

# Clase 5

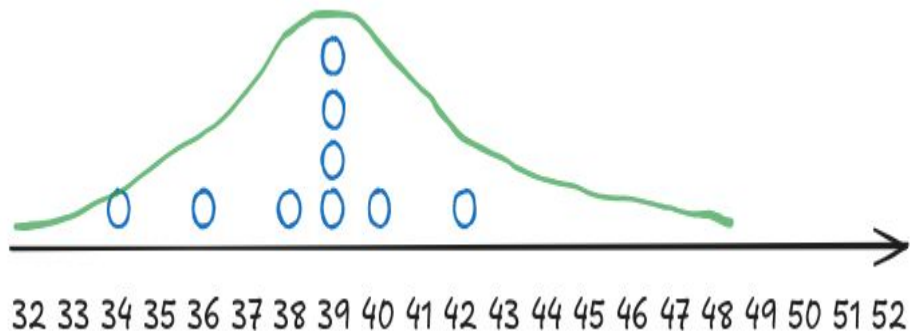
**unab**

VISUALIZACIÓN DE LA INFORMACIÓN 2024

**Visualizar distribuciones**

# ¿Que es una *distribución*?

```
# A tibble: 344 × 3
  especie isla    largo_pico_mm
  <fct>   <fct>      <dbl>
1 Adelia Torgersen 39.1
2 Adelia Torgersen 39.5
3 Adelia Torgersen 40.3
4 Adelia Torgersen NA
5 Adelia Torgersen 36.7
6 Adelia Torgersen 39.3
7 Adelia Torgersen 38.9
8 Adelia Torgersen 39.2
9 Adelia Torgersen 34.1
10 Adelia Torgersen 42
# i 334 more rows
```



Es una representación de los datos que se recopilaron de un muestreo, o del total de la población. Pueden ser datos **discretos** o **continuos**. Y pueden tener más de una dimensión.

Es una forma de simplificar la organización de grandes volúmenes de datos.

# ¿Qué datos representan a una *distribución*?

## Media, Mediana, Moda y Rango

Primero, ordena los números de menor a mayor.

Ejemplo: 3, 5, 5, 6, 8, 10, 12

### Media

el promedio de los números

1. Suma los números.
2. Divide entre la cantidad de números en el conjunto.

$$3+5+5+6+8+10+12=49$$

$$49 / 7 = 7$$

La media es 7

### Mediana

el número de la mitad

1. Coloca los números en orden de valor y encuentra el número del medio
- \*Si hay dos números en el medio, la mediana es la media de los dos números.

3, 5, 5, 6, 8, 10, 12

La mediana es 6

### Moda

el número que aparece con más frecuencia

1. Halla el número que repite más en el conjunto de datos (puede haber más que un solo número).
- \*Hay dos 5s y uno de cada otro número.

3, 5, 5, 6, 8, 10, 12

La moda es 5

### Rango

La diferencia entre el máximo y el mínimo

1. Resta el mínimo (número menor) del máximo (número mayor)

3, 5, 5, 6, 8, 10, 12

$$12 - 3 = 9$$

El rango es 9

## MEDIDAS DE DISPERSIÓN

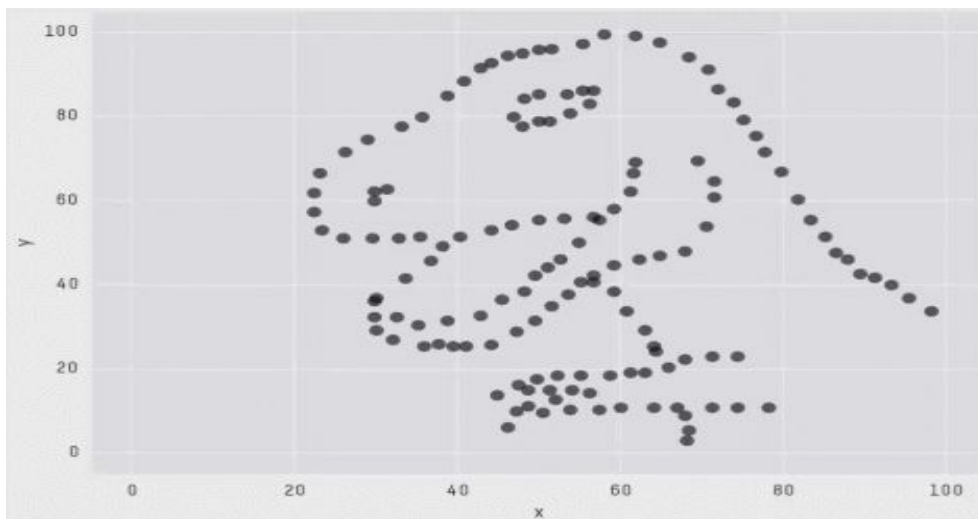
VARIANZA	DESVIACIÓN ESTÁNDAR
$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$	$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2}{N}}$

- **X** → Variable sobre la que se pretenden calcular la varianza.
- **$x_i$**  → Observación número  $i$  de la variable  $X$ .  $i$  puede tomar valores entre 1 y  $n$ .
- **N** → Número de observaciones.
- **$\bar{x}$**  → Es la media de la variable  $X$ .

RANGO ESTADÍSTICO	COEFICIENTE DE VARIACIÓN
$R = Máx_x - Mín_x$	$CV = \frac{\sigma_x}{ \bar{X} }$
<ul style="list-style-type: none"><li>• <b>R</b> → Es el rango.</li><li>• <b>Máx</b> → Es el valor máximo de la muestra o población.</li><li>• <b>Mín</b> → Es el valor mínimo de la muestra o población estadística.</li><li>• <b>x</b> → Es la variable sobre la que se pretende calcular esta medida.</li></ul>	<ul style="list-style-type: none"><li>• <b>X</b> → Variable sobre la que se pretenden calcular la varianza.</li><li>• <b><math>\sigma_x</math></b> → Desviación típica de la variable <math>X</math>.</li><li>• <b><math> \bar{x} </math></b> → Es la media de la variable <math>X</math> en valor absoluto con <math>\bar{x} \neq 0</math>.</li></ul>

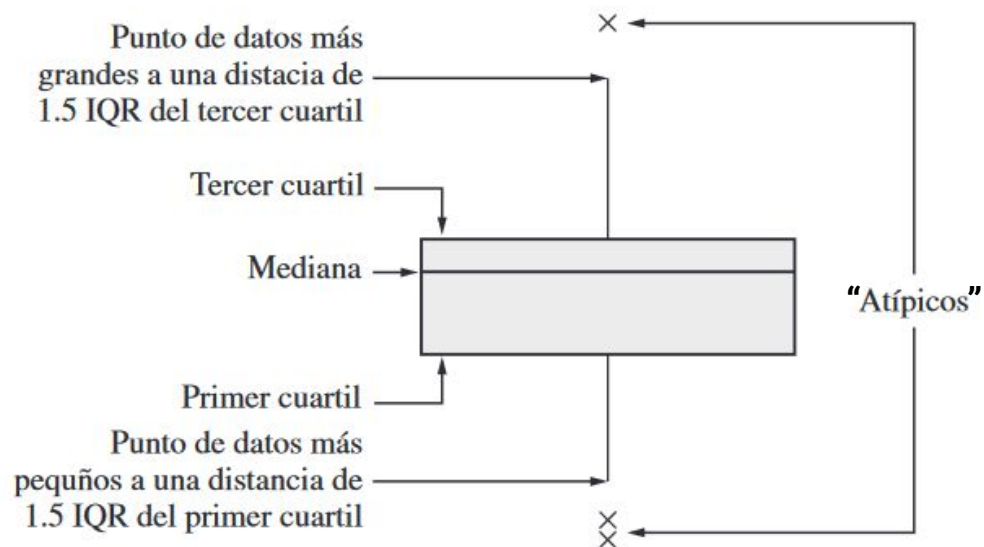
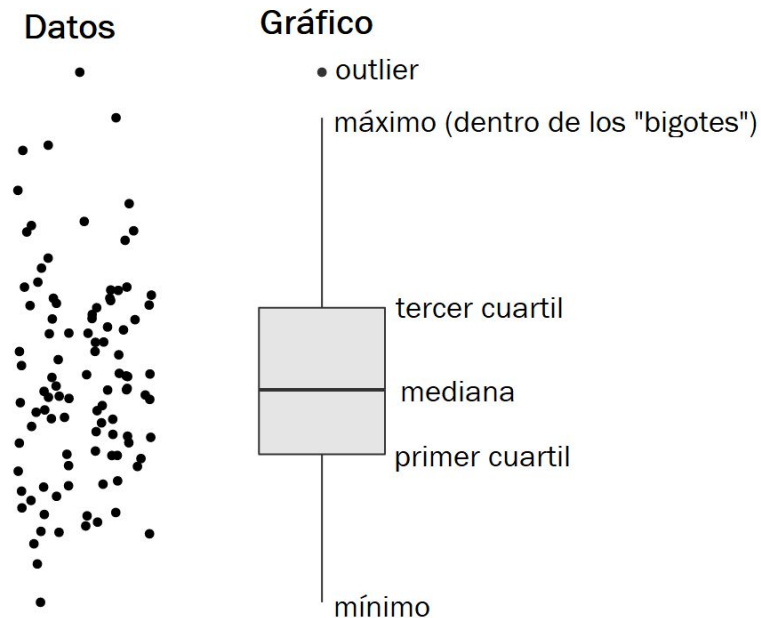
# La importancia de la visualización

datasauRus



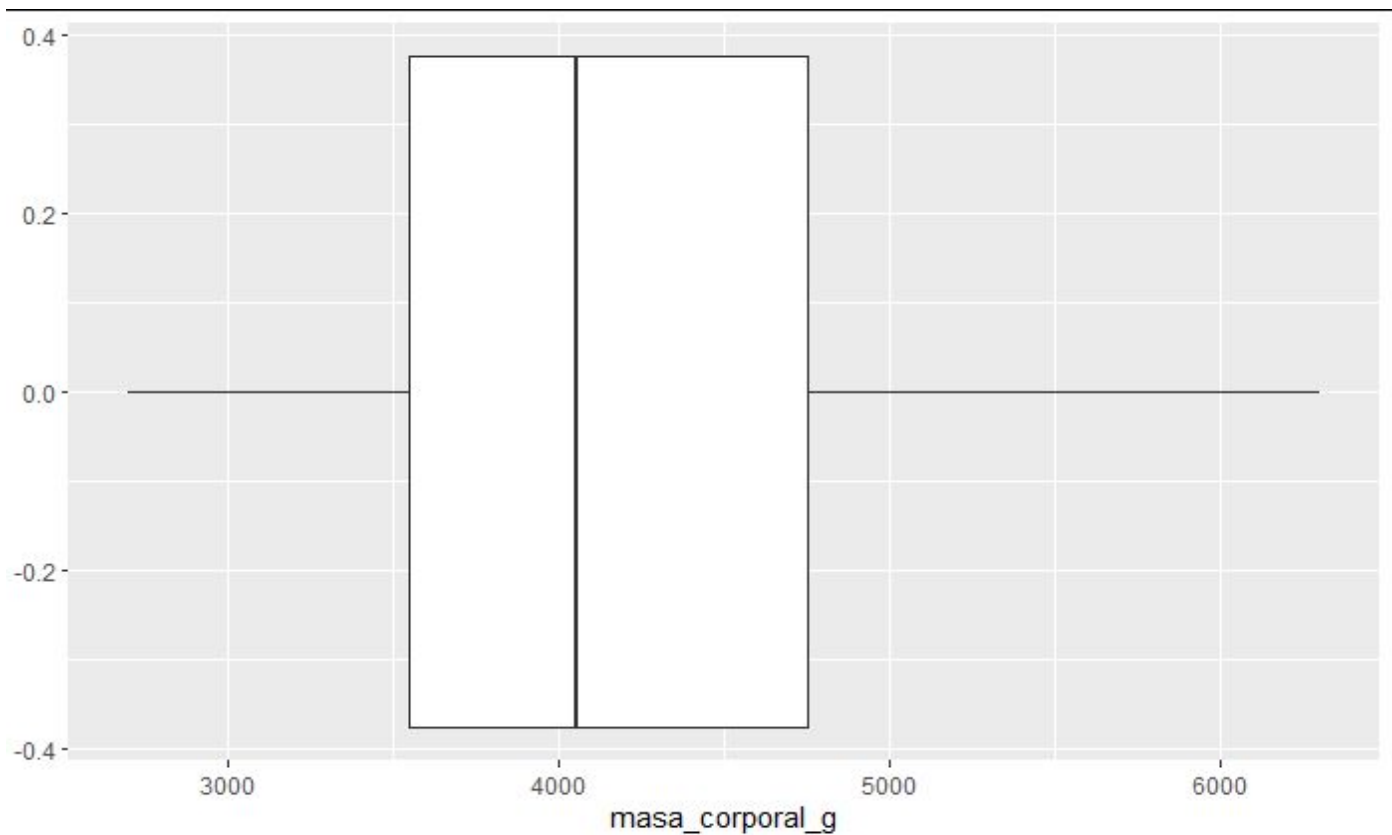
```
X Mean: 54.2659224
Y Mean: 47.8313999
X SD   : 16.7649829
Y SD   : 26.9342120
Corr.  : -0.0642526
```

# geom\_boxplot() - ¿cómo leerlos?

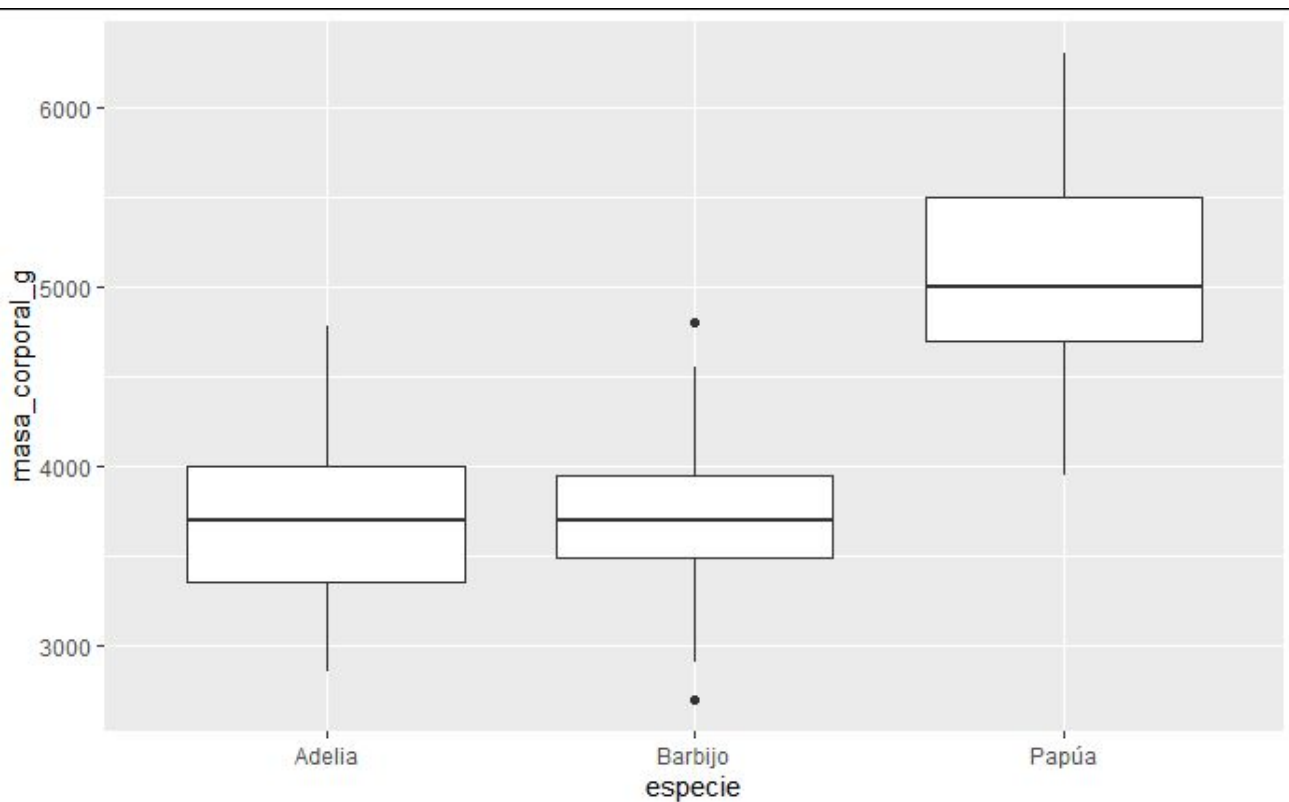


**FIGURA 1.14** Anatomía de un diagrama de caja.

# geom\_boxplot()

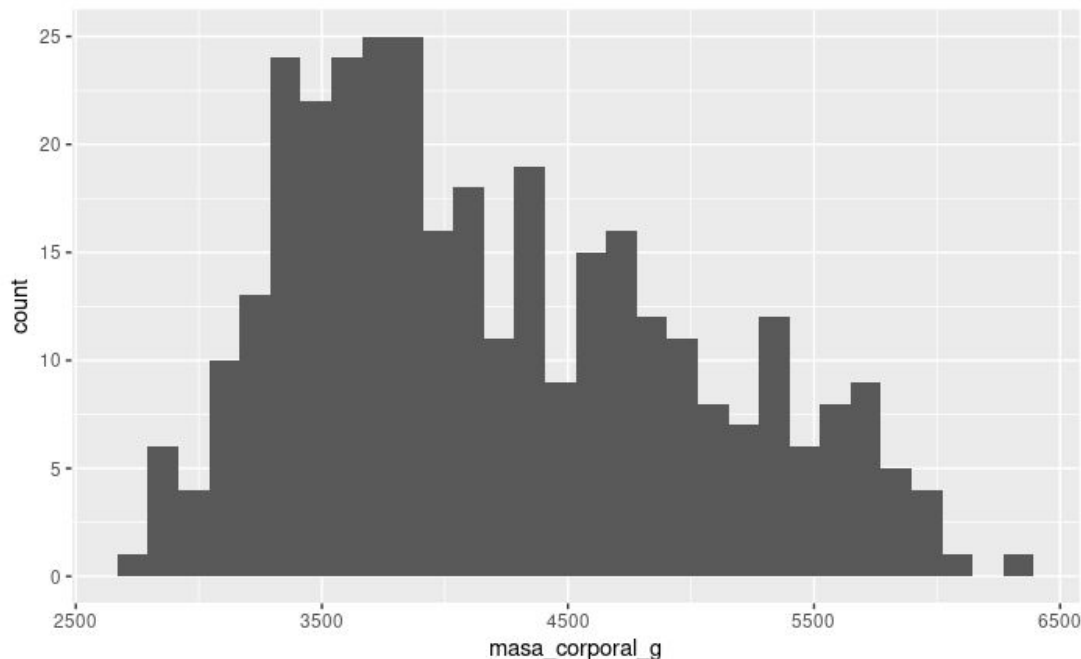


# geom\_boxplot() - comparando distribuciones





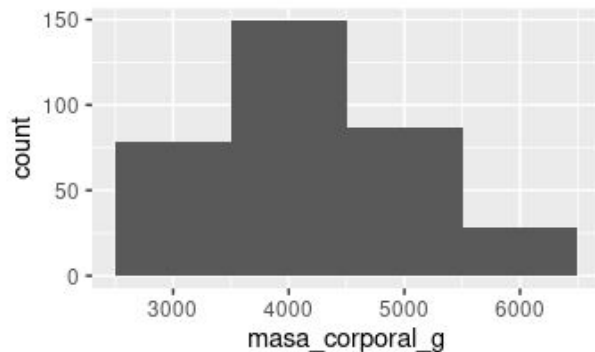
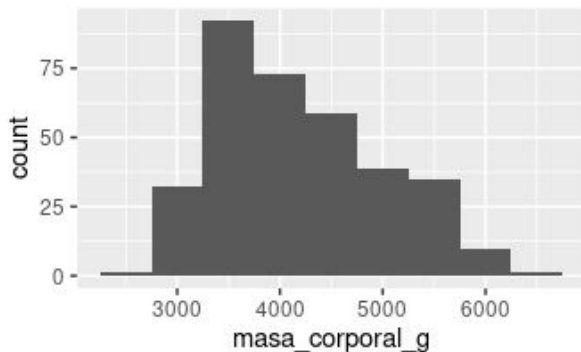
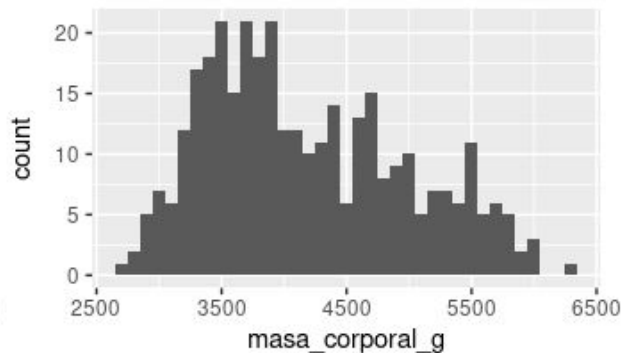
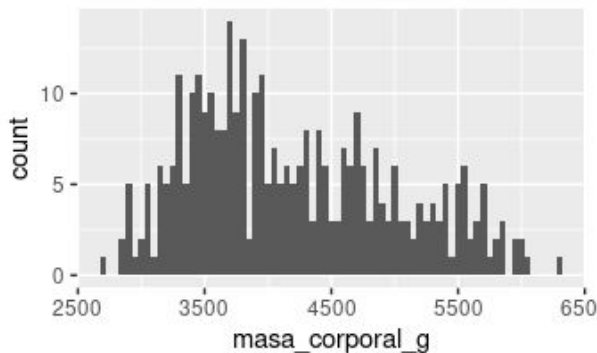
# geom\_histogram()



→ Se definen intervalos y se cuentan los casos.

→ Dependen del valor del ancho del intervalo (binwidth) o la cantidad de intervalos (bin).

# Diferente cantidad de intervalos

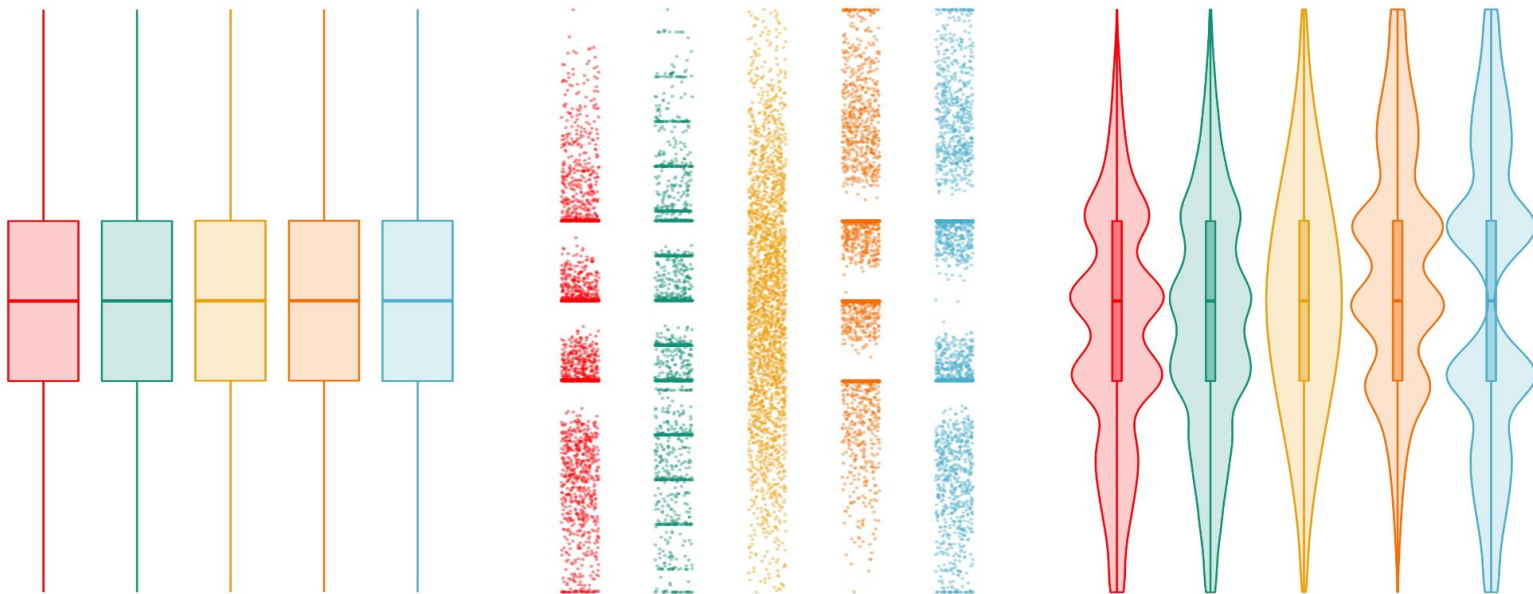


→ Por defecto  $\text{bin} = 30$  (y el  $\text{binwidth} = \text{rango}/\text{bin}$ ).

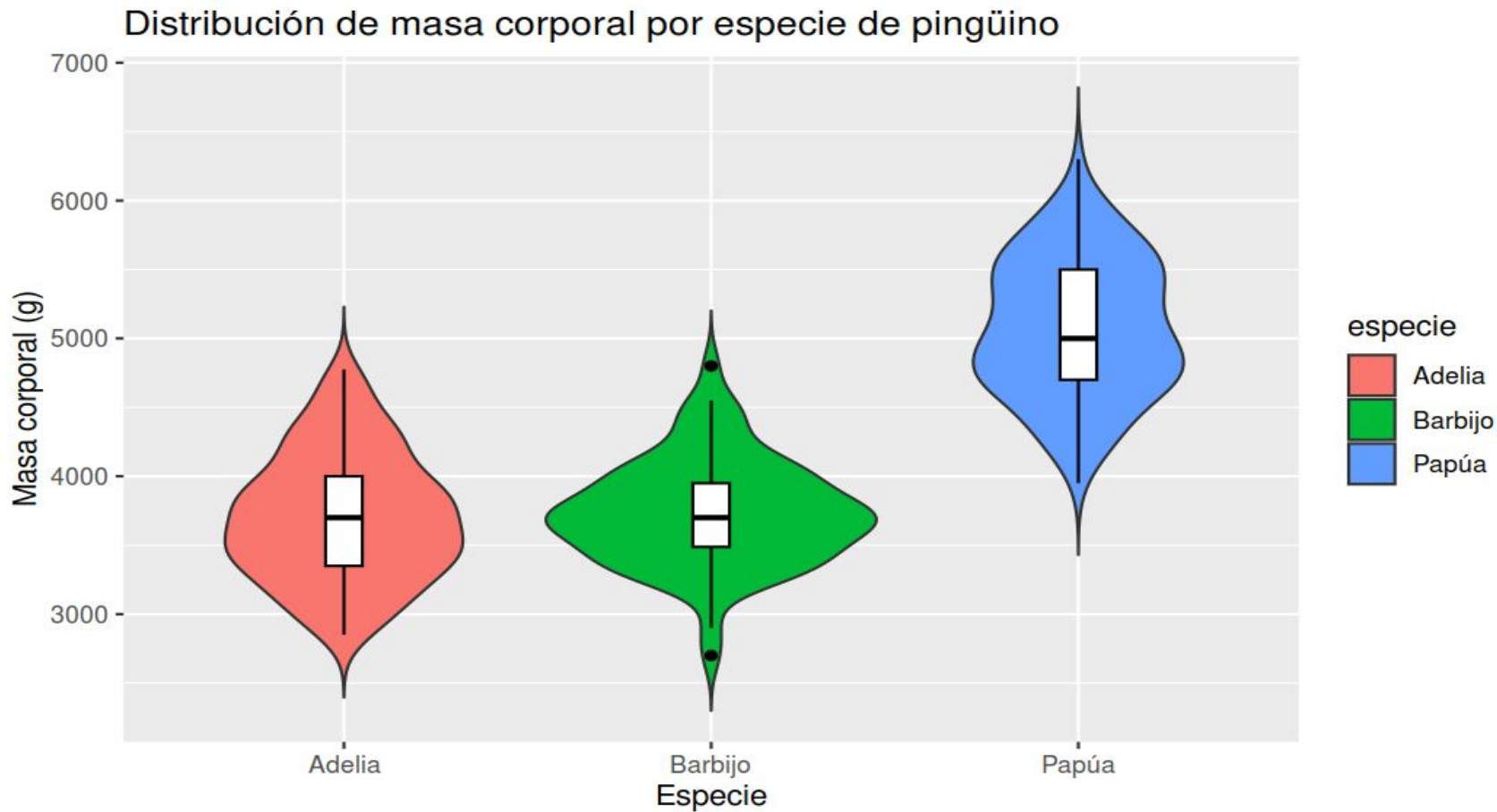
# Sobre elegir gráficos

## Identical boxplots, different distributions

Boxplots are great. They show medians and ranges and enable comparison of different groups. However, boxplots can be misleading. Different datasets can have the same descriptive statistics (left), but quite different underlying distributions (middle). Therefore, it is crucial to visualize the distribution in addition to descriptive statistics. Violin plots with integrated boxplots are great for this.

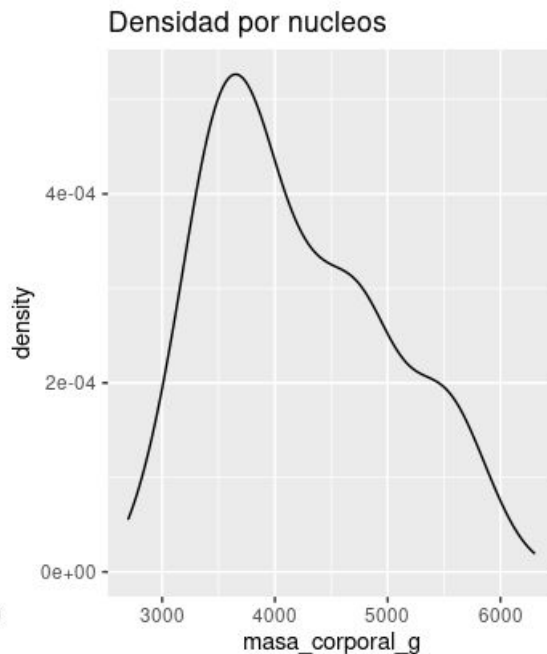
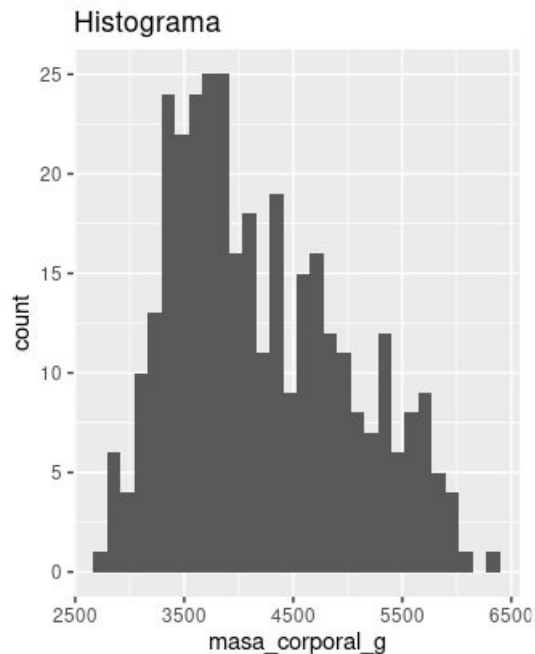


# Combinando información



is.

# geom\_density()

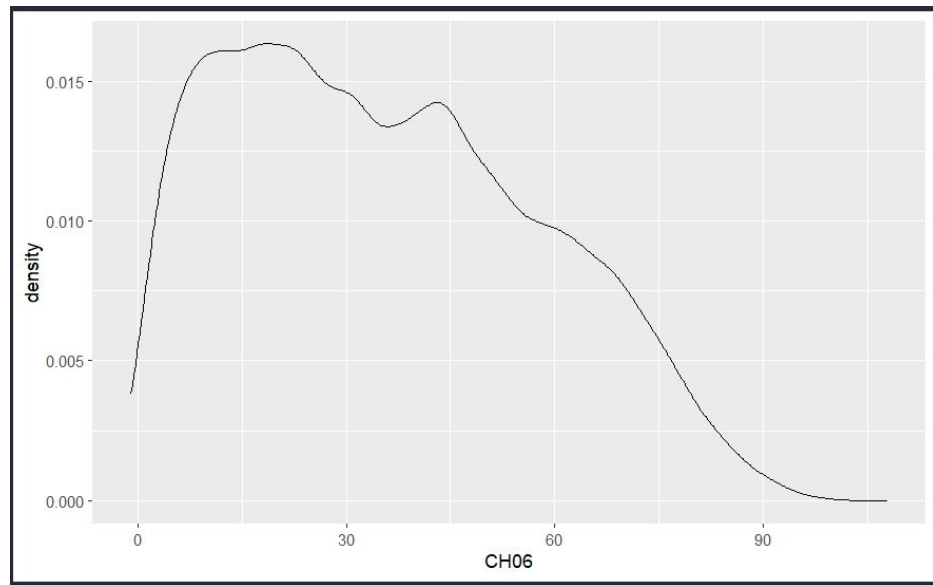
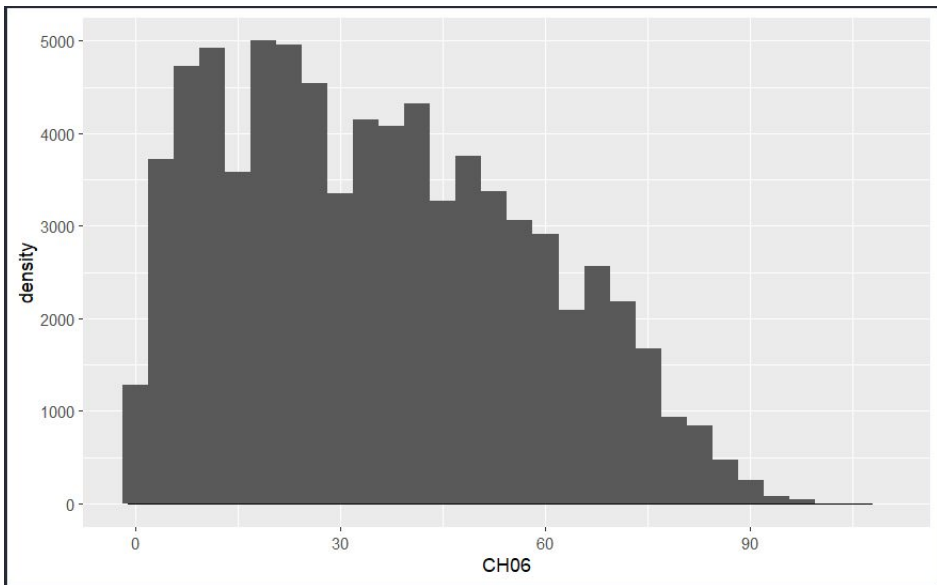


→ El histograma cuenta repeticiones, el gráfico de densidad calcula proporciones. El área bajo la curva es igual a 1.

# geom\_density()

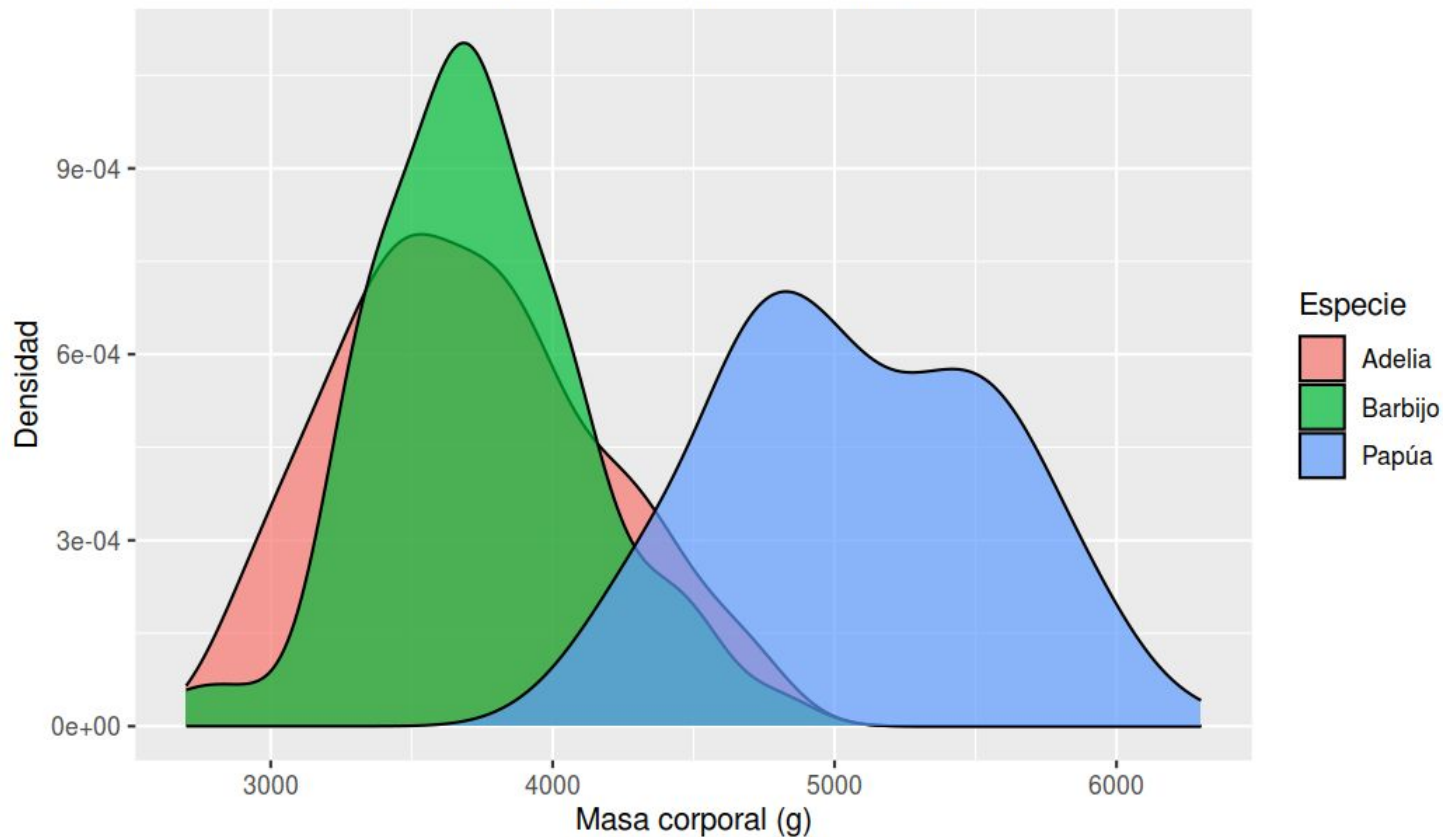
variable EDAD (CH06)

Encuesta Permanente de Hogares

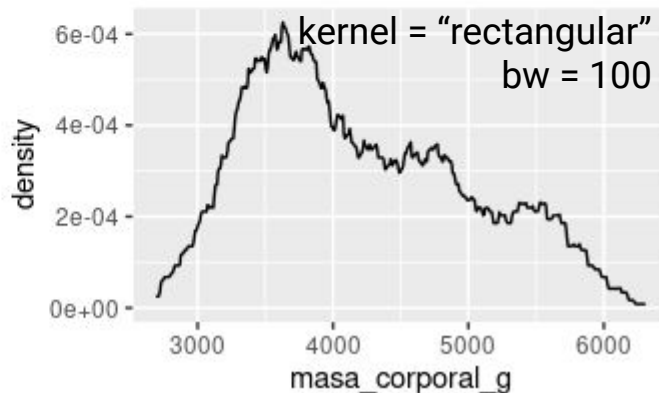
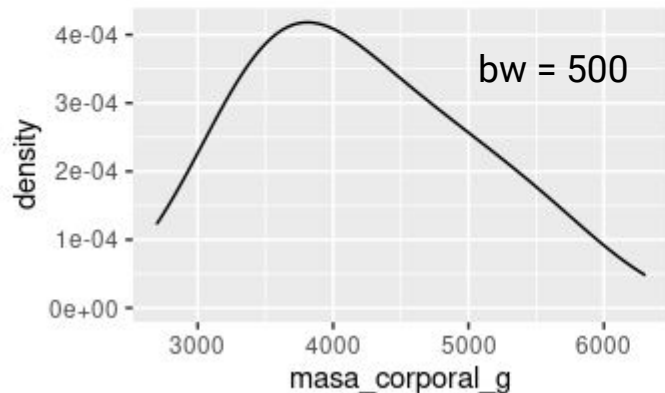
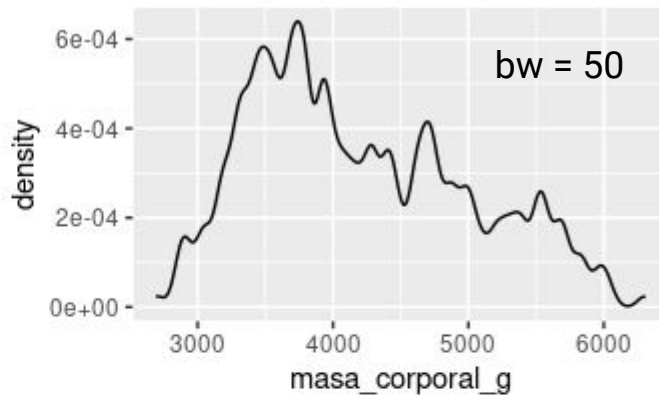
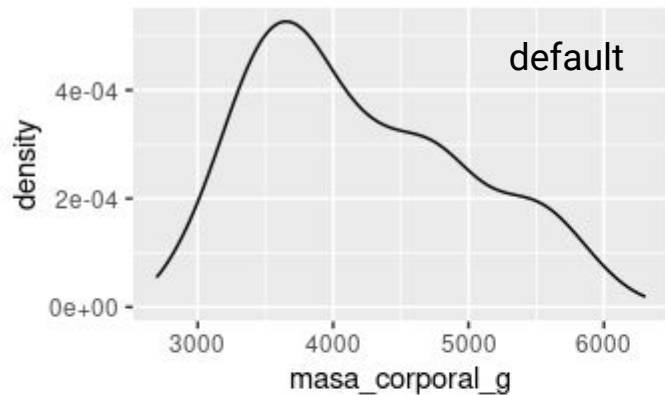


# geom\_density() con los pingüinos

Densidad de masa corporal por especie de pingüino

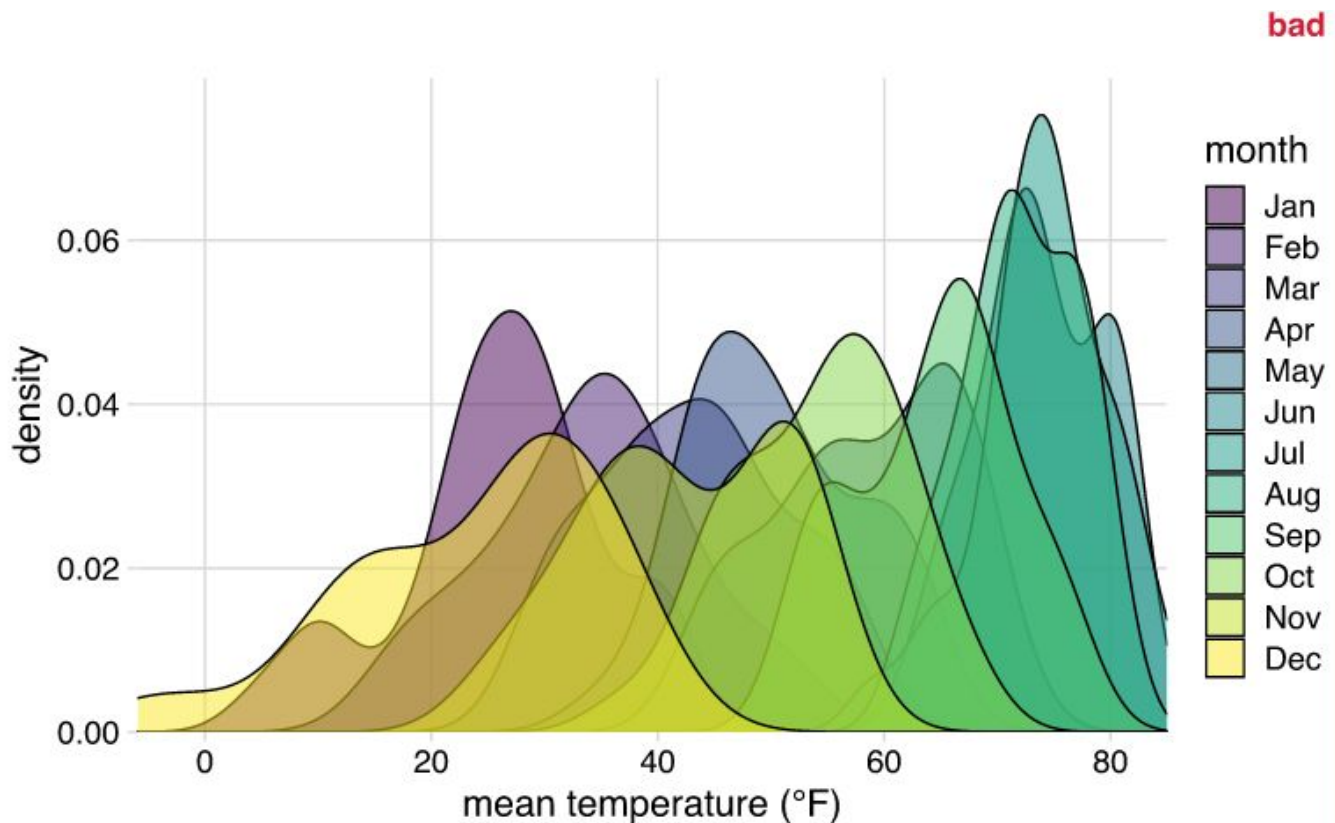


# También depende de parámetros



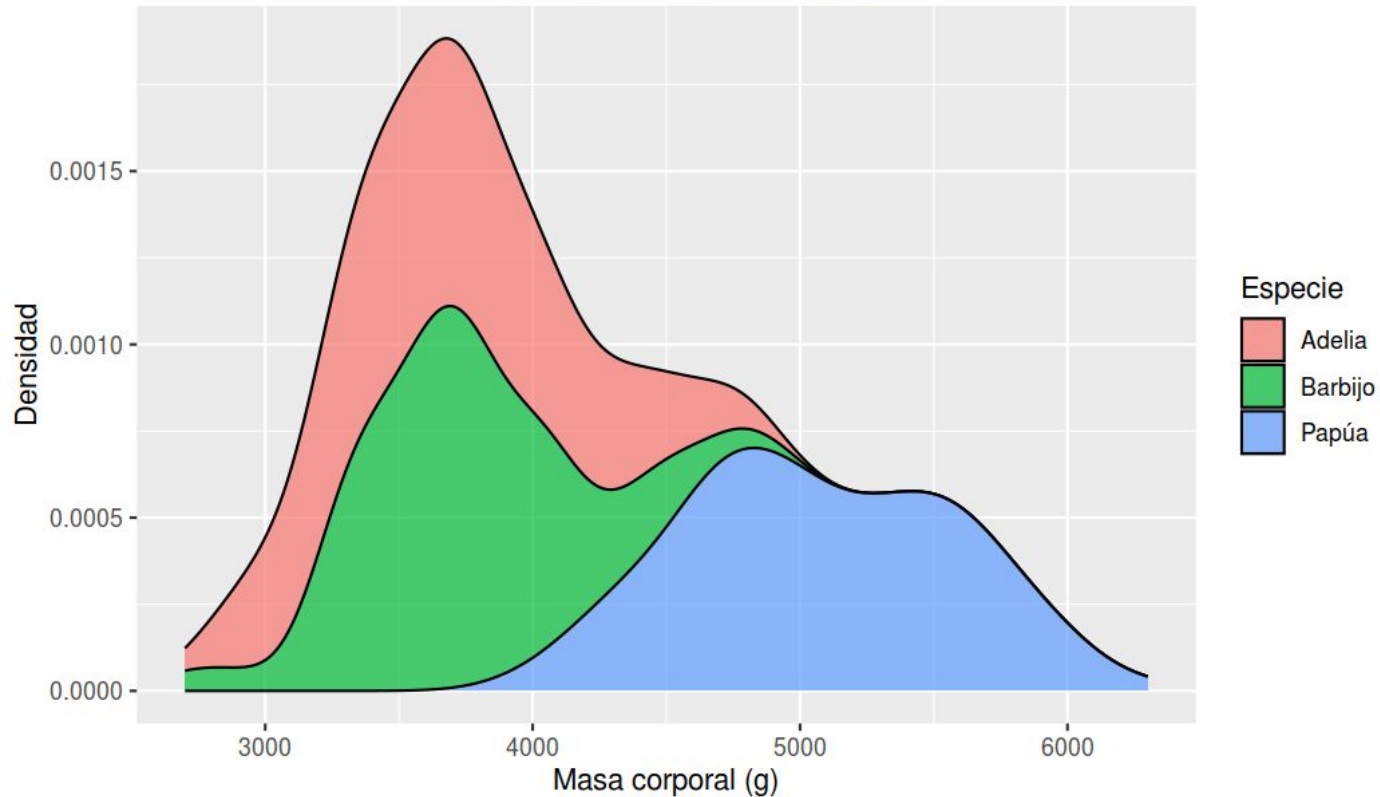


# Cuando son demasiados datos...

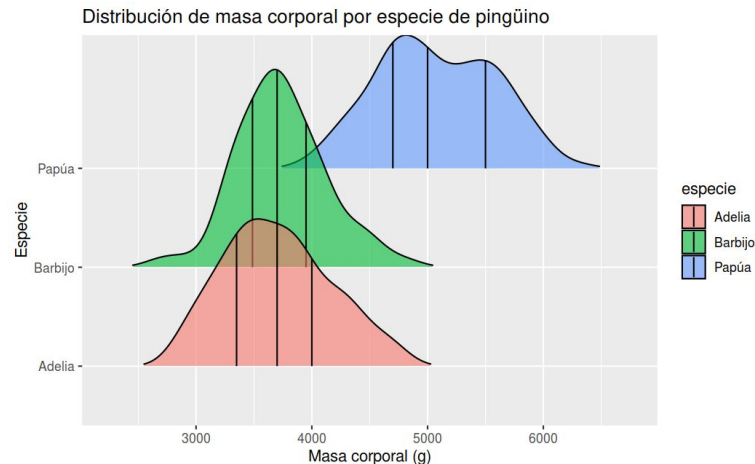
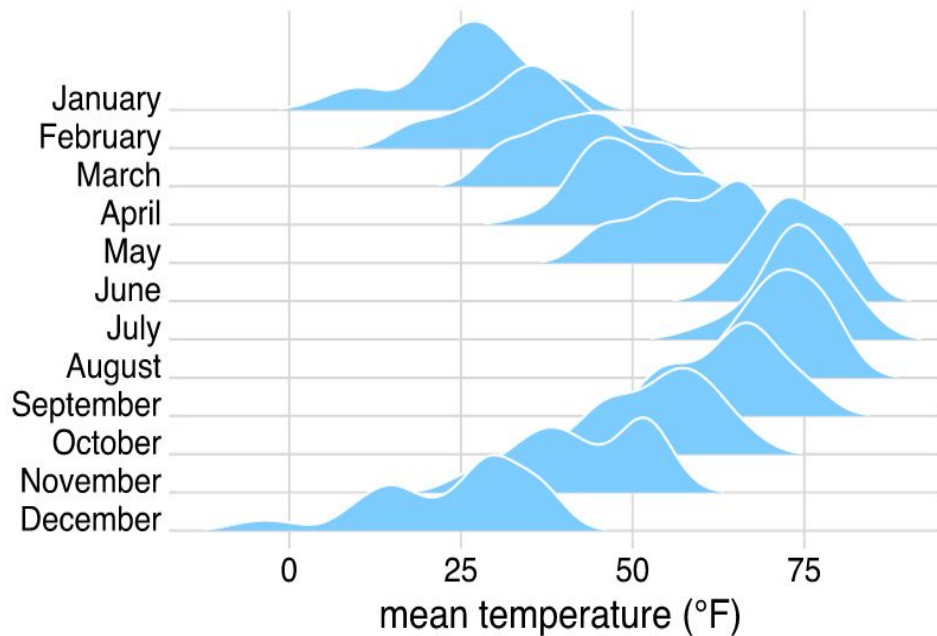


# Esto menos...

Densidad de masa corporal por especie de pingüino



# geom\_density() - comparar distribuciones



→ Gráficos de densidad “apilados” se llaman ridgelines (paquete ggridges)

# **Pausa**

10 minutos

**No se  
desconecten  
pero  
retirensé de las  
pantallas.**

# Principios de la visualización de datos

Según Kai Xu, profesor asociado de la Universidad de Nottingham

<https://www.youtube.com/watch?v=qQ9Wu1IxsYw> 👉

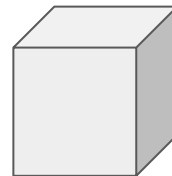
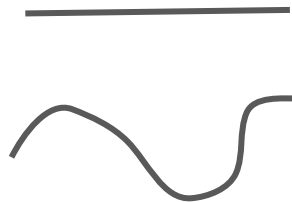
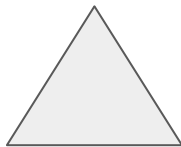
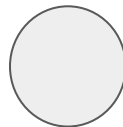
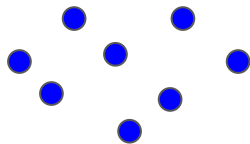


**Graphics = Marks + Channels**

# Marks (marcas)

- 0D marca: punto;
- 1D marca: linea;
- 2D marca: area;
- 3D marca : volumen

Representan elementos  
en los datos (filas en una  
tabla)



# Channels (canales)

Representan  
**atributos** de  
los datos  
(columnas en  
una tabla)

→ Position

→ Horizontal



→ Vertical



→ Both



→ Color



→ Shape



→ Tilt



→ Size

→ Length



→ Area



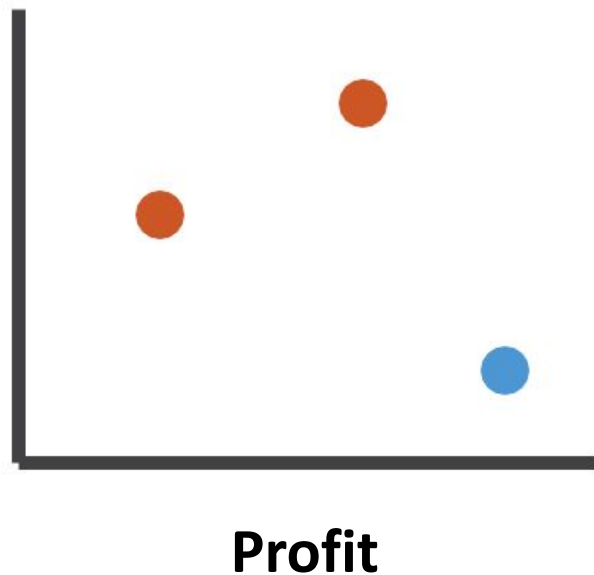
→ Volume





# Ejemplo

Sales      • Furniture    • IT



Marca: punto (0D)

Canales:

- Posición X: profit
- Posición Y: sales
- Color: type

## Canales de magnitud: para atributos con orden

### ➔ **Magnitude Channels: Ordered Attributes**

Position on common scale 

Position on unaligned scale 

Length (1D size) 

Tilt/angle 


Area (2D size) 

Depth (3D position) 

Color luminance 

Color saturation 

Curvature 

Volume (3D size) 

Same  
Same

Best  
Effectiveness  
Least

## Canales de identidad: para atributos categóricos

### ➔ **Identity Channels: Categorical Attributes**

Spatial region 

Color hue 

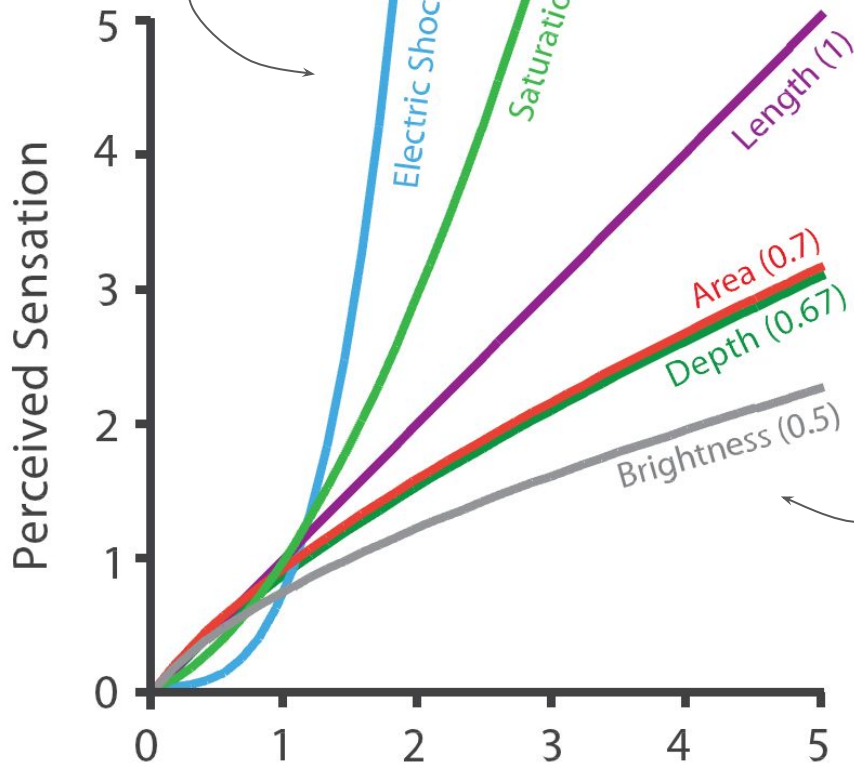
Motion 

Shape 

Ranking de la  
efectividad de  
los canales ;

Sensación percibida

Las personas  
sobreestiman  
cambios



Las personas  
subestiman  
cambios

Physical Intensity → Intensidad/cambio real

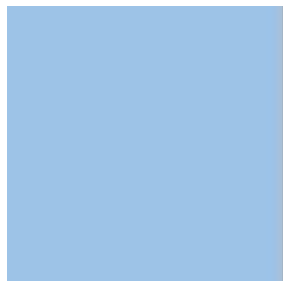
# Saturación, Largo, Área

A - 50% más

B - 100% más

C - 150% más

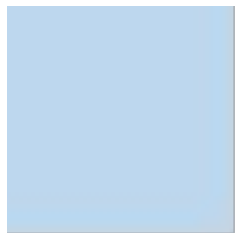
1:



2:



3:



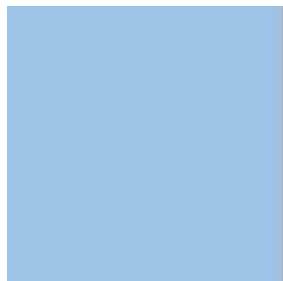
# Saturación, Largo, Área

A - 50% más

**B - 100% más**

C - 150% más

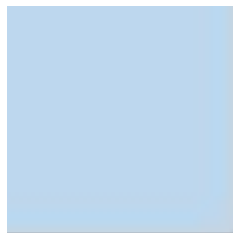
1:



2:

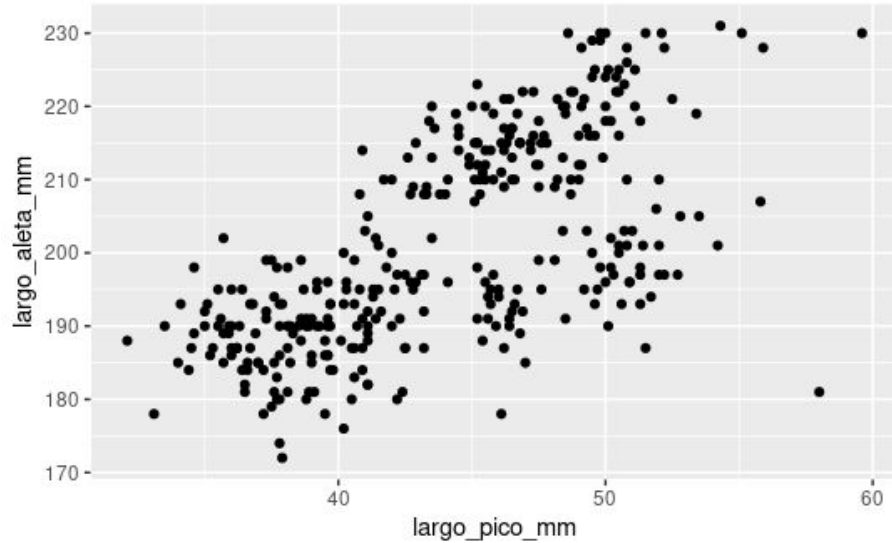


3:



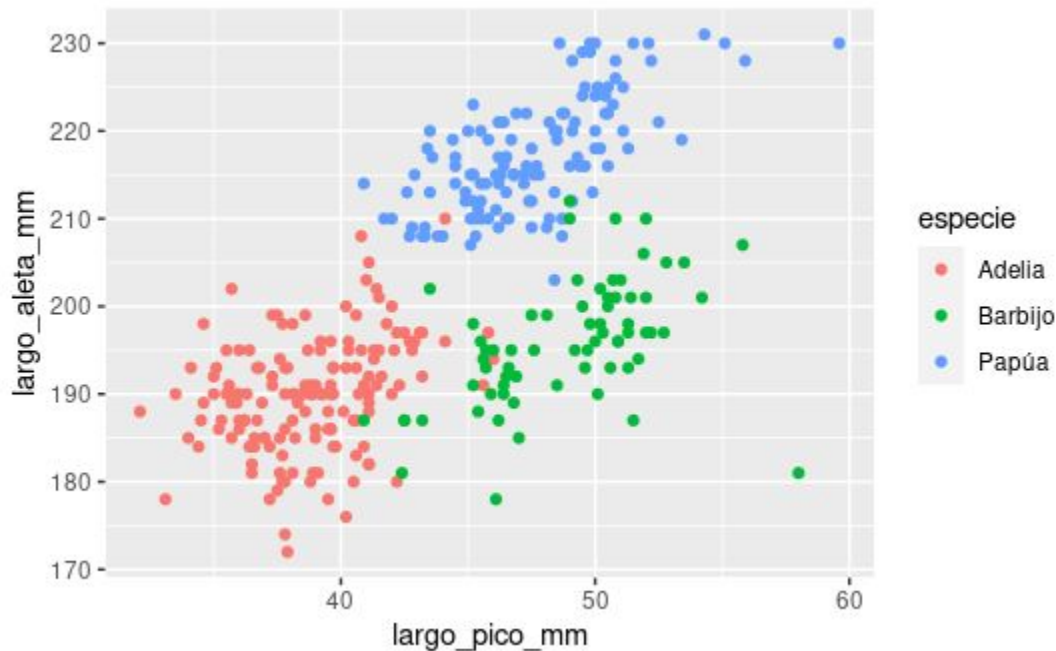
**Visualizar relaciones**

# geom\_point()



→ Mientras más larga la aleta, más largo el pico y viceversa.

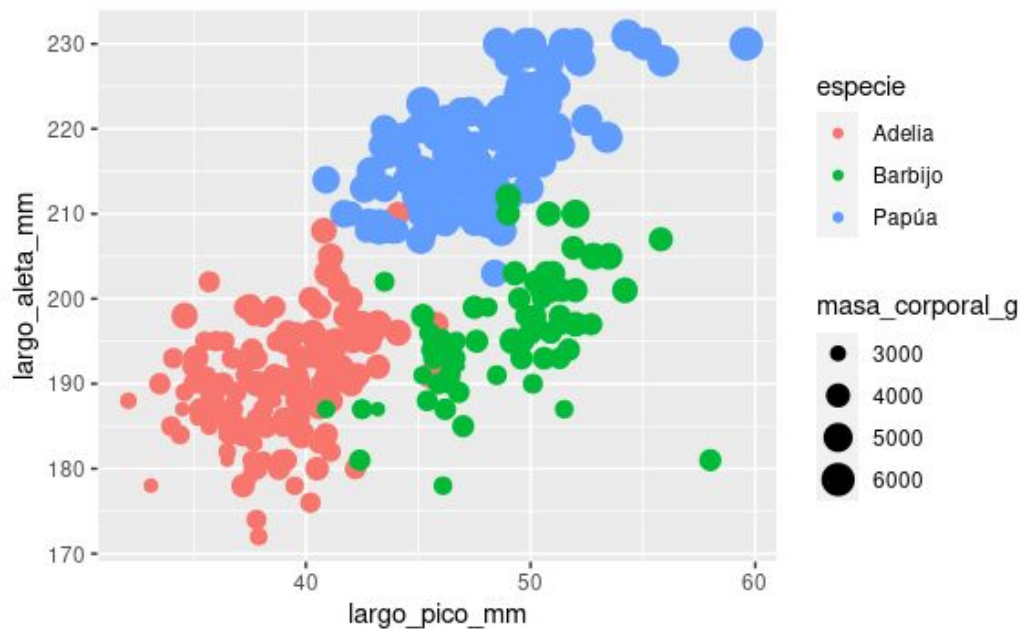
# Más de dos variables



Mapeamos la tercer variable a un elemento estético de los puntos



# Más de tres variables?



La cuarta variable, está mapeada al tamaño de los puntos

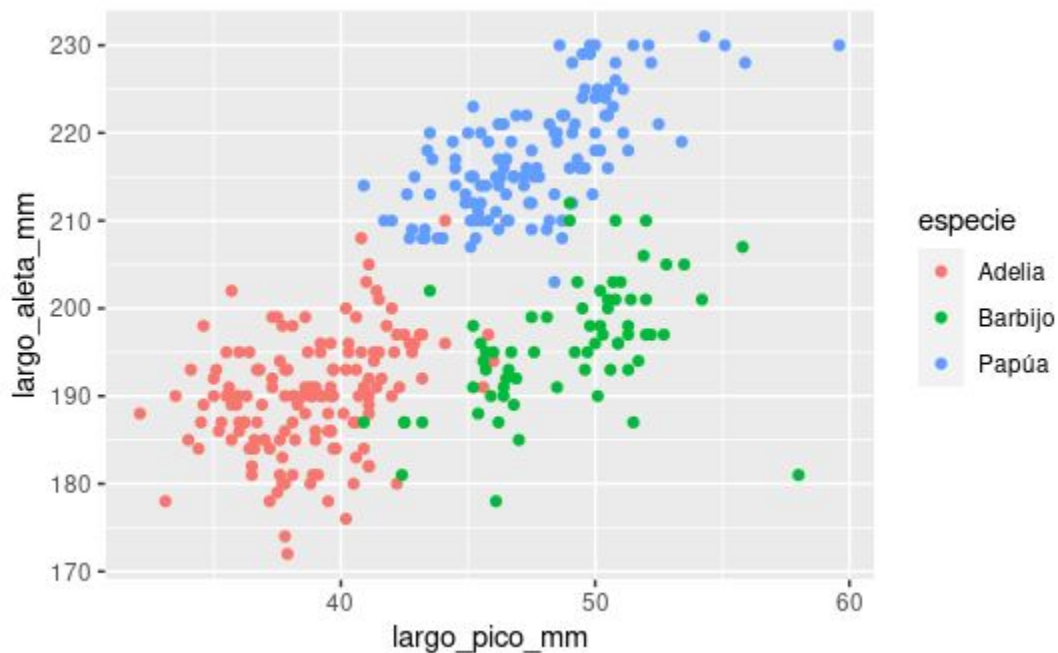
# **Pausa**

5 minutos

**No se  
desconecten  
pero  
retirensé de las  
pantallas.**

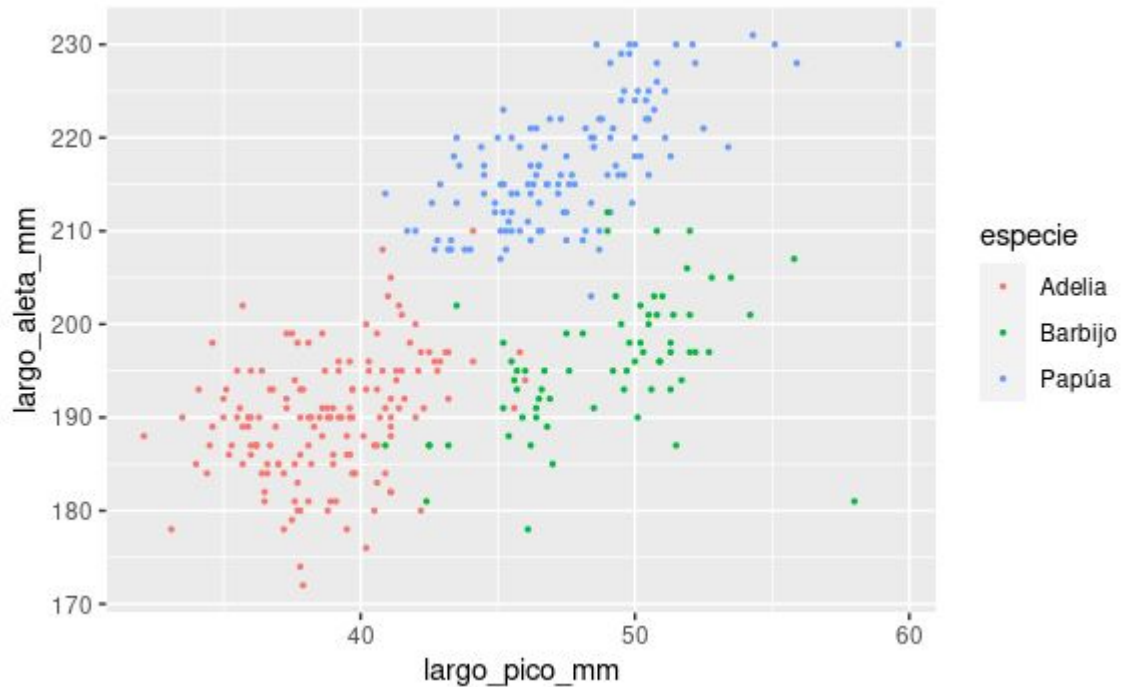
# Overplotting

# Mejorar la presentación



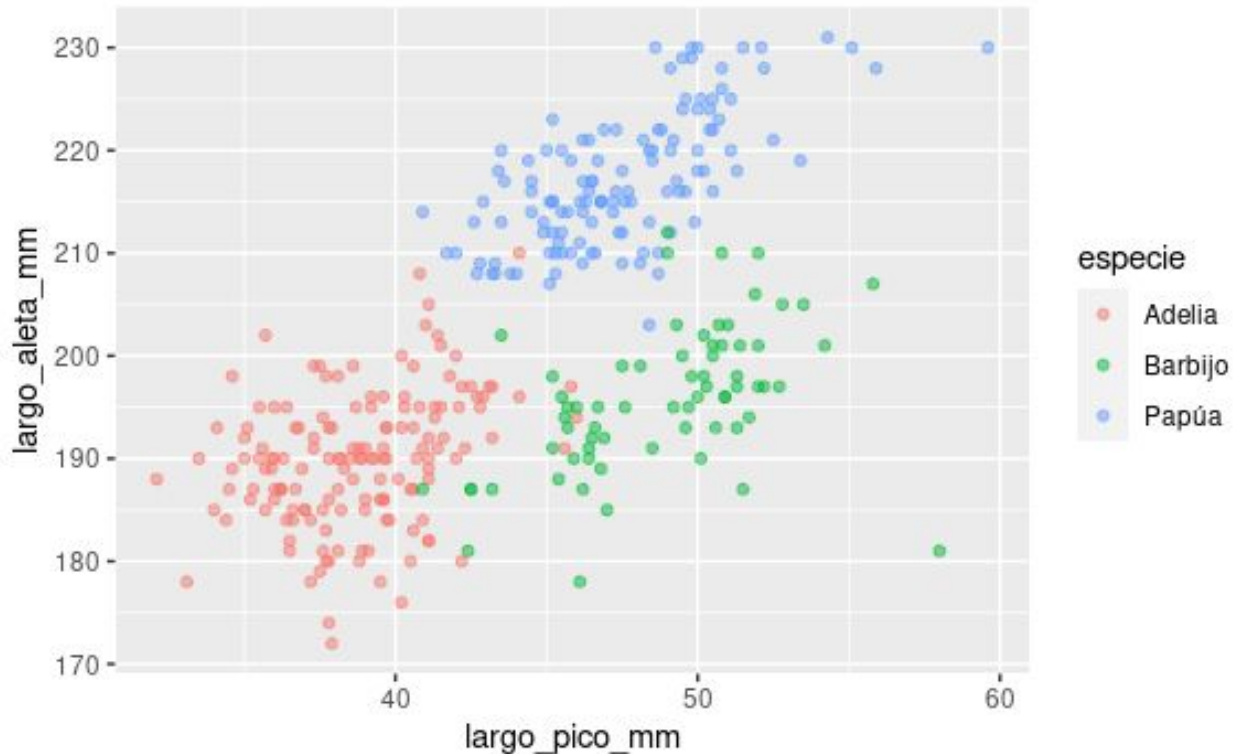
¿Cómo podemos visualizar mejor ese montón de puntos?

# Tamaño de los puntos (size)

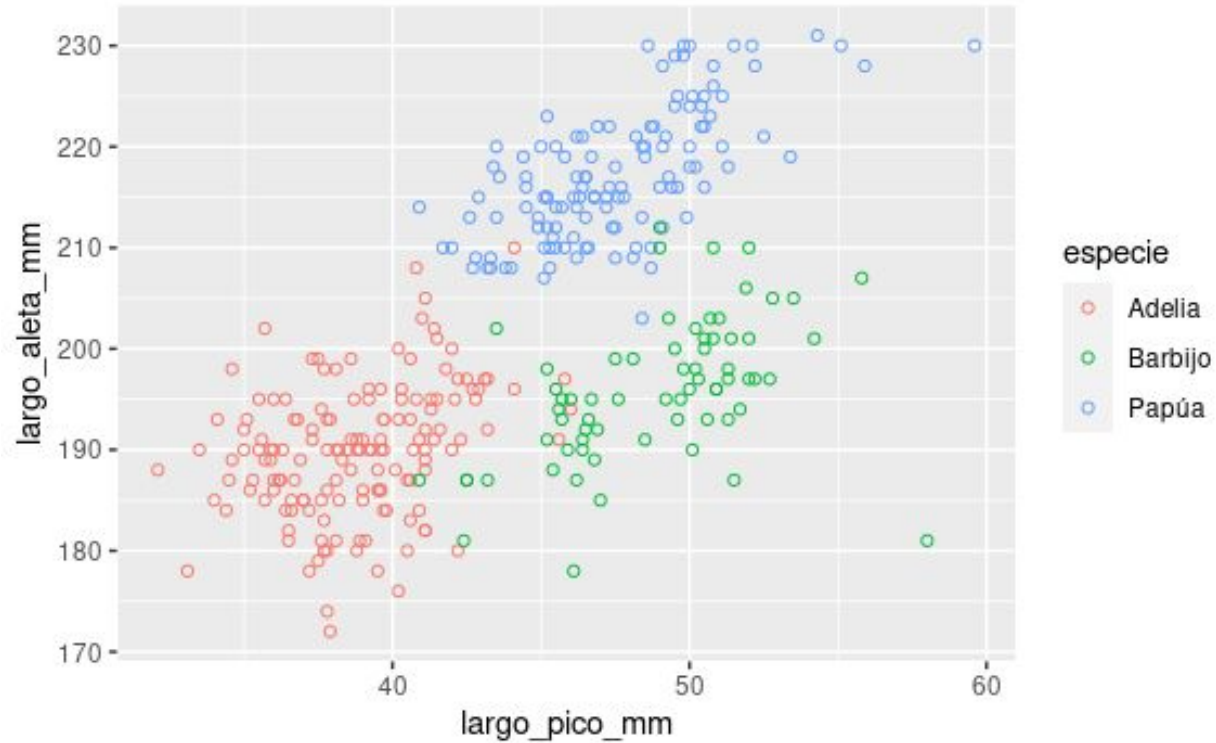


# Transparencia (alpha)

rango de 0 a 1

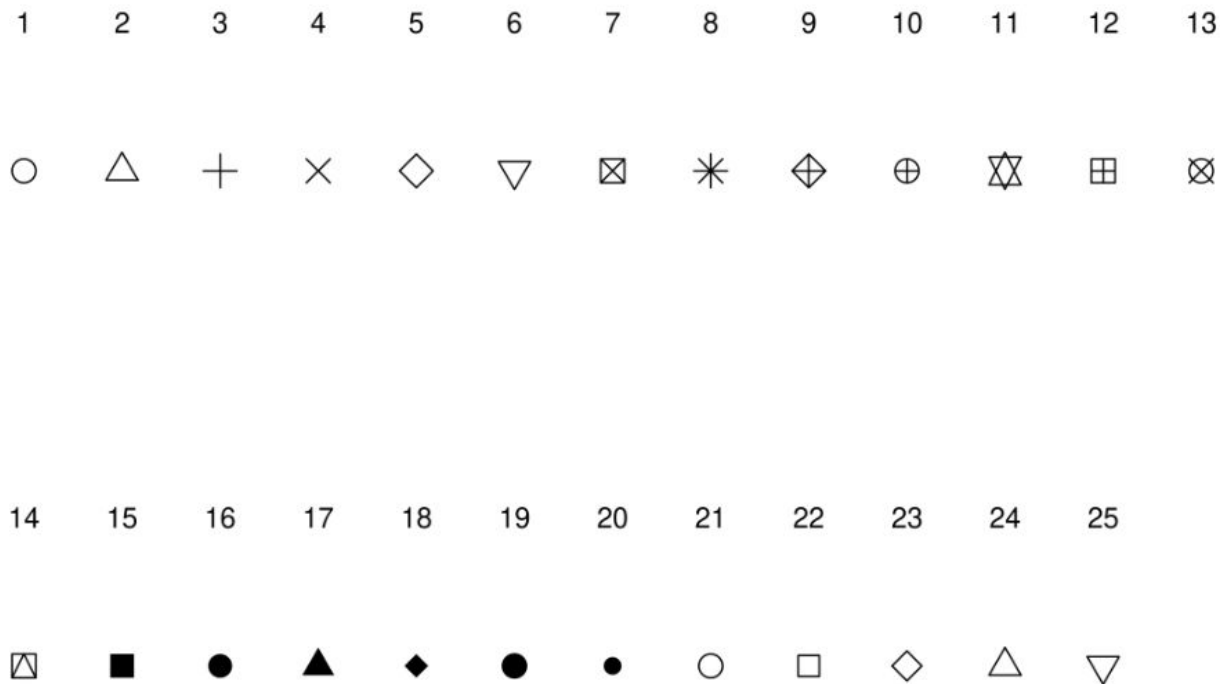


# Forma del punto (shape)



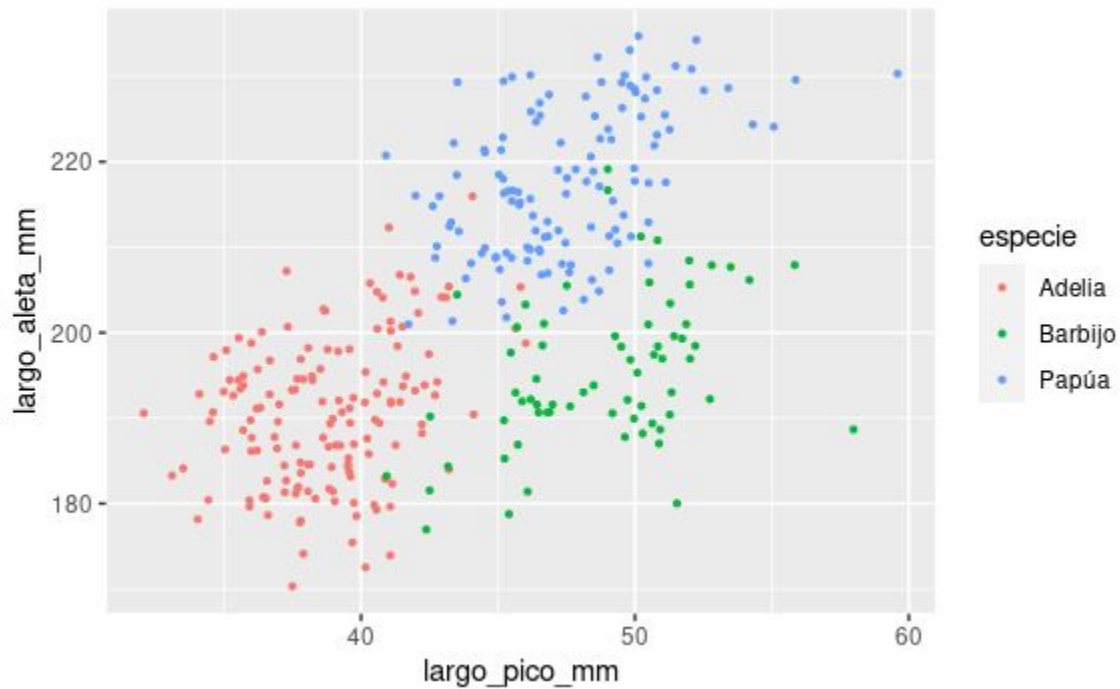
# 25 tipos de shapes

Se recomienda  
usar **máximo 6**  
shapes  
distintas

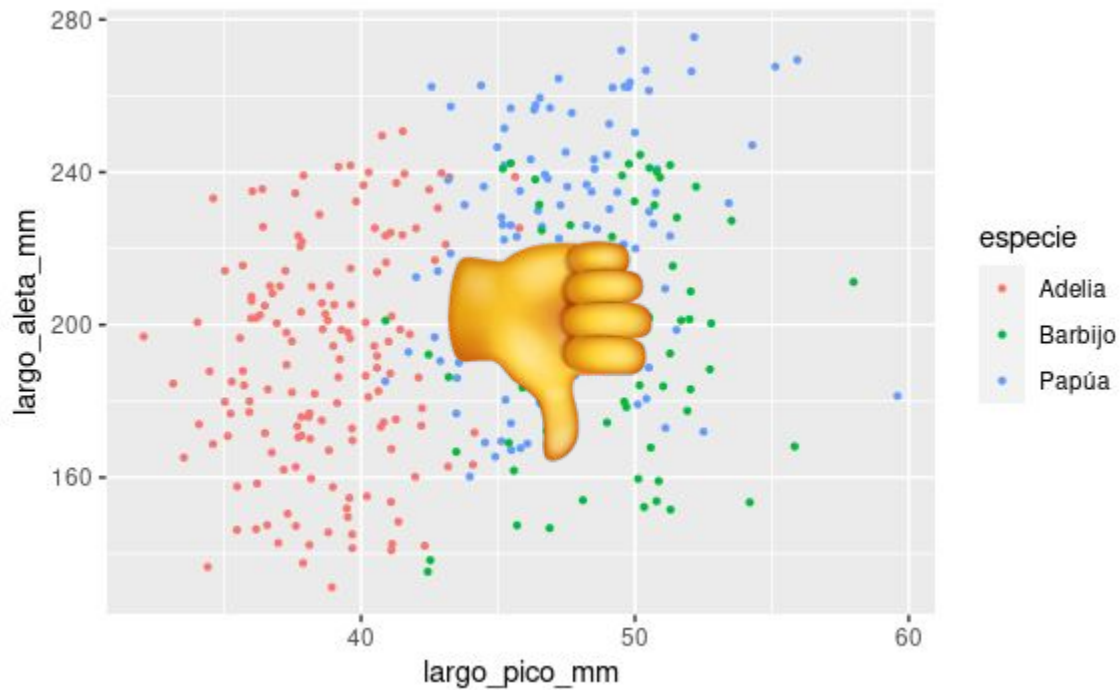




# geom\_jitter()

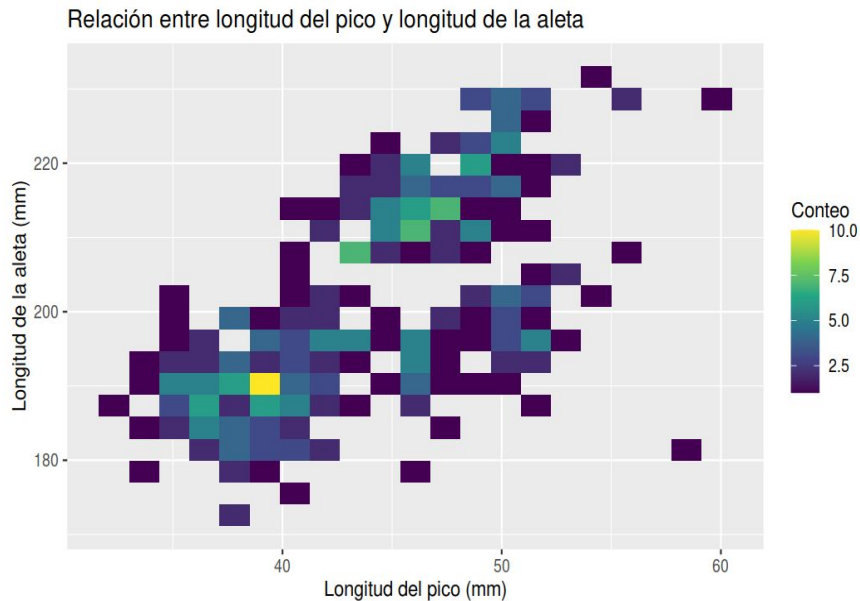


# geom\_jitter()

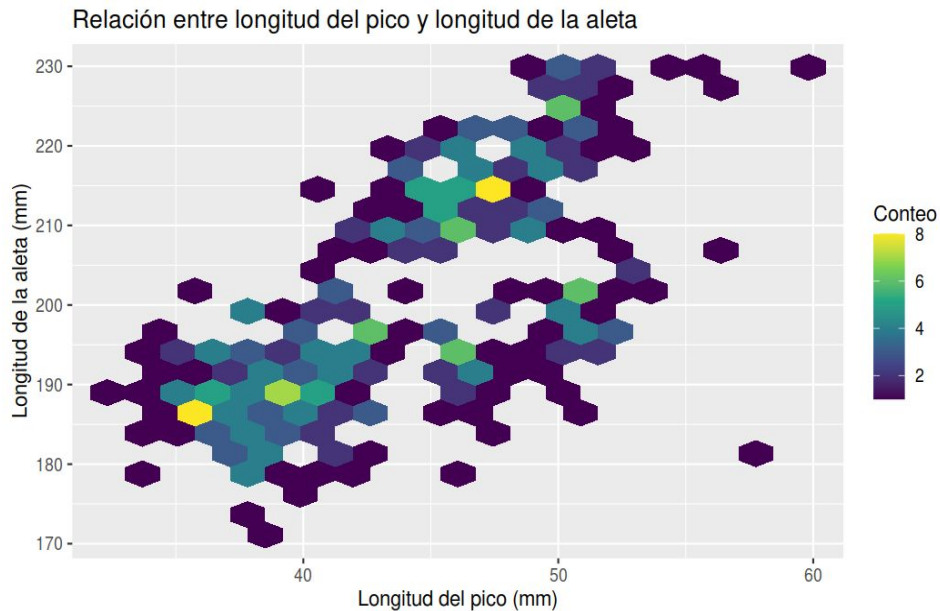


# Otras geometrías para explorar

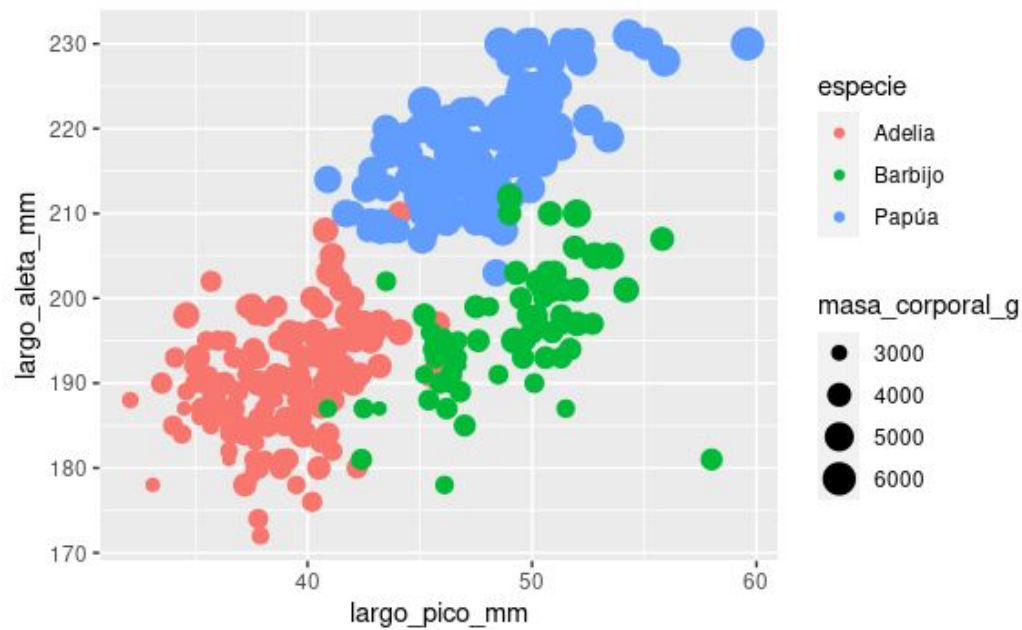
`geom_bin2d()`



`geom_hex()`

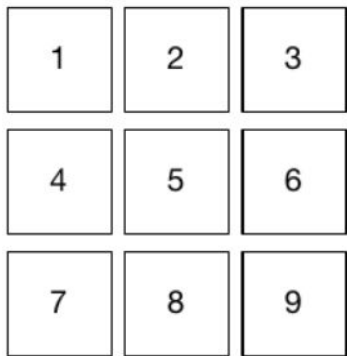


# Más de cuatro variables?

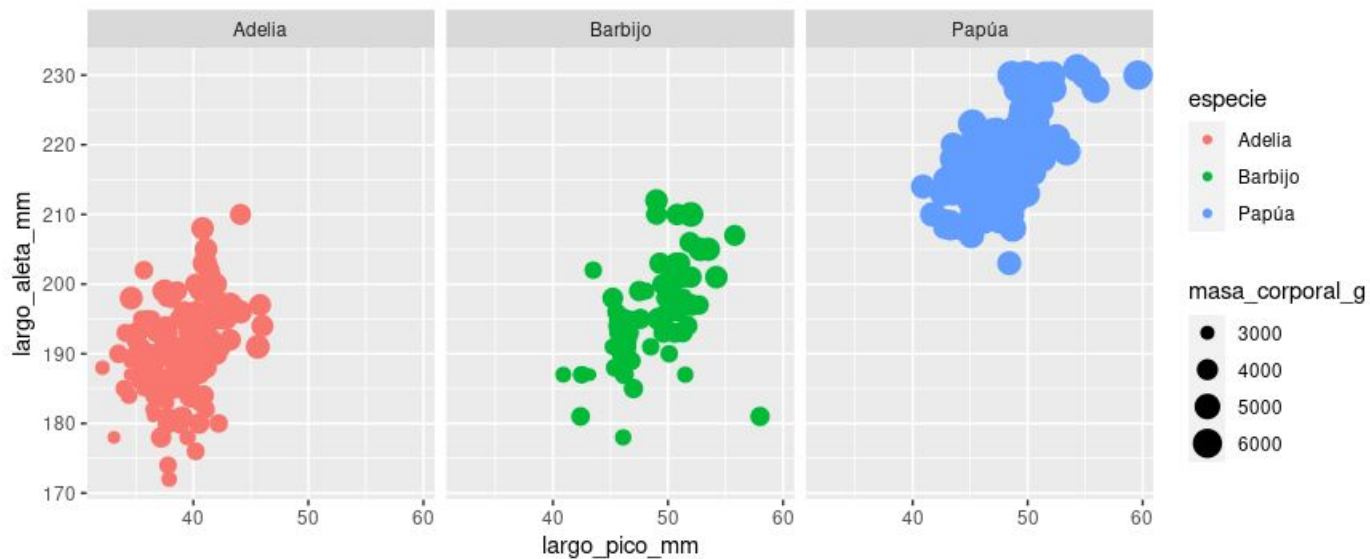


# Paneles

# facet\_wrap()

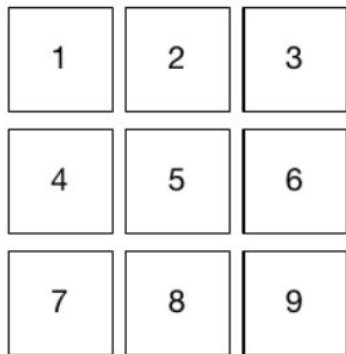


**facet\_wrap**

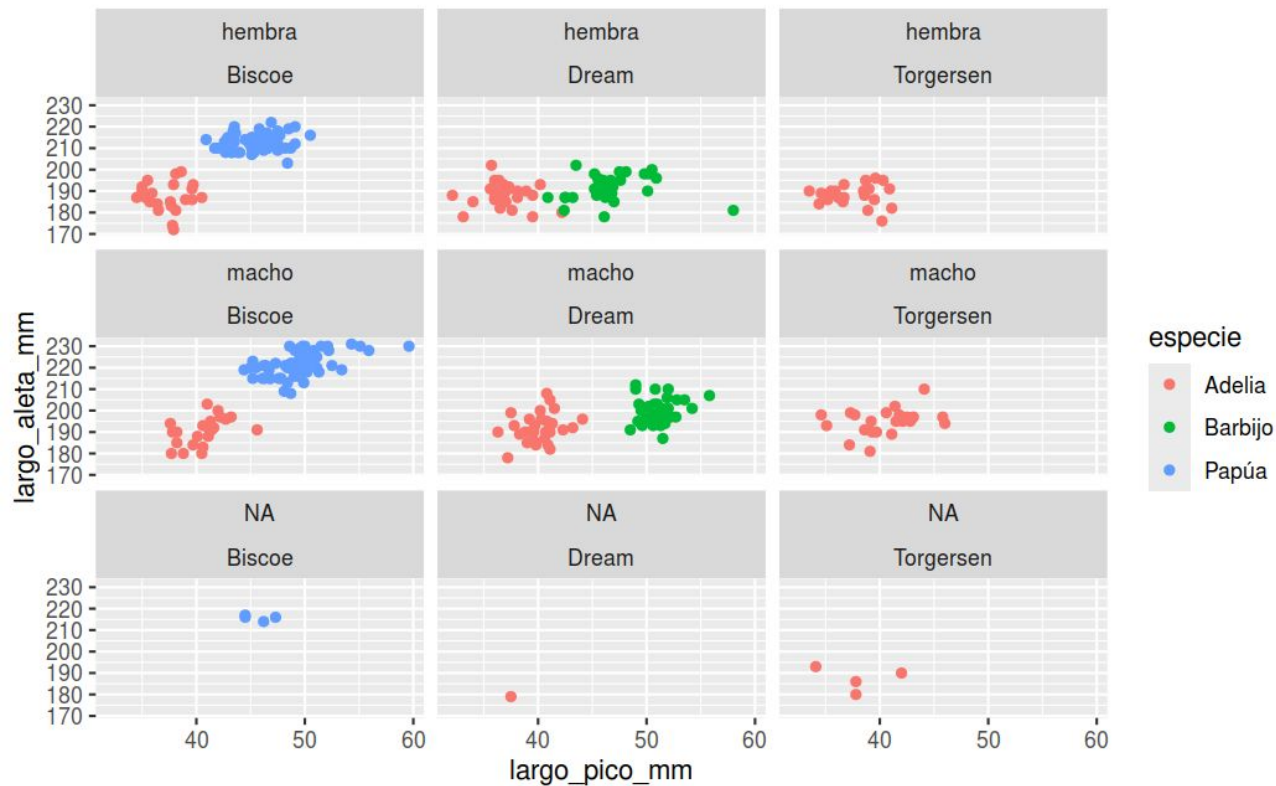


# facet\_wrap()

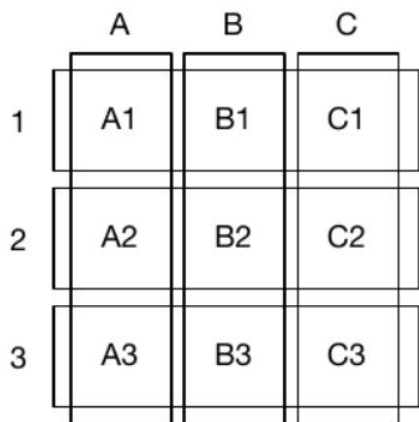
Se puede con 2  
variables pero  
no queda claro



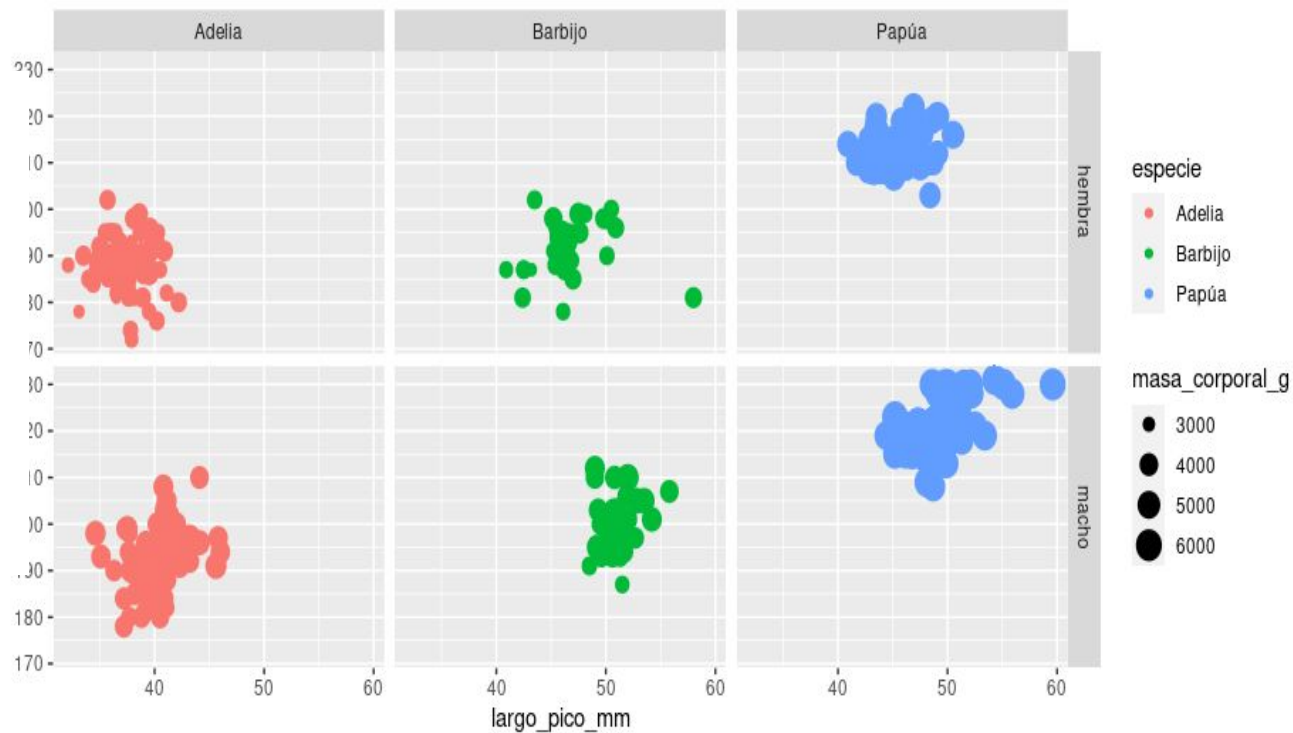
facet\_wrap



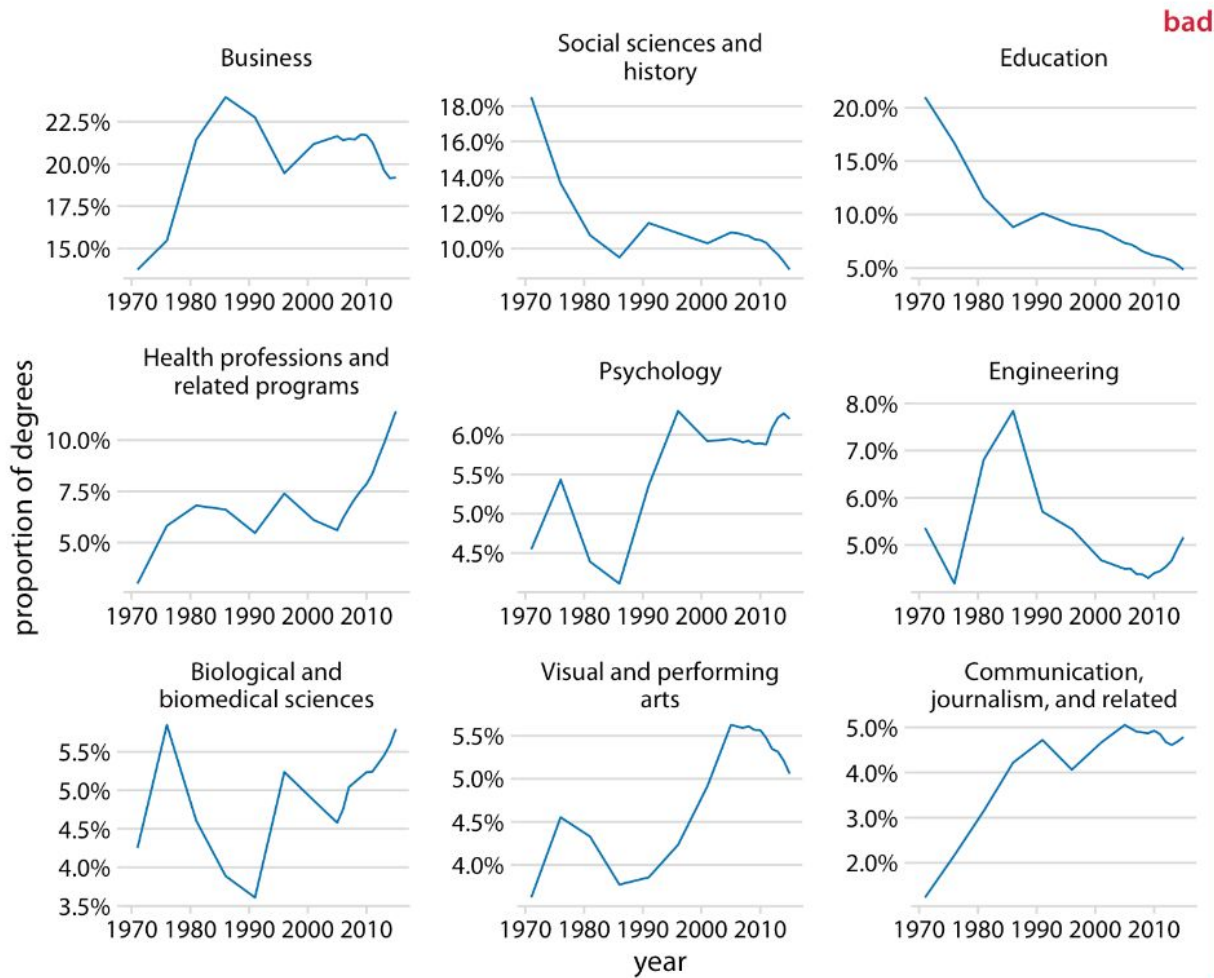
# facet\_grid()

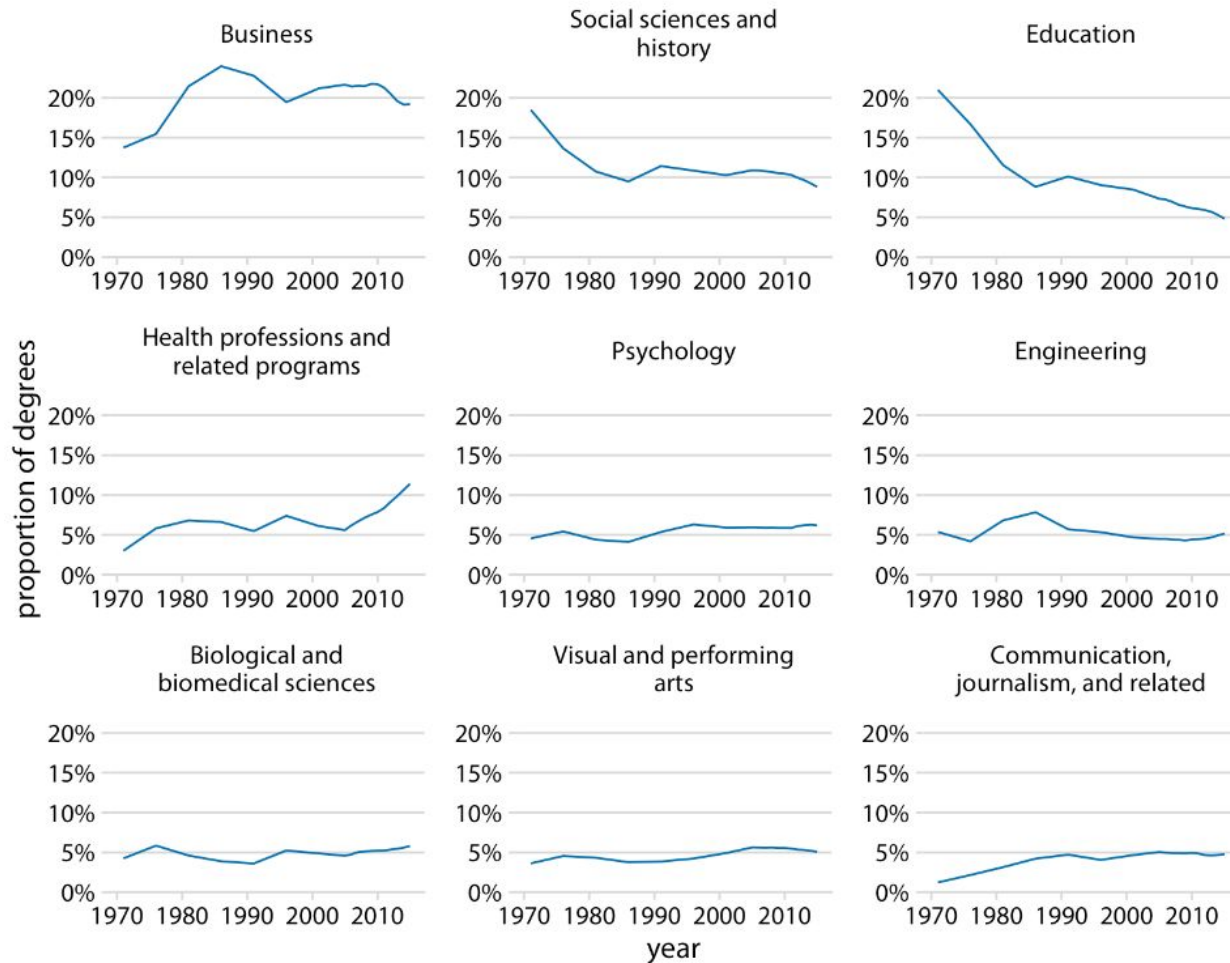


facet\_grid









**¡Nos vemos la semana próxima!**

Estén atentos al campus.