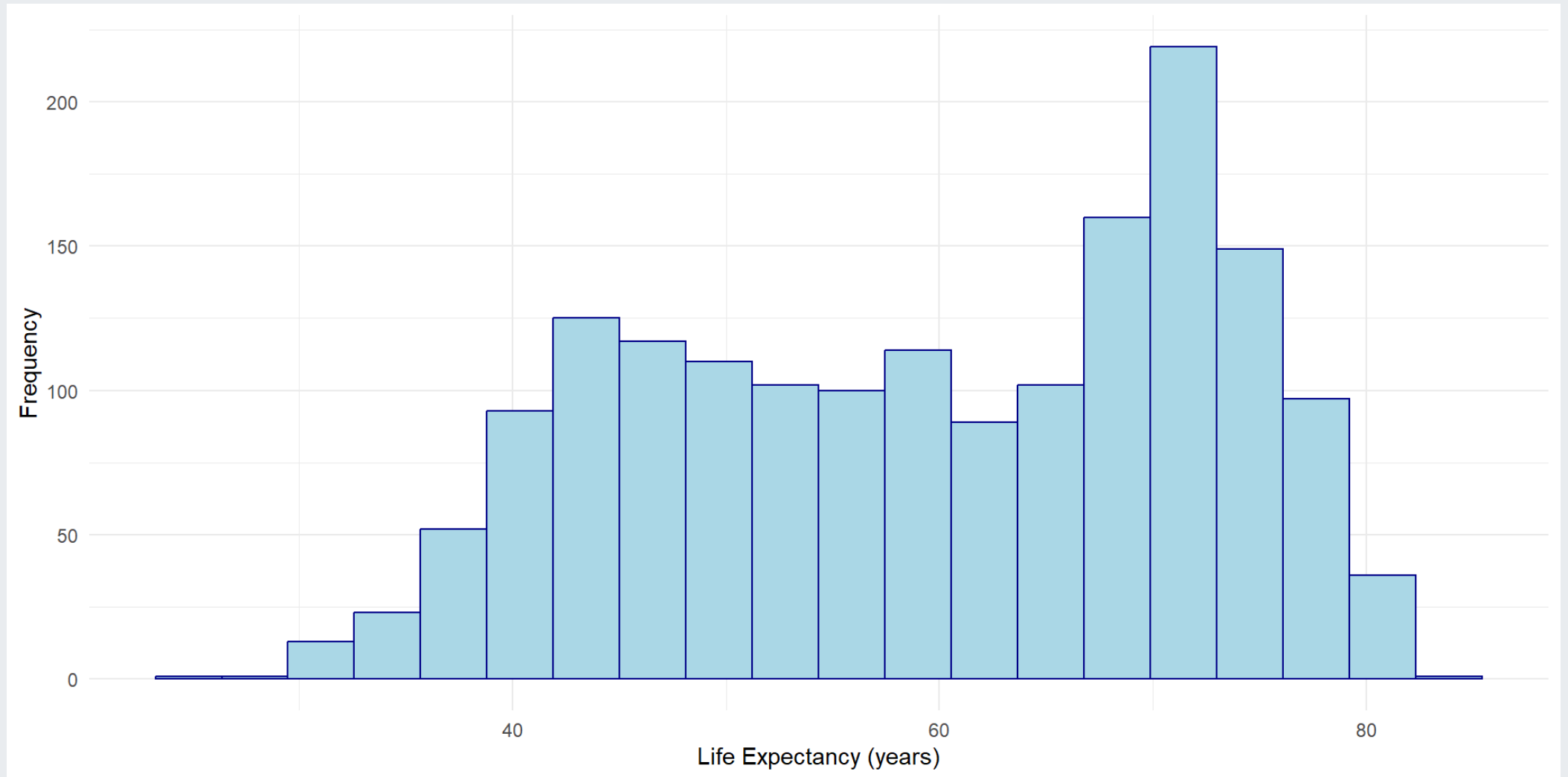
Moda en los Histogramas

La moda es el dato que más se repite en la distribución

Un ejemplo de distribución unimodal, bimodal y multimodal



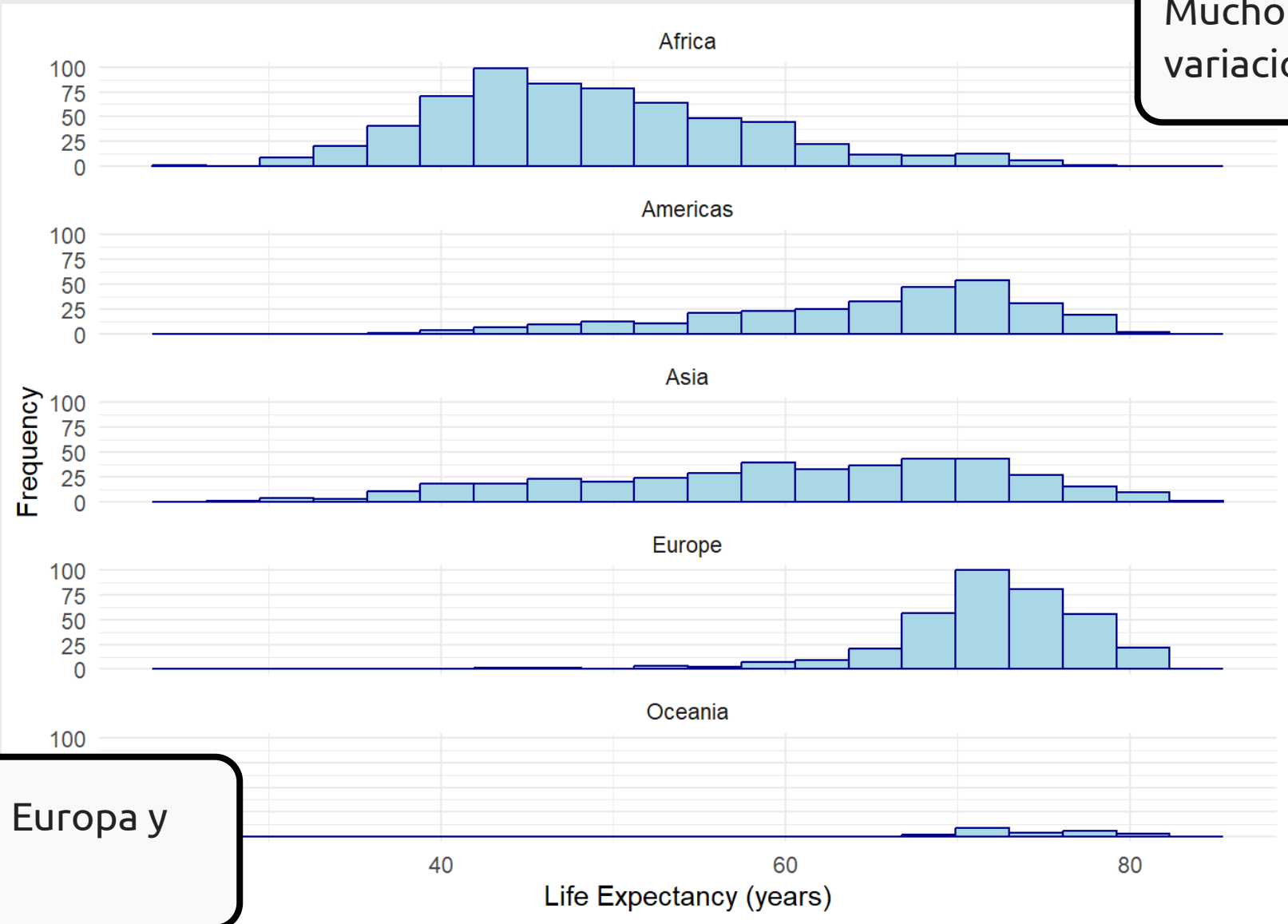
Ejemplo con la Expectativa de vida



Adicional al pico cerca a los 70 años, pareciera existir otro pico a los 40-45 años.

La expectativa de vida es...

Mucho más baja y con
variación en África



uniforme y alta en Europa y
Oceanía

Ejercicio 2

1. Dibujen un histograma para la variable `price`. ¿Existe sesgo?
2. Intenten el punto anterior para distinto número de *bins*.
3. Por la dificultad de graficar el anterior histograma, censuremos un poco los datos con el siguiente código:

```
1 newdata <- airbnb |>
2   filter(price < 3 * sd(price, na.rm = TRUE))
```

4. Intenten graficar el histograma nuevamente y elijan el número de *bins* que para ustedes provea más información.

Medidas de tendencia central

La media

Es la suma de todos los valores dividida entre el número de valores observados:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Ejemplo. Supongamos que tenemos los siguientes valores:

4, 8, 3, 5, 13

La media de la variable será:

$$\bar{x} = \frac{4 + 8 + 3 + 5 + 13}{5} = \frac{33}{5} = 6.6$$

La mediana

Es el valor que denota el punto medio en una distribución ordenada. En otras palabras, 50% de los valores están por debajo de este valor.

Ejemplo 1 Supongamos que tenemos los siguientes valores: 4, 8, 3, 5, 13.

datos	→	4	8	3	5	13
organizados	→	3	4	5	8	13

La mediana es 5.

La mediana

Es el valor que denota el punto medio en una distribución ordenada. En otras palabras, 50% de los valores están por debajo de este valor.

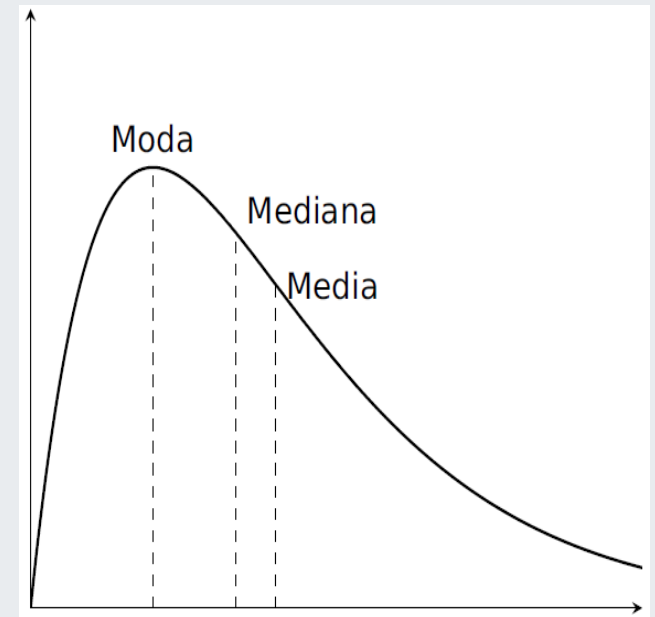
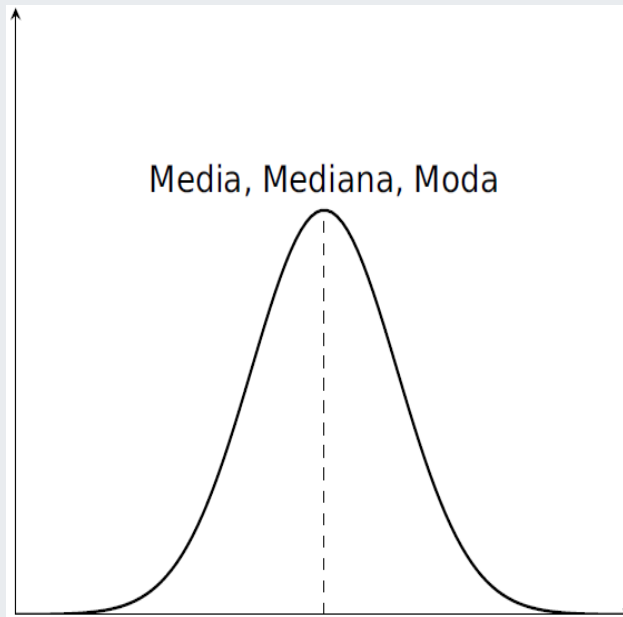
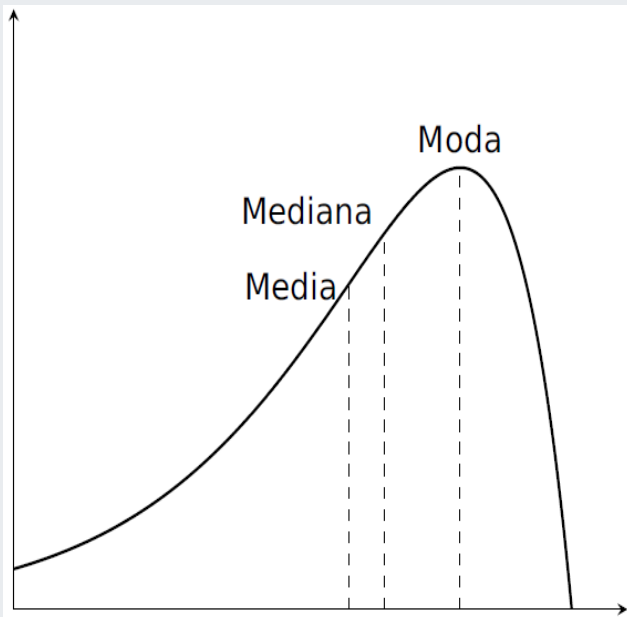
Ejemplo 2 Supongamos que tenemos los siguientes valores: 4, 8, 3, 5, 13, 12..

datos	→	4	8	3	5	13	12
organizados	→	3	4	5	8	12	13

La mediana es $\frac{5+8}{2} = 6,5$.

Tendencia central y sesgo

La asimetría de una distribución está relacionada a la ubicación de las medidas de tendencia central dentro de la misma distribución.

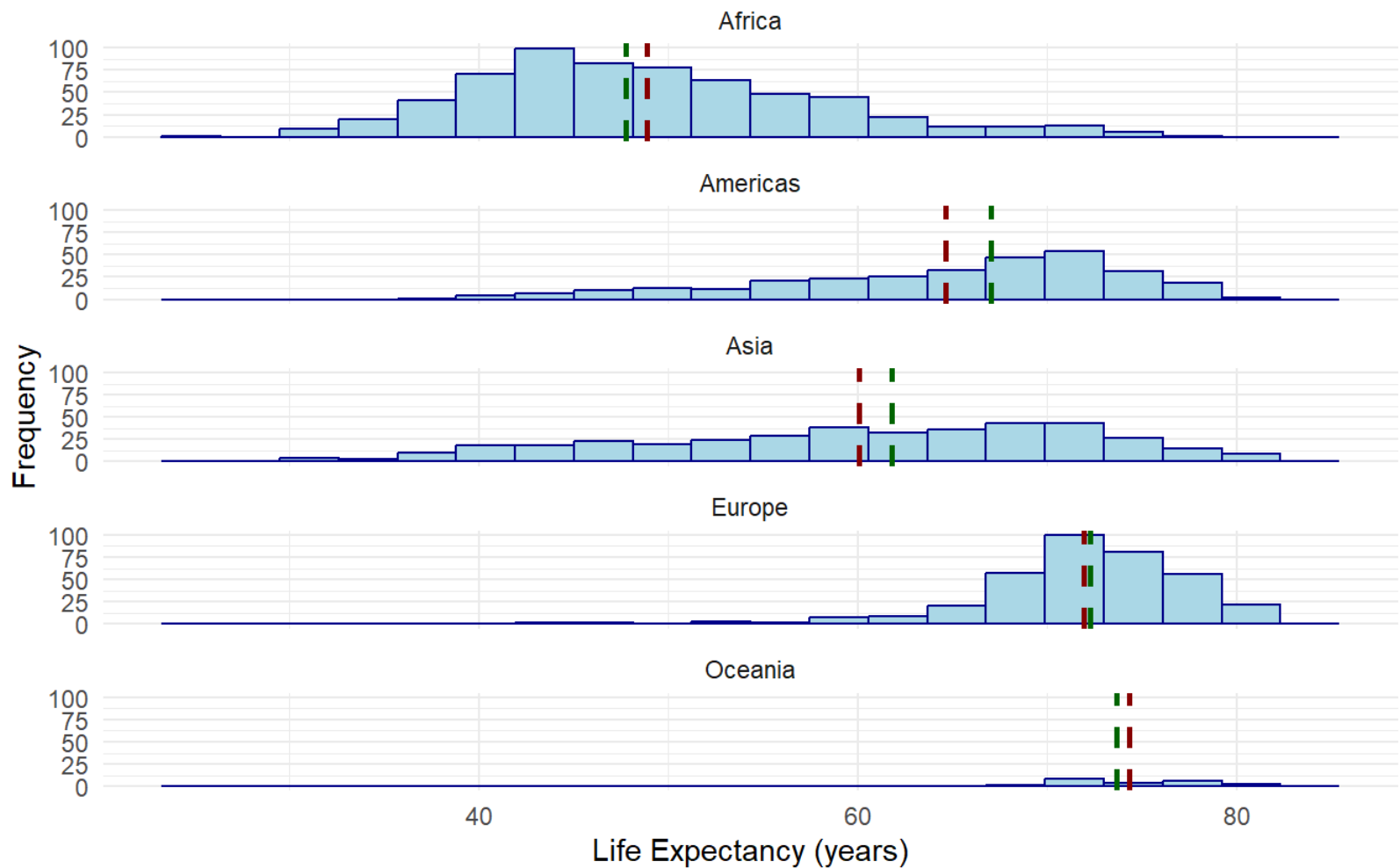


Media vs Mediana

¿Cuál distribución es simétrica?

¿Cuál distribución es sesgada?

Mean Median



La media y la mediana en R

La media se calcula con la función `mean()`:

```
1 mean(gapminder$lifeExp)
[1] 59.47444
```

La mediana se calcula con la función `median()`:

```
1 median(gapminder$lifeExp)
[1] 60.7125
```

Comparemos para cada continente:

```
1 gapminder |>
2   group_by(continent) |>
3   summarise(mean_lexp = mean(lifeExp),
4             median_lexp = median(lifeExp))
```

A tibble: 5 × 3

	continent	mean_lexp	median_lexp
	<fct>	<dbl>	<dbl>
1	Africa	48.9	47.8
2	Americas	64.7	67.0
3	Asia	60.1	61.8



Ejercicio 3

1. Calculen la media y la mediana para cada `neighbourhood_group`.
2. Basado en el punto anterior, ¿qué tipo de sesgo tiene cada barrio?

Resumen de los Cinco Números

Cuartiles y Resumen de Cinco Números

- Los cuartiles dividen los datos en 4 partes iguales.
 - Primer cuartil (Q_1) = percentil 25:
El 25% de los datos están por debajo y el 75% por encima.
 - Segundo cuartil (Q_2) = mediana = percentil 50
 - Tercer cuartil (Q_3) = percentil 75:
El 75% de los datos están por debajo y el 25% por encima.
- Rango intercuartílico (RIC) = $Q_3 - Q_1$

Cuartiles

Se calculan de acuerdo a la posición en los datos ordenados:

$$Q_1 : \frac{n + 1}{4}$$

$$Q_2 : \frac{n + 1}{2}$$

$$Q_3 : \frac{3(n + 1)}{4}$$

Donde n es el número de valores.

Cuartiles

- está en la posición
- Calculamos el promedio de los valores en la posición 2 y 3.

Cuartiles

- posición
- posición

Resumen de los cinco números

- Min: 11
- : 12.5
- Mediana: 16
- : 19.5
- Max: 22

RIC

- : $19.5 - 12.5 = 7$

Cálculo de los cuartiles

De hecho, no existe un consenso sobre el cálculo de los cuartiles. Hay varias fórmulas para los cuartiles, que varían de un libro a otro y de un software a otro.

Retomemos el ejemplo de la diapositiva anterior, donde $= 12.5$, $= 16$ y $= 19.5$

```
1 x = c(11,12,13,16,16,17,18,21,22)
2 summary(x)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
11.00	13.00	16.00	16.22	18.00	22.00

```
1 fivenum(x)
```

```
[1] 11 13 16 18 22
```

Cálculo de los cuartiles

Incluso diferentes comandos en  a veces reportan diferentes cuartiles:

```
1 y = c(43, 35, 43, 33, 38, 53, 64, 27, 34, 27)
2 summary(y)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
27.00	33.25	36.50	39.70	43.00	64.00

```
1 fivenum(y)
```

```
[1] 27.0 33.0 36.5 43.0 64.0
```

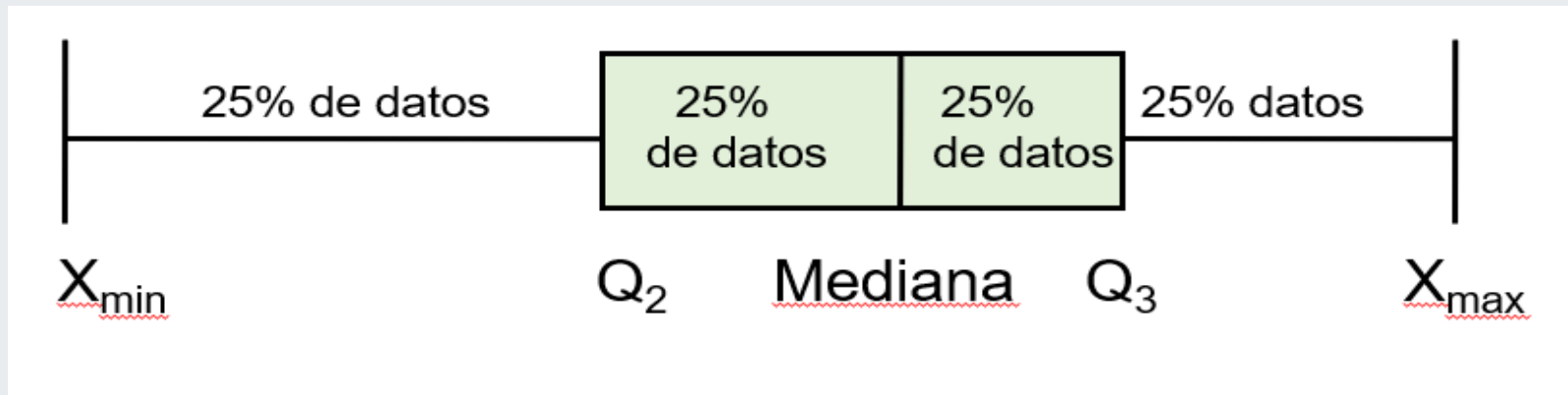
No se preocupen por la fórmula. Simplemente tengan en cuenta que:

Los cuartiles dividen los datos en 4 partes iguales

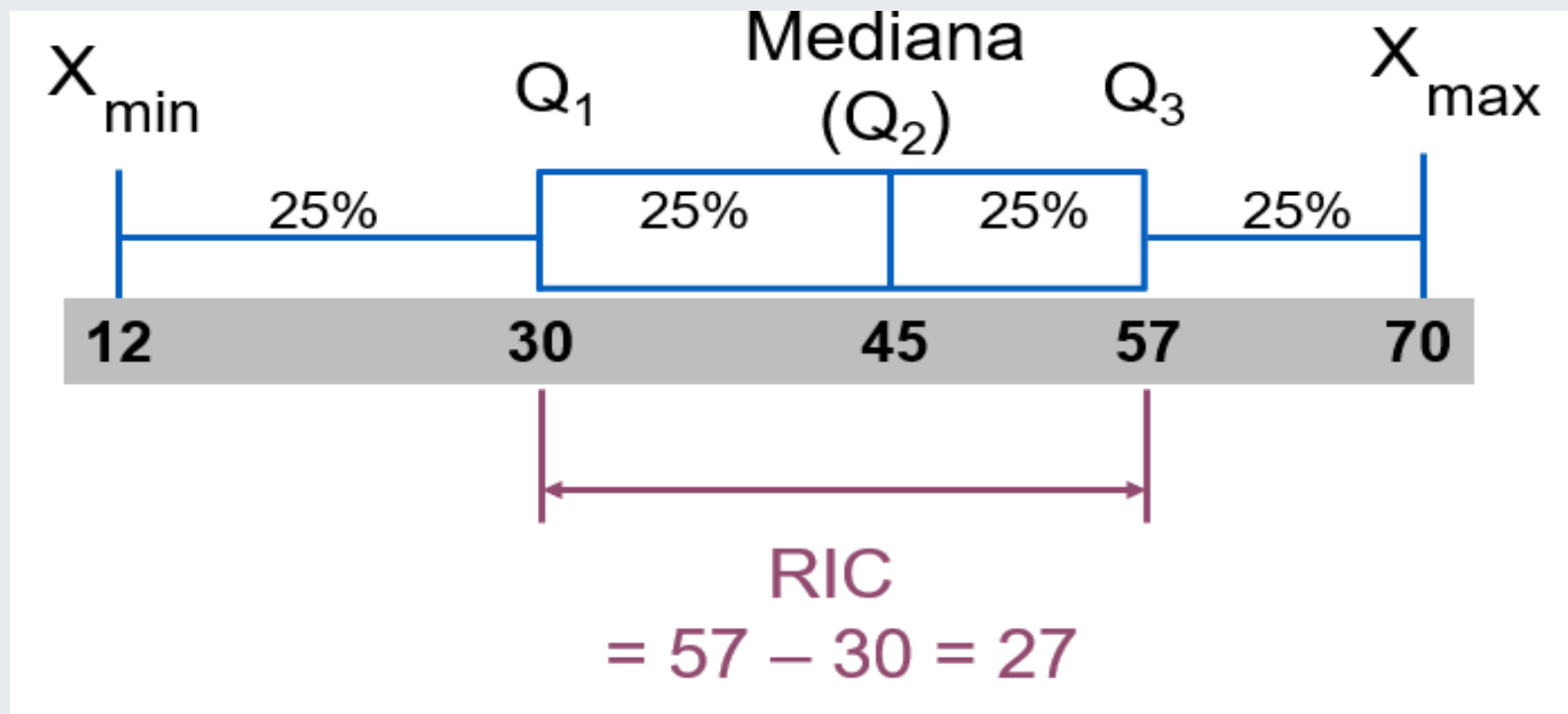
Diagrama de caja

El diagrama de caja

El diagrama de caja es la manera más común de visualizar los 5 estadísticos que explicamos anteriormente. Al mismo tiempo, un diagrama de caja identifica observaciones *inusuales*.

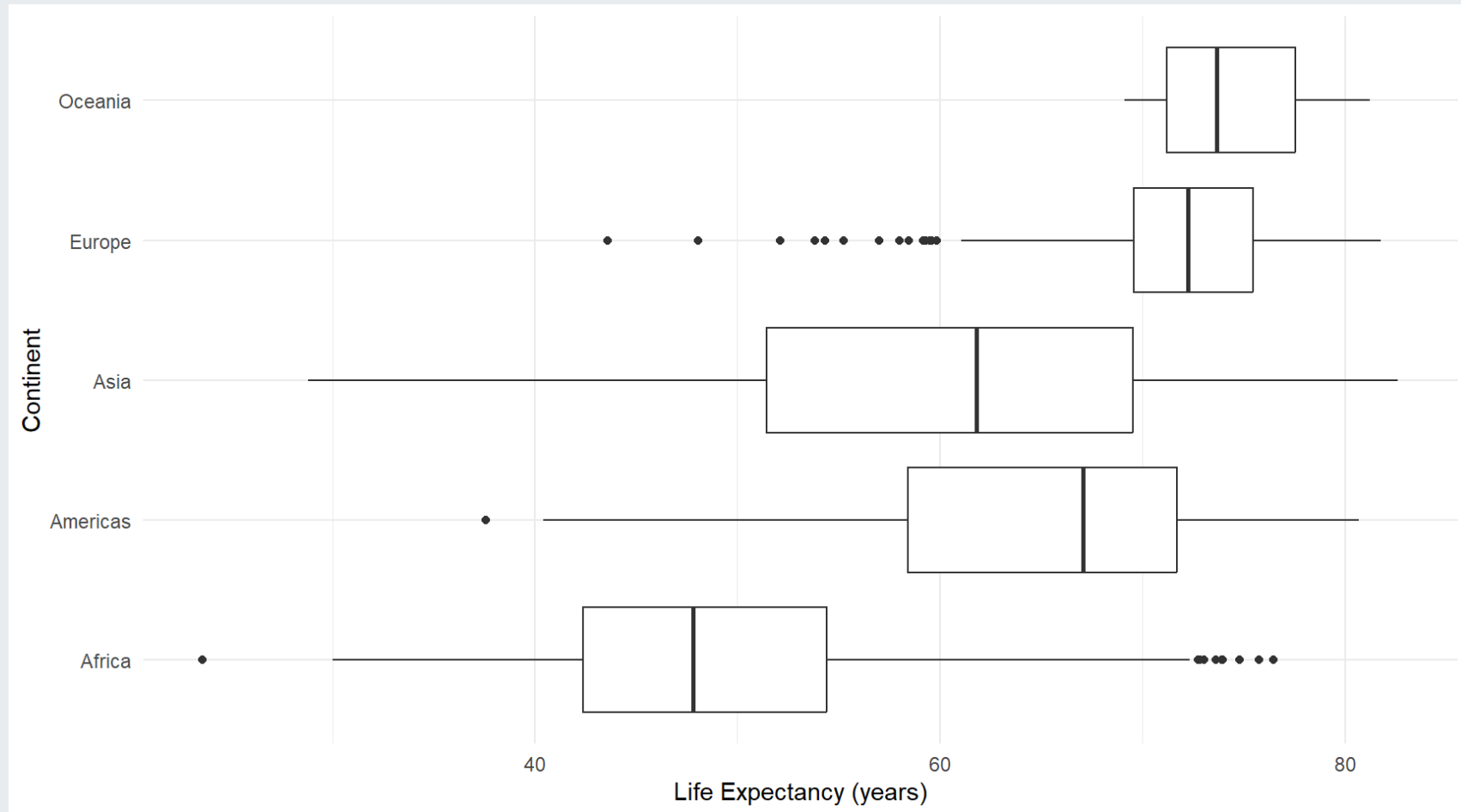


El diagrama de caja - RIC



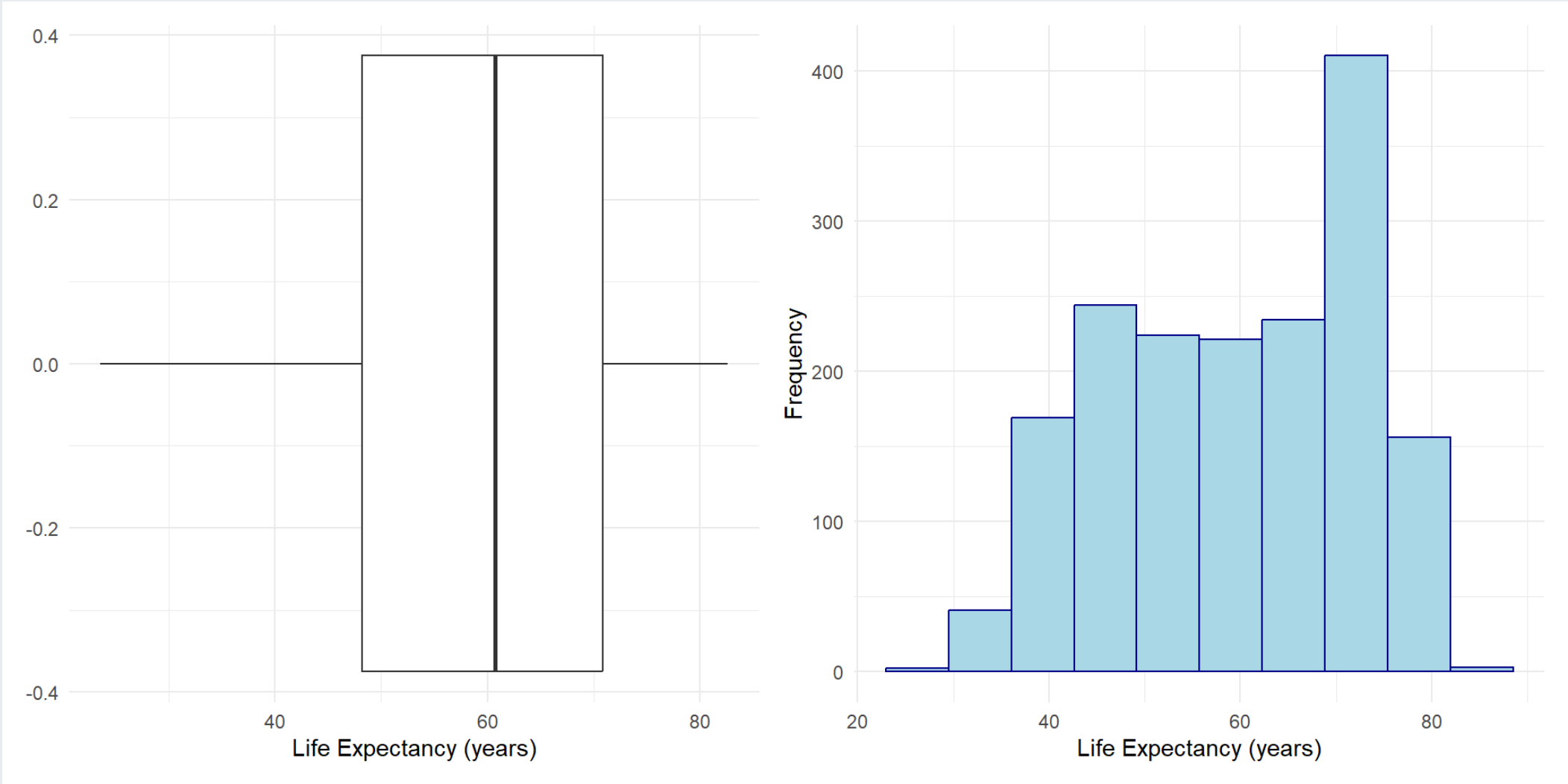
El **RIC** es una medida de la variabilidad de los datos. El RIC tiende a ser mayor si la variación de los datos también es mayor.

Diagrama de caja - Outliers



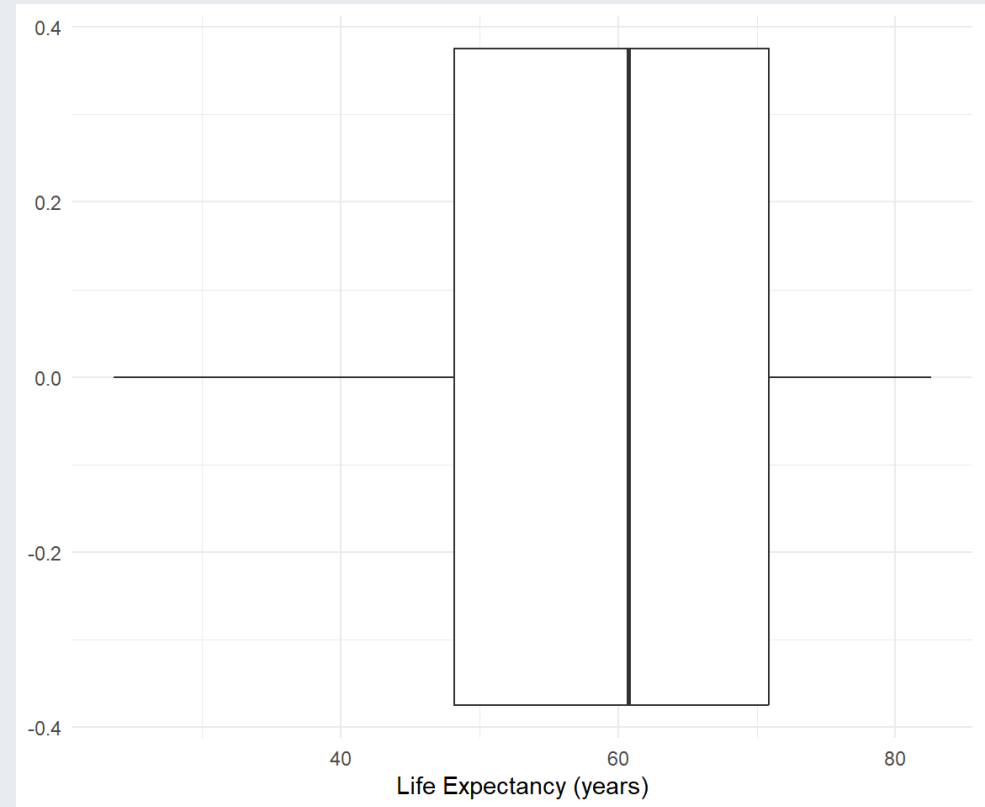
Una observación se identifica como un posible valor atípico (*outlier*) si se encuentra más de $1.5 \times \text{RIC}$ por debajo de o por encima de

Diagrama de caja vs Histograma



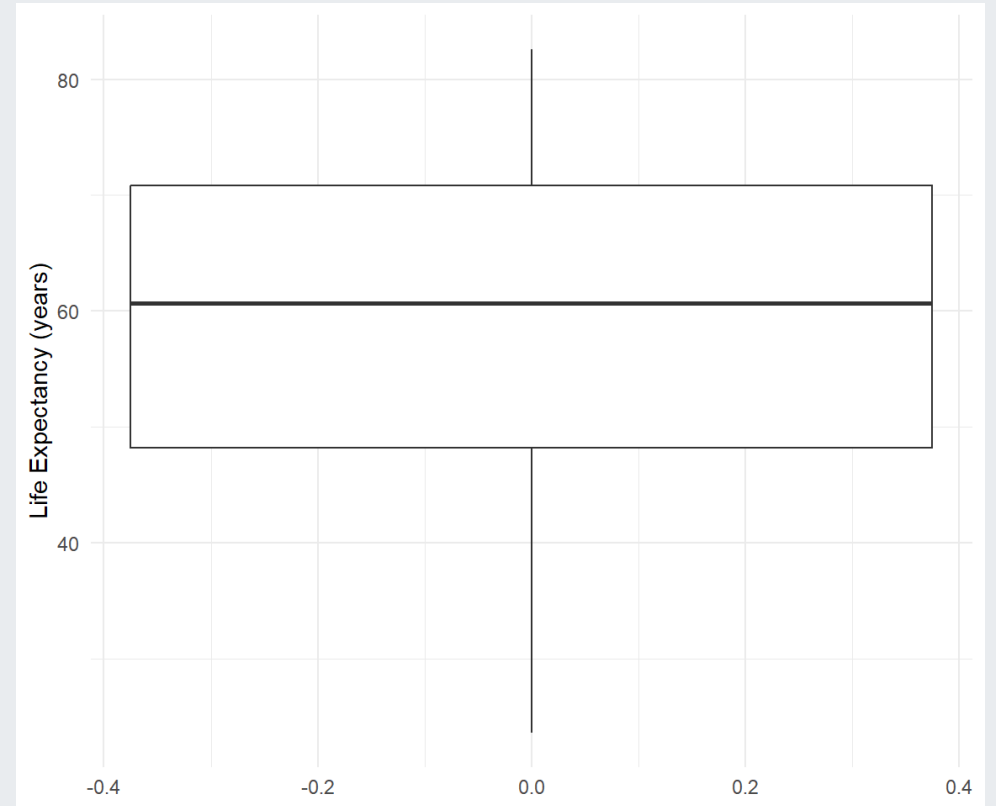
El diagrama de caja en R

```
1 ggplot(gapminder) +  
2   geom_boxplot(aes(x=lifeExp)) +  
3   labs(x = "Life Expectancy (years"  
4   theme_minimal()
```



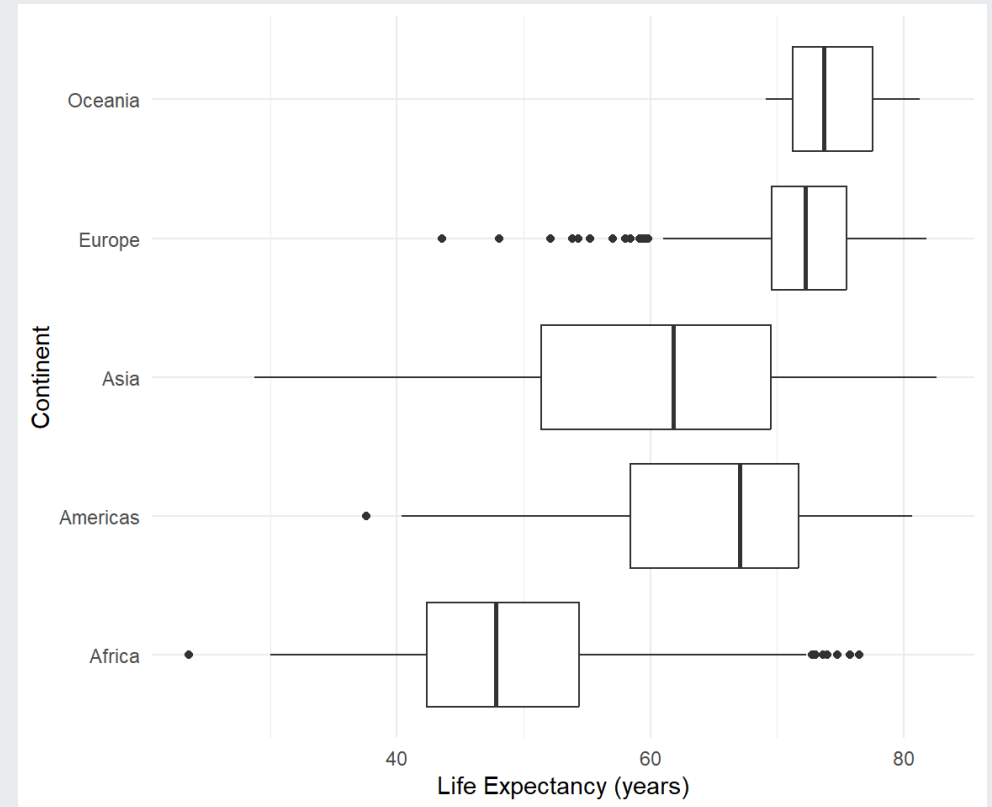
El diagrama de caja en R

```
1 ggplot(gapminder) +  
2   geom_boxplot(aes(y=lifeExp)) +  
3   labs(y = "Life Expectancy (years)  
4   theme_minimal()
```



El diagrama de caja en R

```
1 ggplot(gapminder) +  
2   geom_boxplot(aes(x=lifeExp , y=c  
3     labs(x = "Life Expectancy (years  
4         y = "Continent") +  
5     theme_minimal()
```



Ejercicio 4

1. Grafiquen los diagramas de caja para cada `neighbourhood_group`.
2. Describan con 3 aspectos las 5 gráficas. Dos ejemplos: todos los barrios tienen una presencia fuerte de outliers y Manhattan tiene la mediana mayor.
3. Usando la función `IQR()`, calculen el rango intercuartílico para cada barrio.
4. ¿Qué pueden decir a partir del punto anterior?

Medidas de dispersión

El Rango

Es la distancia cubierta por los valores en una distribución, es decir, la distancia entre menor y el mayor valor

Se calcula de la siguiente manera:

Usualmente es más útil reportar el mínimo y el máximo que reportar el rango

La desviación estándar

Describe la forma en que los valores de una variable se dispersan a lo largo de la distribución en relación a la media.

Se calcula siguiendo la fórmula:

La desviación estándar

Tomemos el siguiente conjunto de datos como ejemplo:

La media es .

La **desviación** es la distancia de un valor a la media. El siguiente paso es calcular la desviación de cada valor:

La desviación estándar

Si calculamos el cuadrado de estas desviaciones y luego calculamos su promedio hallaremos la varianza muestral:

Dividimos por n para que la varianza muestral sea más confiable y útil, según ciertas propiedades estadísticas.

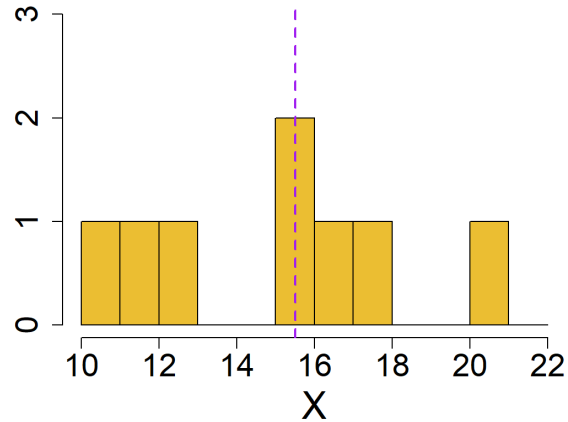
La desviación estándar se define como la raíz cuadrada de la varianza:

Algunas características de la desviación estándar

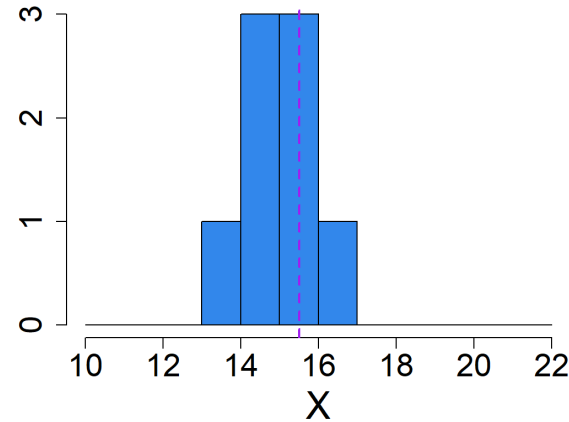
- Sumar la misma constante a cada valor, NO modifica la desviación estándar
- Multiplicar cada valor por la misma constante, aumenta la desviación estándar en la misma proporción
- A diferencia de la varianza, la desviación estándar está en las mismas unidades que la variable original

La desviación estándar vs la distribución de los datos

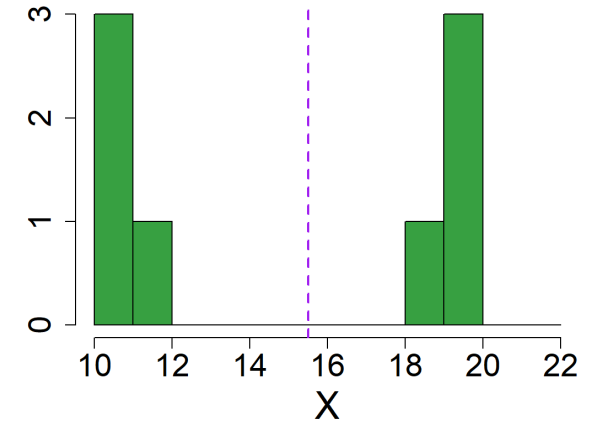
Datos A = 11,12,13,16,16,17,18,



Datos B = 14,15,15,15,16,16,16,



Datos C = 11,11,11,12,19,20,20,

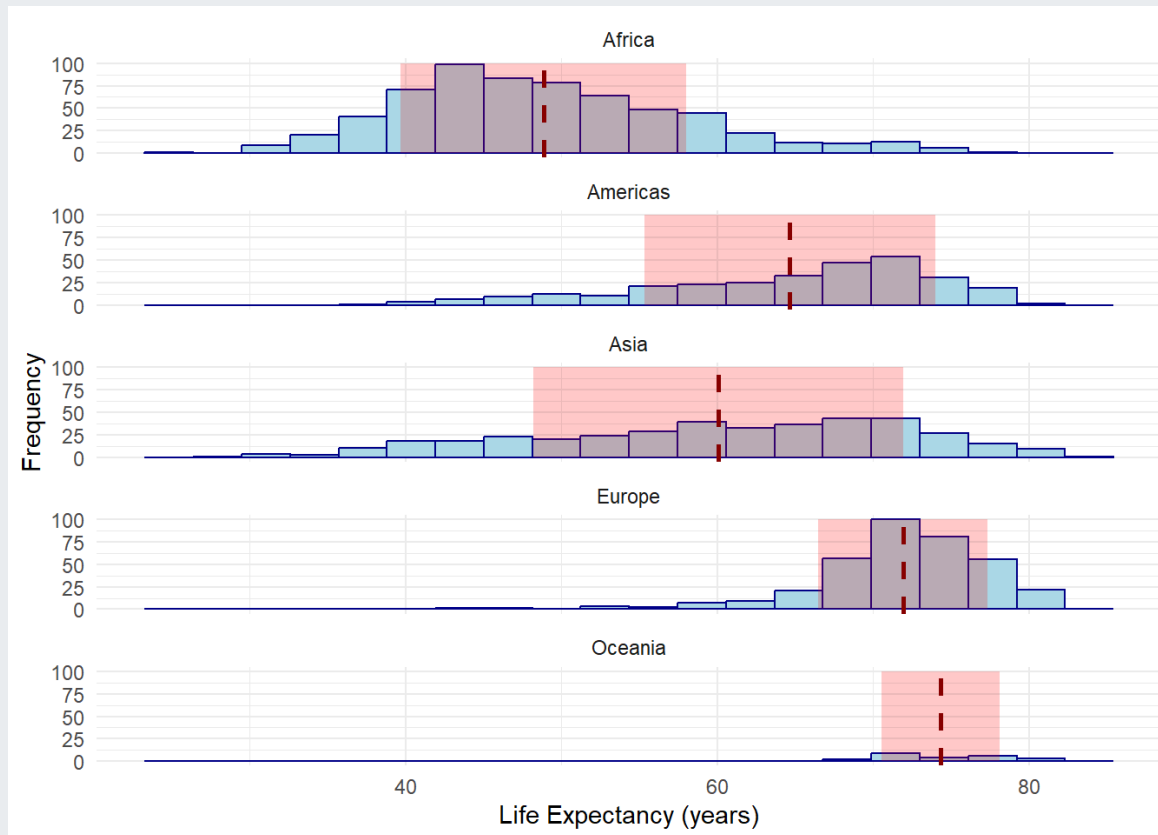


- Noten que las 3 muestras tienen la misma media:
- Las desviaciones estándar son $S=3.33$, $S=0.92$ and $S=4.56$.

La regla 68% y 95%

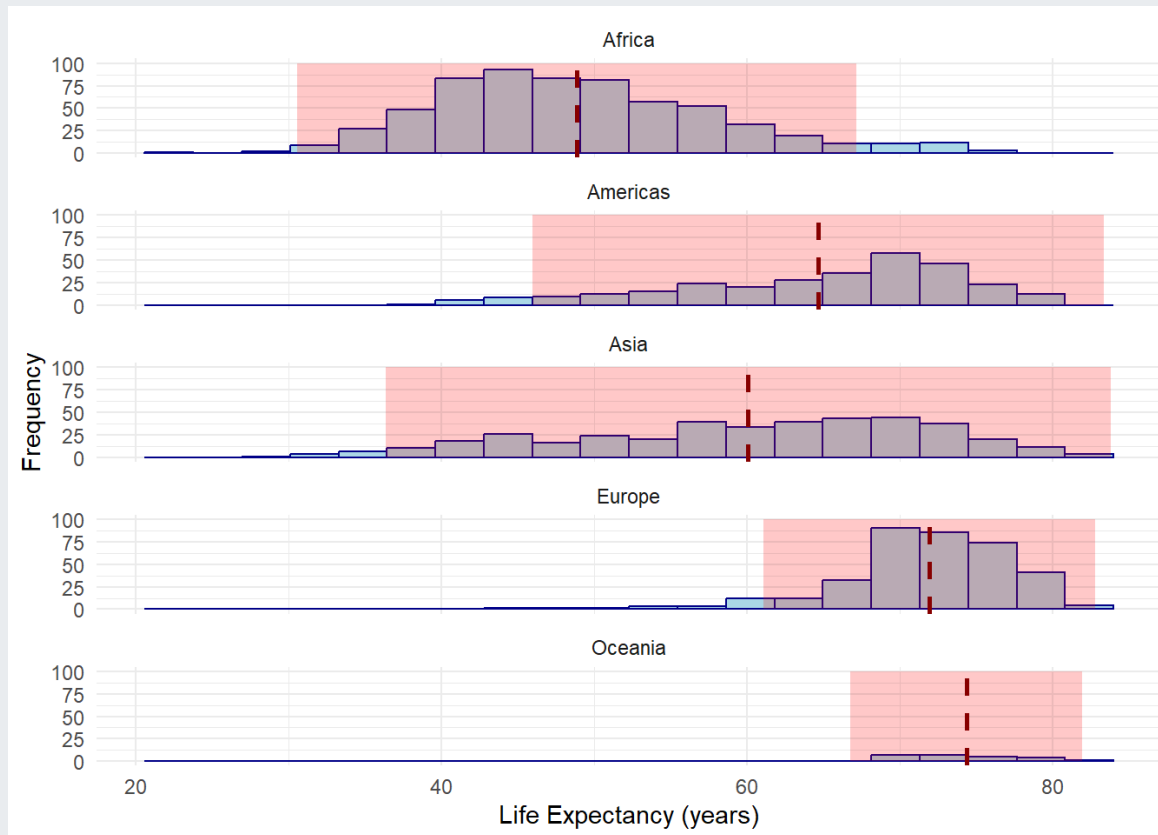
- Aproximadamente el 68% de las observaciones estarán dentro de 1 desviación estándar de la media.
- Aproximadamente el 95% de las observaciones estarán dentro de 2 desviaciones estándar de la media.
- Ambas reglas funcionan **muy bien** para datos con forma de campana y **razonablemente bien** para datos unimodales y no muy sesgados, pero no para todos los datos.

La regla 68% y 95%



```
# A tibble: 5 × 2
  continent proportion_1SD
  <fct>         <dbl>
1 Africa      0.691
2 Americas    0.683
3 Asia        0.641
4 Europe      0.742
5 Oceania     0.625
```

La regla 68% y 95%



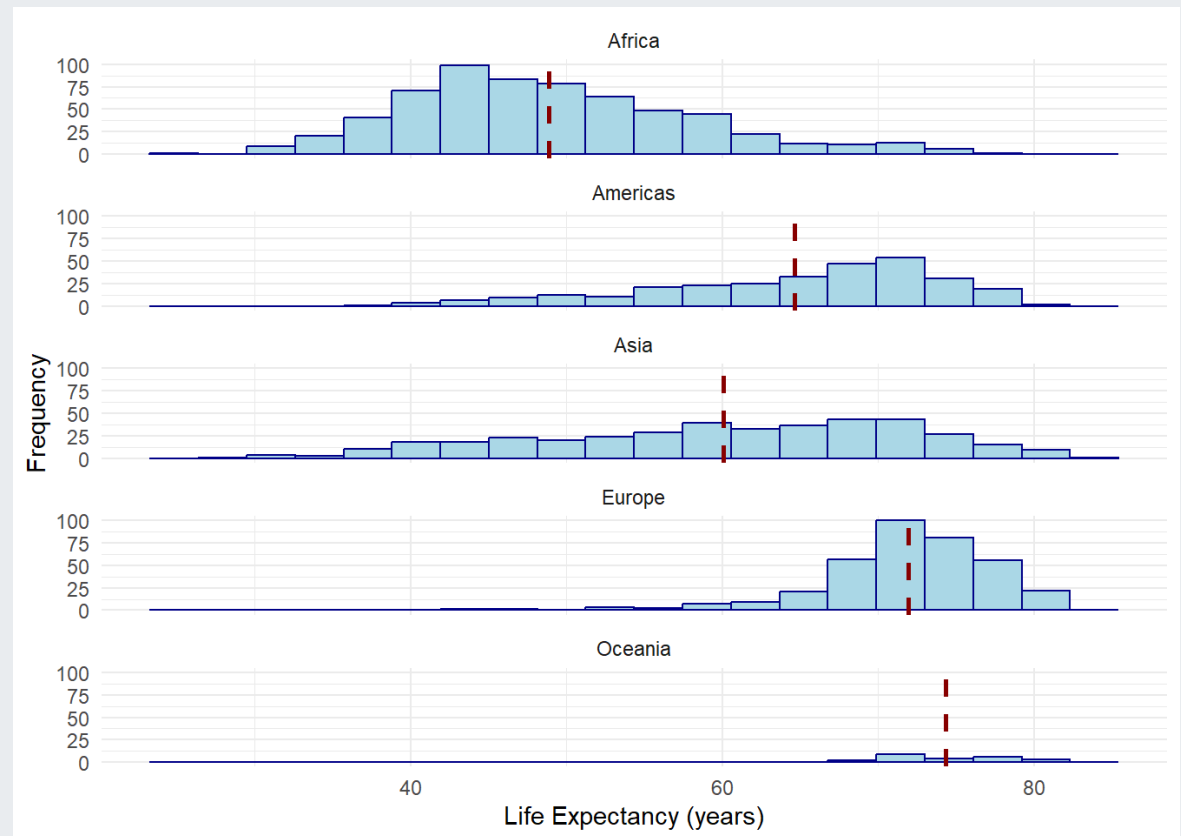
```
# A tibble: 5 × 2
  continent proportion 2SD
  <fct>          <dbl>
1 Africa         0.947
2 Americas        0.947
3 Asia           0.972
4 Europe          0.958
5 Oceania         1
```

La desviación estándar en R

La desviación estándar se calcula con la función **sd**:

```
1 gapminder |>
2   group_by(continent) |>
3   summarise(
4     mean_lExp = mean(lifeExp),
5     sd_lExp = sd(lifeExp)
6   )
```

```
# A tibble: 5 × 3
  continent mean_lExp sd_lExp
  <fct>      <dbl>    <dbl>
1 Africa      48.9      9.15
2 Americas    64.7      9.35
3 Asia        60.1     11.9
4 Europe      71.9      5.43
5 Oceania     74.3      3.80
```

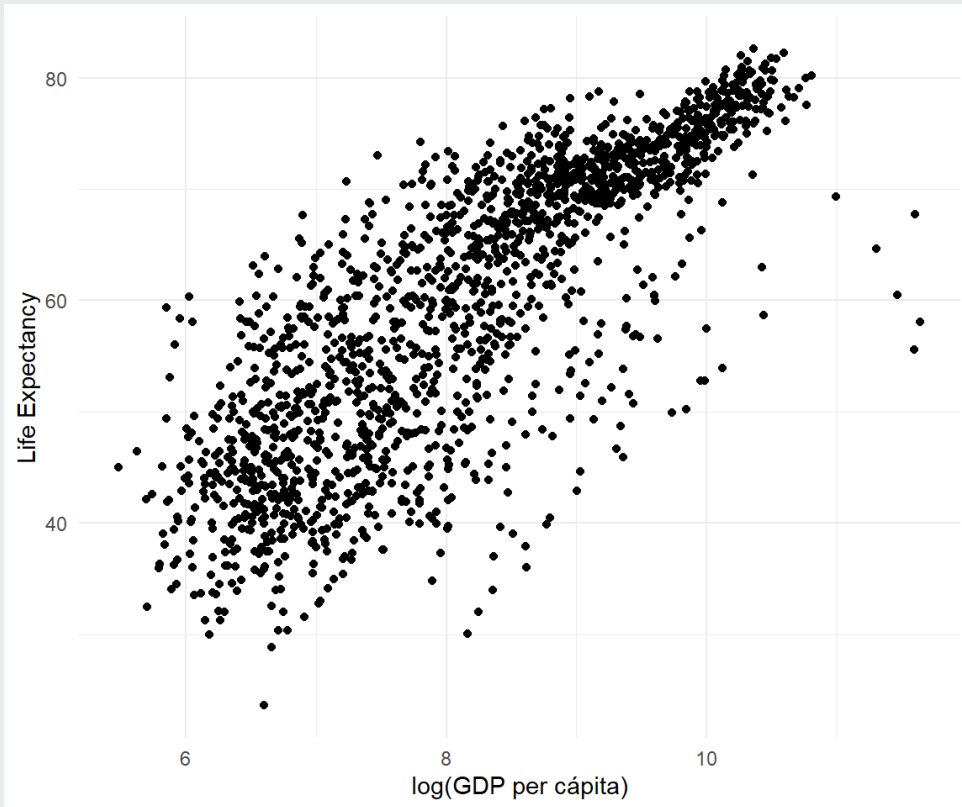


Ejercicio 5

1. Usando la función `mutate`, creen una nueva variable llamada `within_1SD` que sea igual a TRUE si el precio se encuentra dentro de una desviación estándar de la media, y FALSE en caso contrario. Es decir, la variable debe ser TRUE si el precio está en el rango de $[media - 1 * \text{desviación estándar}, media + 1 * \text{desviación estándar}]$, y FALSE si no está en este rango.
2. Calculen la media de `within_1SD`. Recuerden que el promedio de una variable que toma valores 0 o 1 corresponde a la proporción de observaciones que cumplen la condición. ¿La proporción obtenida se aproxima al 68%?

Medidas de relación entre dos variables

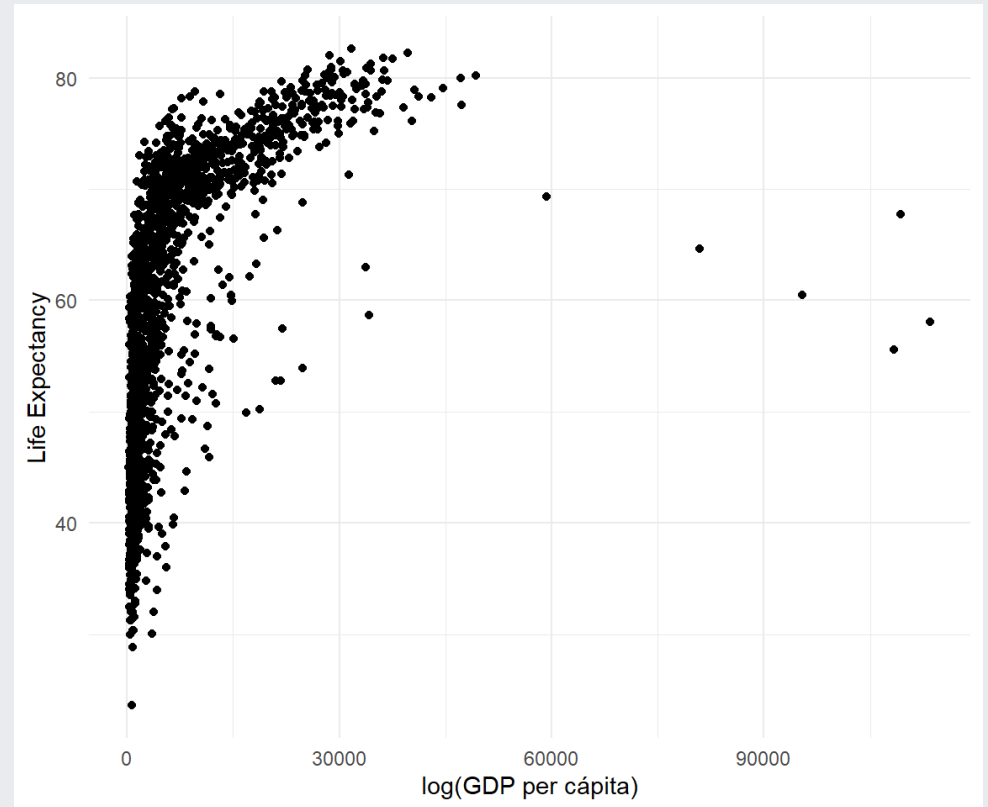
Diagramas de dispersión



- Los gráficos de dispersión nos ayudan a visualizar y examinar la relación entre dos variables numéricas
- Discutiremos dos medidas cuantitativas de esta relación:
 1. la covarianza
 2. la correlación

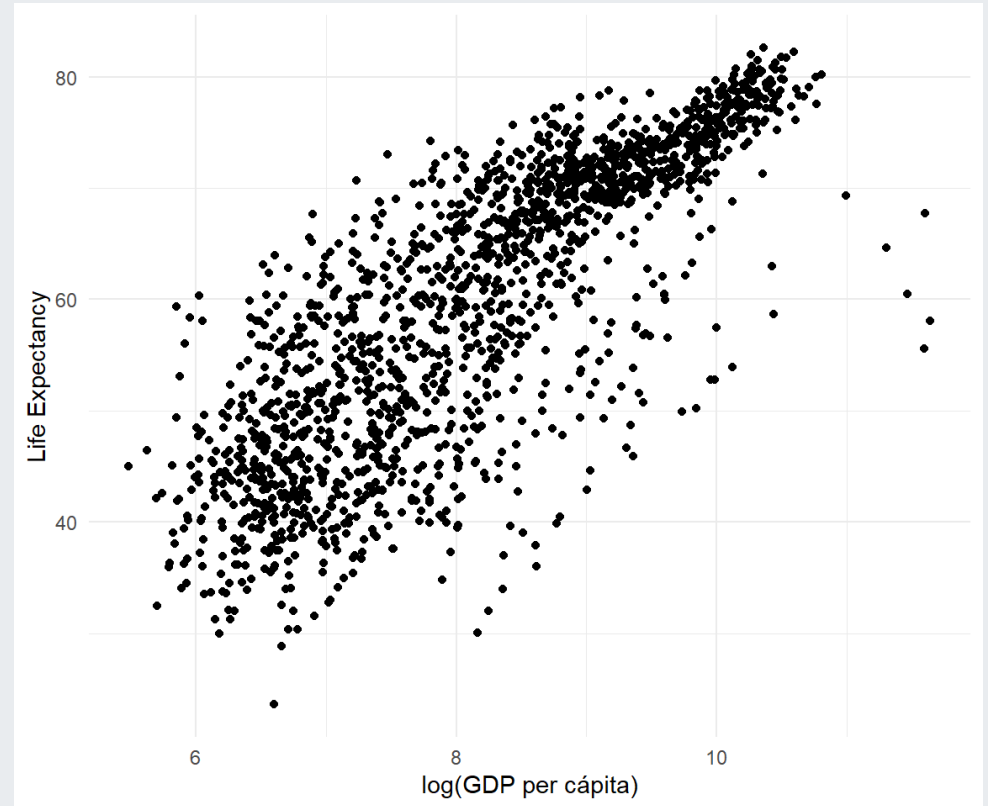
El gráfico de dispersión en R

```
1 ggplot(gapminder) +  
2   geom_point(aes(x=gdpPercap, y=lifeExp)) +  
3   labs(x = "GDP per cápita",  
4        y = "Life Expectancy") +  
5   theme_minimal()
```



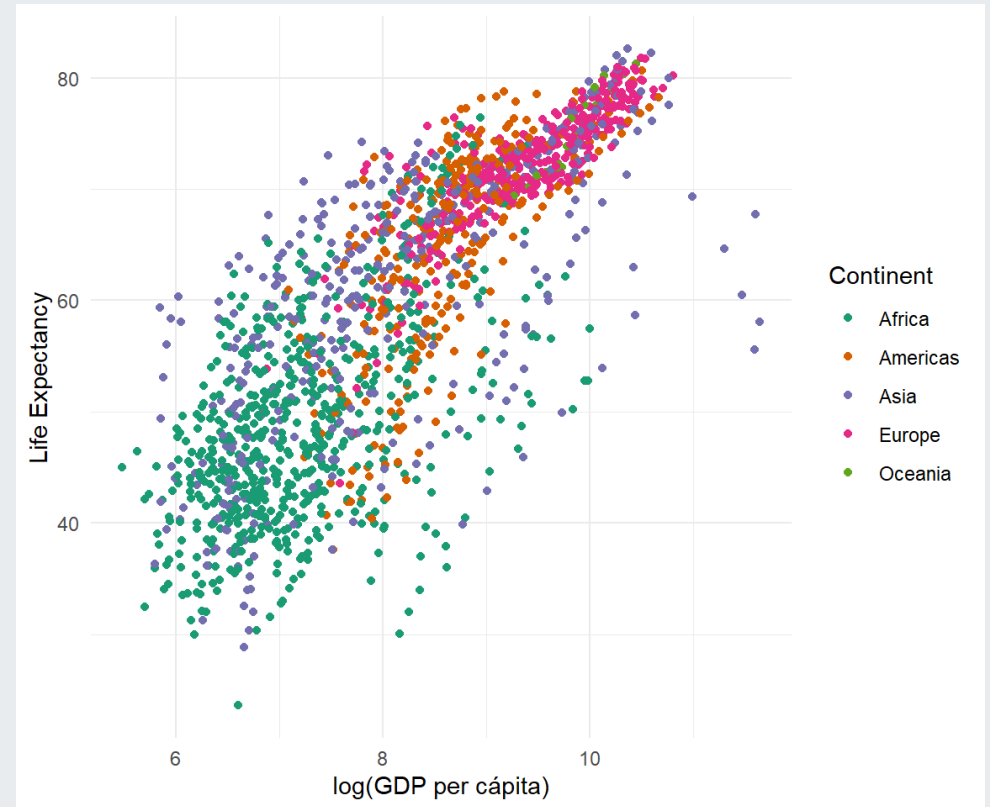
El gráfico de dispersión en R

```
1 ggplot(gapminder) +  
2   geom_point(aes(x=log(gdpPercap),  
3     labs(x = "log(GDP per cápita)",  
4       y = "Life Expectancy")) +  
5   theme_minimal()
```



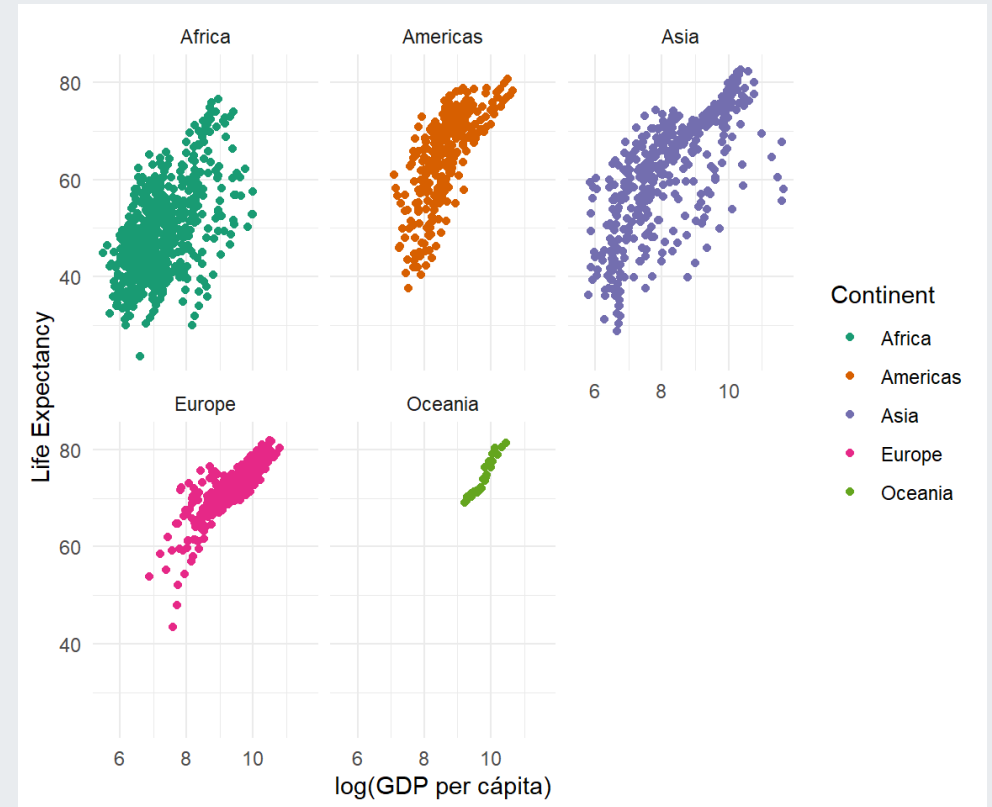
El gráfico de dispersión en R

```
1 ggplot(gapminder) +  
2   geom_point(aes(x=log(gdpPercap  
3     labs(x = "log(GDP per cápita)"  
4       y = "Life Expectancy") +  
5     scale_color_brewer(name = "Con  
6     theme_minimal()
```



El gráfico de dispersión en R

```
1 ggplot(gapminder) +  
2   geom_point(aes(x=log(gdpPercap  
3     labs(x = "log(GDP per cápita)"  
4       y = "Life Expectancy") +  
5     scale_color_brewer(name = "Con  
6     theme_minimal() +  
7     facet_wrap(~continent, ncol =
```



La opción `facet_wrap()` separa el gráfico por grupos. Este [link](#) explica cómo usarla y da ejemplos.

Covarianza

- La covarianza mide qué tan fuerte es la relación (lineal) de *dos variables numéricas*.
- La covarianza se calcula de la siguiente manera:
- Solamente mide la “dirección” de la relación.

Interpretación de la covarianza

- La covarianza entre dos variables:
 - y se mueven en la misma dirección.
 - y se mueven en dirección opuesta.
 - y son independientes.
- El defecto de la covarianza es que no indica la intensidad de la relación entre las dos variables

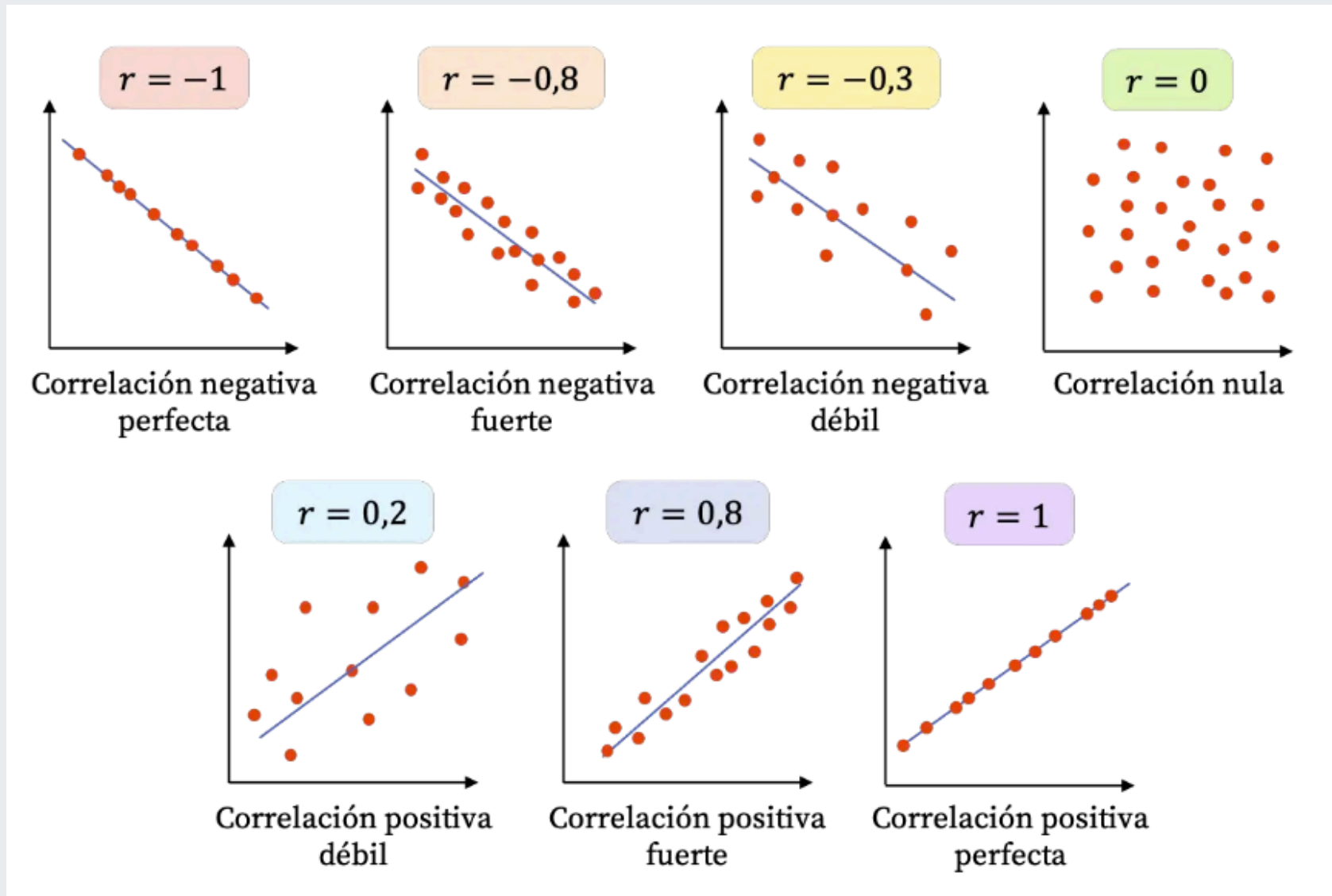
Correlación

- La correlación mide la dirección y la fuerza de la relación lineal entre dos variables numéricas.

Características de la correlación

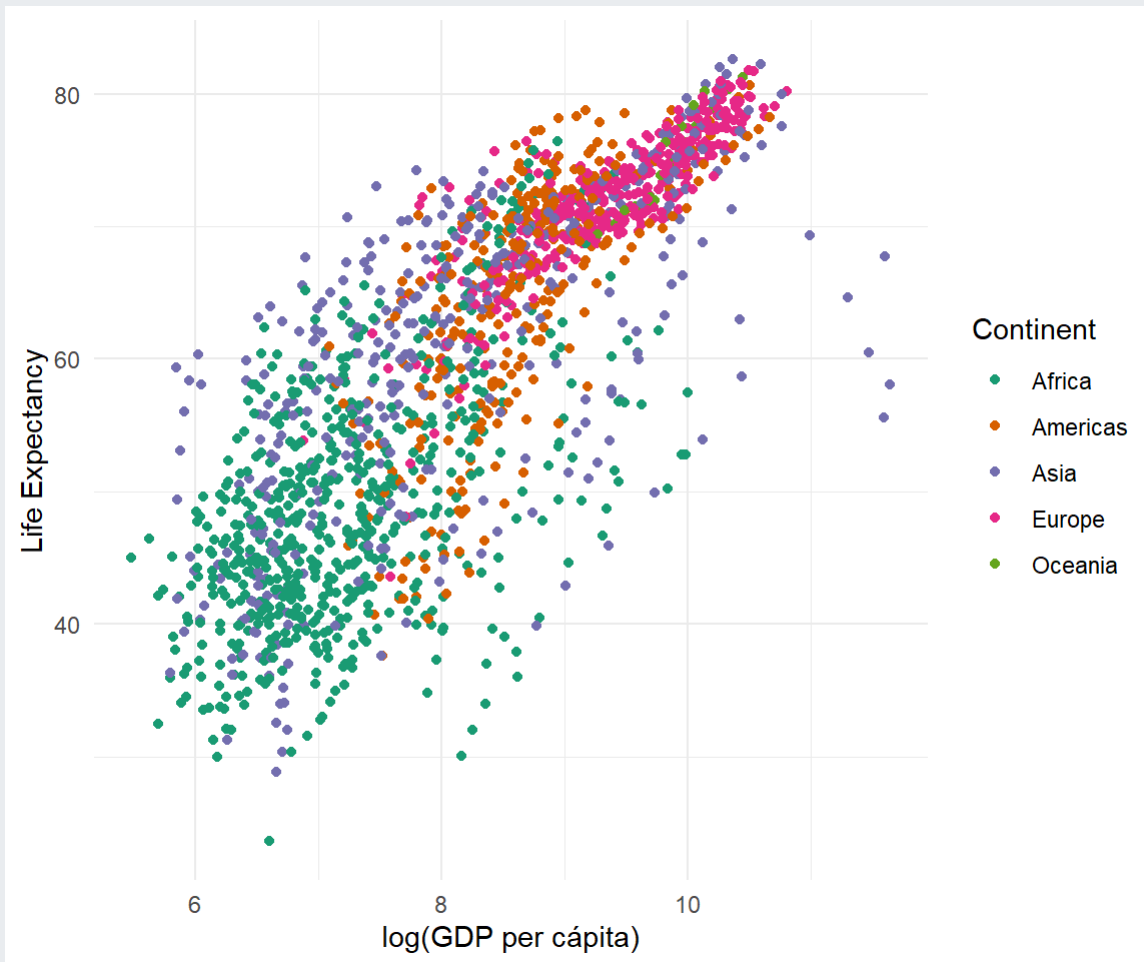
-
- no cambia cuando se cambian las unidades de medida de , o ambas.
- no tiene unidad de medida
- Más cerca a -1, más fuerte la relación lineal negativa.
- Más cerca a 1, más fuerte la relación lineal positiva.
- Más cerca a 0, más débil la relación lineal.

Valores y gráficos de la correlación



La correlación en R

La correlación se calcula con la función `cor`:

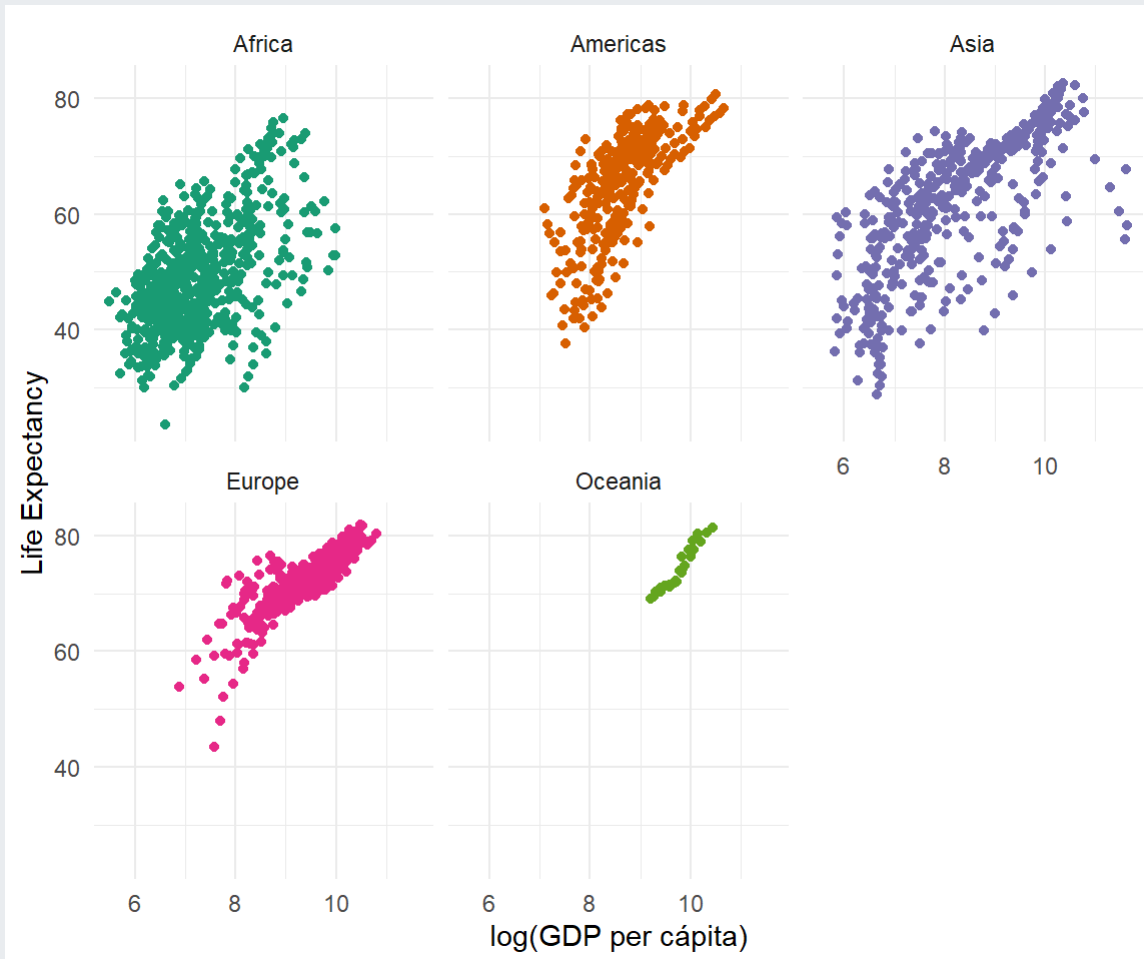


```
1 gapminder |>  
2   group_by(continent) |>  
3   summarise(  
4     corr = cor(lifeExp, gd  
5     )
```

```
# A tibble: 5 × 2  
  continent    corr  
  <fct>      <dbl>  
1 Africa     0.426  
2 Americas   0.558  
3 Asia       0.382  
4 Europe     0.781  
5 Oceania    0.956
```

La correlación en R

En ocasiones es mejor separar los gráficos para cada categoría:



```
1 gapminder |>
2   group_by(continent) |>
3   summarise(
4     corr = cor(lifeExp, gd
5   )
```

A tibble: 5 × 2

continent	corr
<fct>	<dbl>
1 Africa	0.426
2 Americas	0.558
3 Asia	0.382
4 Europe	0.781
5 Oceania	0.956

Ejercicio 6

1. Grafiquen la dispersión entre el precio y el número de reseñas (reviews). ¿Qué tipo de relación visualizan entre estas dos variables?
2. Calculen la correlación entre el precio y el número de reseñas. ¿El valor obtenido tiene sentido con lo que observan en la gráfica del punto anterior?
3. En ocasiones, la relación entre dos variables no es clara a primera vista. La función `geom_smooth()` facilita visualizar posibles relaciones. Por ejemplo:

```
1 ggplot(newdata) +  
2   geom_smooth(aes(y=number_of_reviews, x=price))
```

¿Qué historia nos cuenta esta nueva gráfica? ¿Tiene sentido para ustedes la relación observada entre las variables?

