

Curso de **Análisis Exploratorio de Datos**

Jesús Vélez Santiago
@jvelezmagic

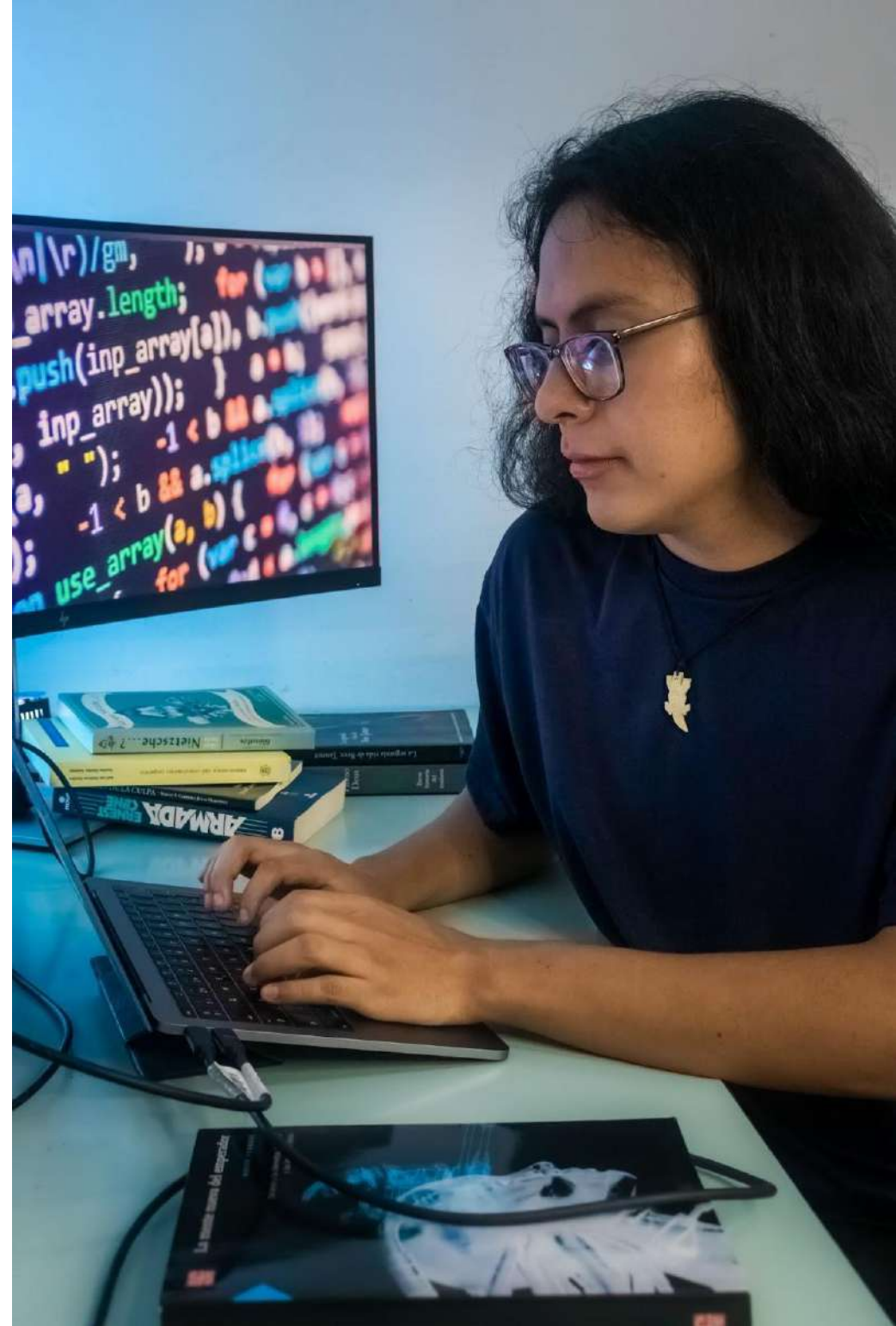


Científico **genómico/datos**.

A través del:

- Desarrollo de software
- Matemáticas
- Estadística

Busco **crear** herramientas
para **enseñar** y **ayudar**
al desarrollo de la ciencia.





¿Qué aprenderás?

- Entenderás qué es y para qué sirve un análisis exploratorio de datos.
- Conocerás los distintos tipos de análisis de datos que existen.
- Identificarás distintos tipos de variables y análisis que puedes realizar con ellas.

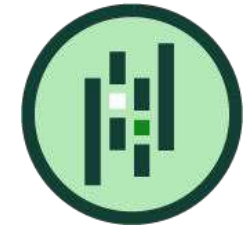


¿Qué aprenderás?

- Serás capaz de explorar conjuntos de datos con 1, 2, 3 o más variables.
- Identificarás las ventajas y desventajas de distintas visualizaciones de datos.

Conocimientos previos

- Python y Jupyter Notebooks.
- Principios de visualización de datos.
- Pandas y NumPy.
- Matplotlib y Seaborn.
- Estadística.



**¿Qué es el análisis
exploratorio de datos?**



**Proceso de conocer
en detalle a tus datos,
darles sentido.**

**Transformarlos
en información
útil.**



**Determinar
cómo tratarlos.**

**Interrogarlos para
obtener las respuestas
que necesites.**

¿Cómo hacer un análisis exploratorio de datos?

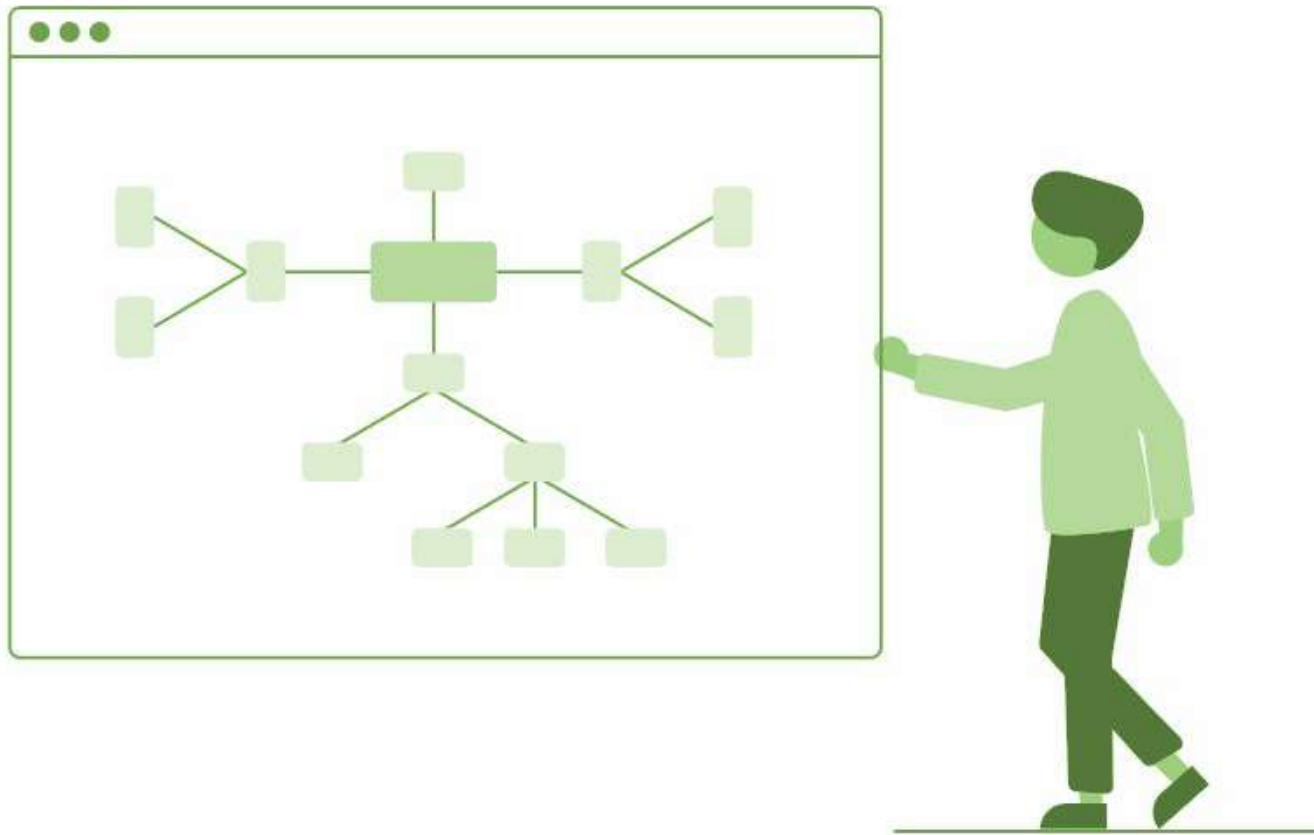
**¿Por qué deberías
realizar un análisis
exploratorio de datos?**



Organizar y entender las variables



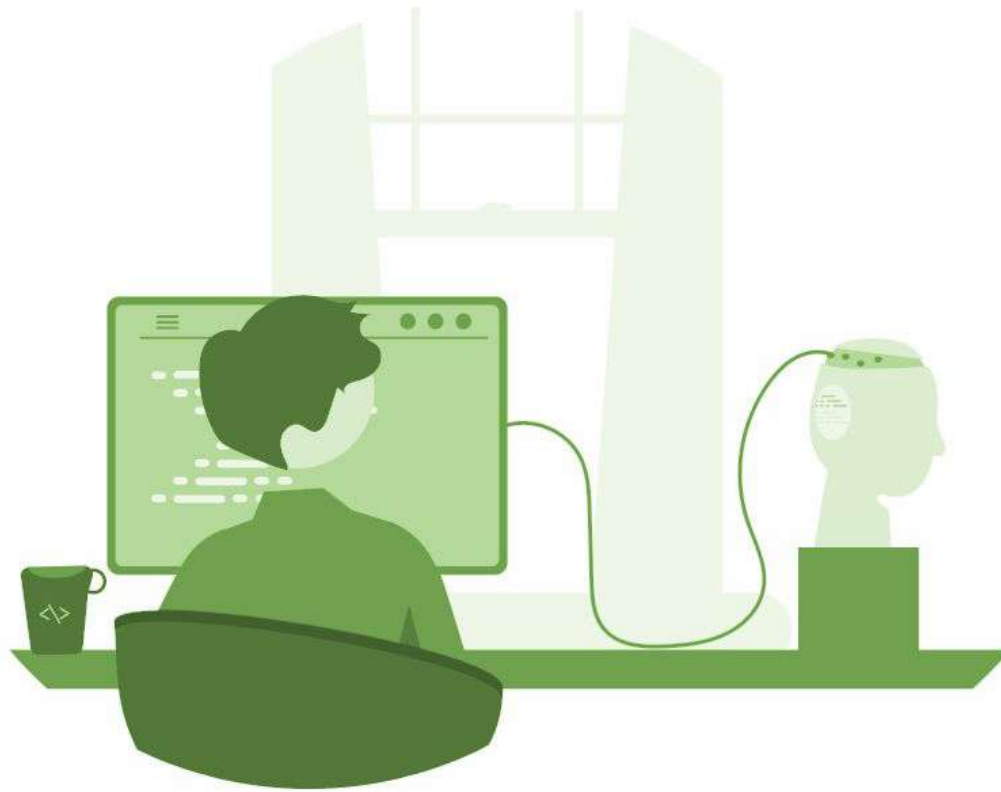
Establecer relaciones entre las variables



Encontrar patrones ocultos en los datos



Ayudarte a escoger el modelo correcto para la necesidad correcta



Ayudarte a tomar una decisión informada



**¿Cuáles son los pasos
de un análisis
exploratorio de datos?**



1

Hacer preguntas

2

**Determinar
el tamaño
de los datos**

3

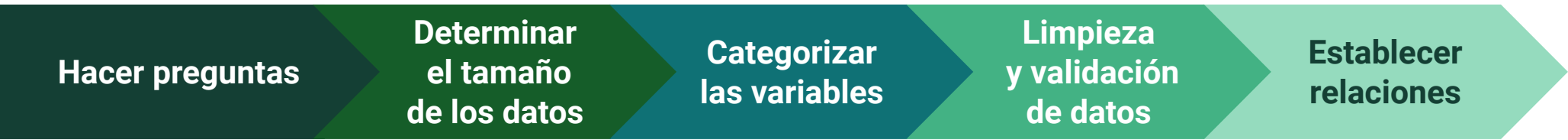
**Categorizar
las variables**

4

**Limpieza
y validación
de datos**

5

**Establecer
relaciones**



1

Hacer preguntas

2

Determinar
el tamaño
de los datos

3

Categorizar
las variables

4

Limpieza
y validación
de datos

5

Establecer
relaciones

- ¿Qué te gustaría encontrar?
- ¿Qué quisieras saber de los datos?
- ¿Cuál es la razón para realizar el análisis?

1

Hacer preguntas

2

**Determinar
el tamaño
de los datos**

3

Categorizar
las variables

4

Limpieza
y validación
de datos

5

Establecer
relaciones

- ¿Cuántas observaciones existen?
- ¿Cuántas variables hay?
- ¿Necesito todas las observaciones?
- ¿Necesito todas las variables?

1

Hacer preguntas

2

Determinar
el tamaño
de los datos

3

**Categorizar
las variables**

4

Limpieza
y validación
de datos

5

Establecer
relaciones

- ¿Cuántas variables categóricas existen?
- ¿Cuántas variables continuas existen?
- ¿Cómo puedo explorar cada variable dependiendo de su categoría?

1

Hacer preguntas

2

Determinar
el tamaño
de los datos

3

Categorizar
las variables

4

Limpieza
y validación
de datos

5

Establecer
relaciones

- ¿Tengo valores faltantes?
- ¿Cuál es la proporción de datos faltantes?
- ¿Cómo puedo tratar a los datos faltantes?
- ¿Cuál es la distribución de los datos?
- ¿Tengo valores atípicos?

1

Hacer preguntas

2

Determinar
el tamaño
de los datos

3

Categorizar
las variables

4

Limpieza
y validación
de datos

5

**Establecer
relaciones**

- ¿Existe algún tipo de relación entre mi variable X y Y ?
- ¿Qué pasa si ahora considero a la variable Z en el análisis?
- ¿Qué significa que las observaciones se agrupen?
- ¿Qué significa el patrón que se observa?

1

Hacer preguntas

- ¿Qué te gustaría encontrar?
- ¿Qué quisieras saber de los datos?
- ¿Cuál es la razón por la que realizas el análisis?

2

Determinar el tamaño de los datos

- ¿Cuántas observaciones existen?
- ¿Cuántas variables hay?
- ¿Necesito todas las observaciones?
- ¿Necesito todas las variables?

3

Categorizar las variables

- ¿Cuántas variables categóricas existen?
- ¿Cuántas variables continuas existen?
- ¿Cómo puedo explorar cada variable dependiendo de su categoría?

4

Limpieza y validación de datos

- ¿Tengo valores faltantes?
- ¿Cuál es la proporción de datos faltantes?
- ¿Cómo puedo tratar a los datos faltantes?
- ¿Cuál es la distribución de los datos?
- ¿Tengo valores atípicos?

5

Establecer relaciones

- ¿Existe algún tipo de relación entre mi variable X y Y?
- ¿Qué pasa si ahora considero a la variable Z en el análisis?
- ¿Qué significa que las observaciones se agrupen?
- ¿Qué significa el patrón que se observa?

A pesar de que pueda parecer un ciclo infinito...

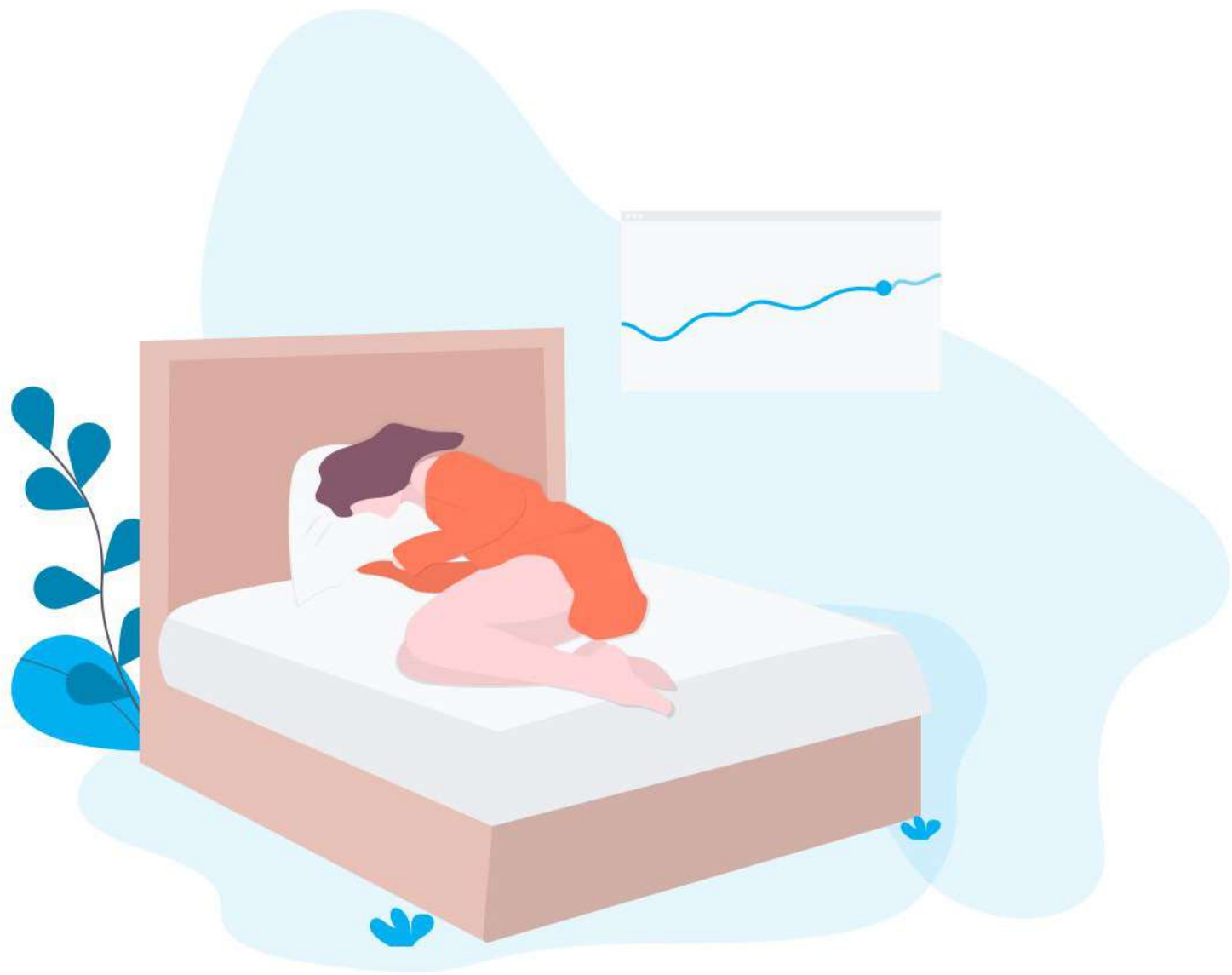


En algún momento debes romperlo y continuar

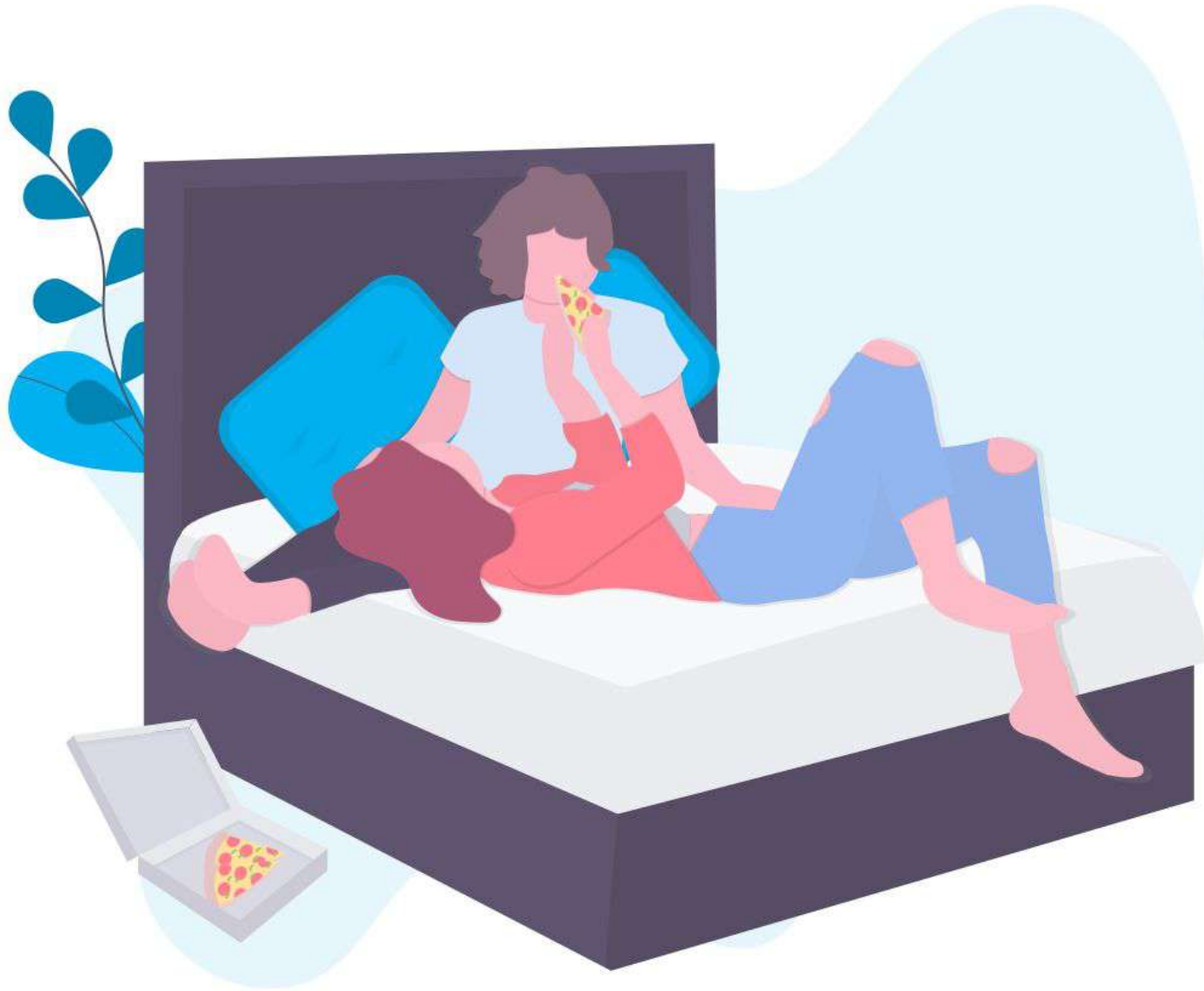


Tipos de analítica de datos



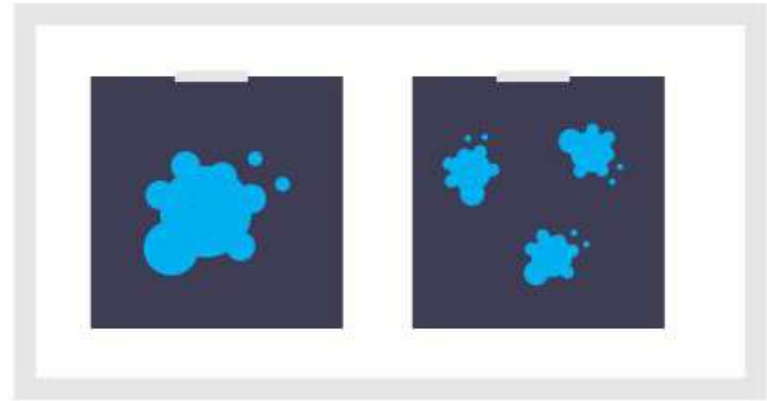














Descriptiva



¿Qué sucedió?

Provee de ideas sobre eventos del pasado.

Diagnóstica



¿Por qué sucedió?

Profundiza para encontrar las causas del evento.

Predictiva



¿Qué podría pasar si?

Utiliza los datos del pasado para predecir un futuro evento.

Prescriptiva

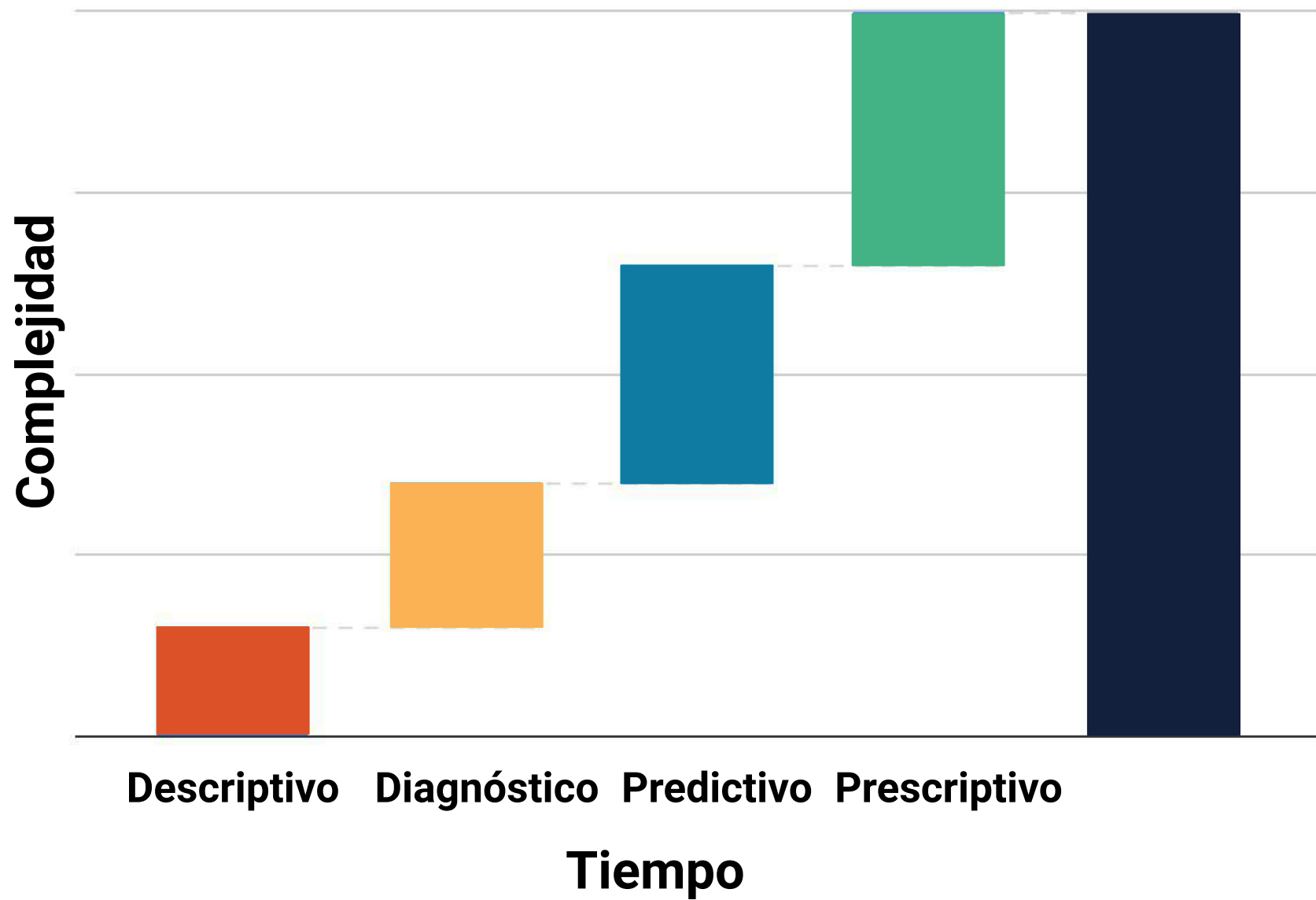


¿Qué debería hacerse?

Analiza decisiones y eventos del pasado para estimar la probabilidad de diferentes resultados.

Pasado

Futuro



Tipos de datos y análisis de variables

Cualitativos

Cuantitativos

Tipos de datos

```
graph TD; A[Tipos de datos] --> B[Categoricos]; A --> C[Numericos]; B --> D[Ordinal]; B --> E[Nominal]; D & E --> F["Género, a favor o en contra, nivel de estudios, categoría de película, día de la semana, sabor, textura."]; C --> G[Discreto]; C --> H[Continuo]; G & H --> I["Altura, peso, longitud, volumen, temperatura, humedad, edad, número de amigos, calificación."];
```

Categoricos

Numéricos

Ordinal

Nominal

Género, a favor o en contra, nivel de estudios, categoría de película, día de la semana, sabor, textura.

Discreto

Continuo

Altura, peso, longitud, volumen, temperatura, humedad, edad, número de amigos, calificación.

Análisis Univariado

Analizar cada variable por separado.



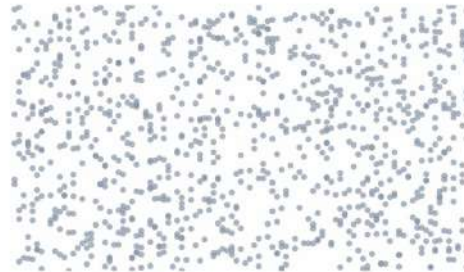
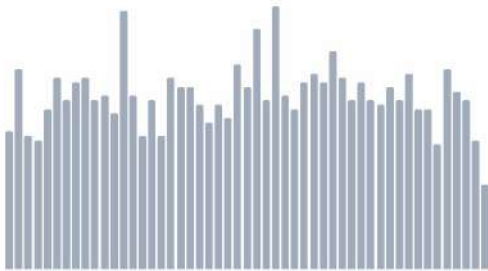
Análisis Bivariado

Analizar la relación de cada par de variables.



Análisis Multivariado

Analizar el efecto simultáneo de múltiples variables.

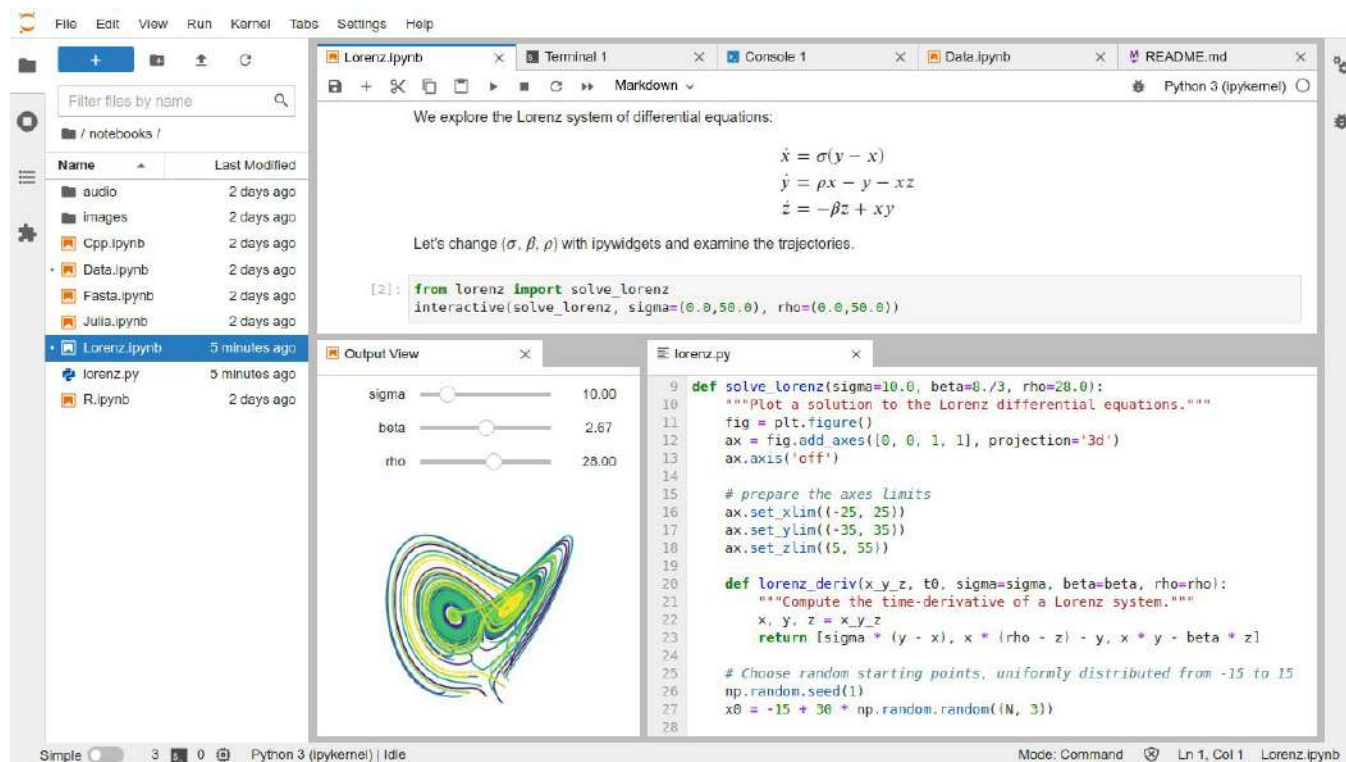


Herramientas de software para el análisis exploratorio de datos

Herramientas para el análisis exploratorio de datos



Jupyter notebooks

A screenshot of a Jupyter Notebook interface. The top menu bar includes File, Edit, View, Run, Kernel, Tabs, Settings, and Help. The left sidebar shows a file browser with a search bar and a list of files and folders. The main area displays a notebook with text, equations, and a plot. The bottom right shows a code editor with Python code for solving the Lorenz system.

File Edit View Run Kernel Tabs Settings Help

Filter files by name

/ notebooks /

Name	Last Modified
audio	2 days ago
images	2 days ago
Cpp.ipynb	2 days ago
Data.ipynb	2 days ago
Fasta.ipynb	2 days ago
Julia.ipynb	2 days ago
Lorenz.ipynb	5 minutes ago
lorenz.py	5 minutes ago
R.ipynb	2 days ago

We explore the Lorenz system of differential equations:

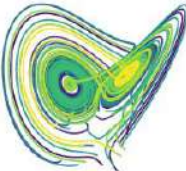
$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

Let's change (σ, β, ρ) with ipywidgets and examine the trajectories.

```
[2]: from lorenz import solve_lorenz
interactive(solve_lorenz, sigma=(0.0, 50.0), rho=(0.0, 50.0))
```

Output View

sigma 10.00
beta 2.67
rho 28.00

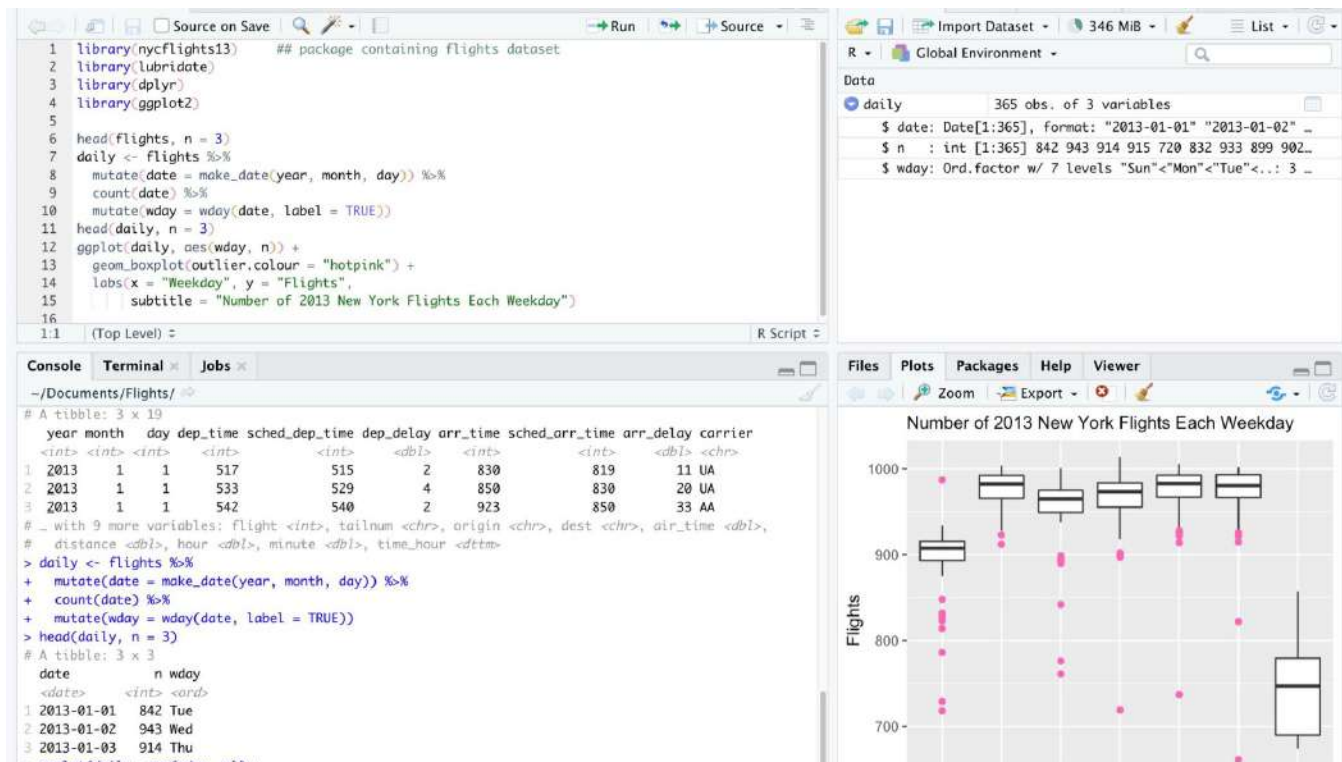


lorenz.py

```
9 def solve_lorenz(sigma=10.0, beta=8./3, rho=28.0):
10     """Plot a solution to the Lorenz differential equations."""
11     fig = plt.figure()
12     ax = fig.add_axes([0, 0, 1, 1], projection='3d')
13     ax.axis('off')
14
15     # prepare the axes limits
16     ax.set_xlim((-25, 25))
17     ax.set_ylim((-35, 35))
18     ax.set_zlim((5, 55))
19
20     def lorenz_deriv(x_y_z, t0, sigma=sigma, beta=beta, rho=rho):
21         """Compute the time-derivative of a Lorenz system."""
22         x, y, z = x_y_z
23         return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]
24
25     # Choose random starting points, uniformly distributed from -15 to 15
26     np.random.seed(1)
27     x0 = -15 + 30 * np.random.random((N, 3))
28
```

Simple 3 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 Lorenz.ipynb

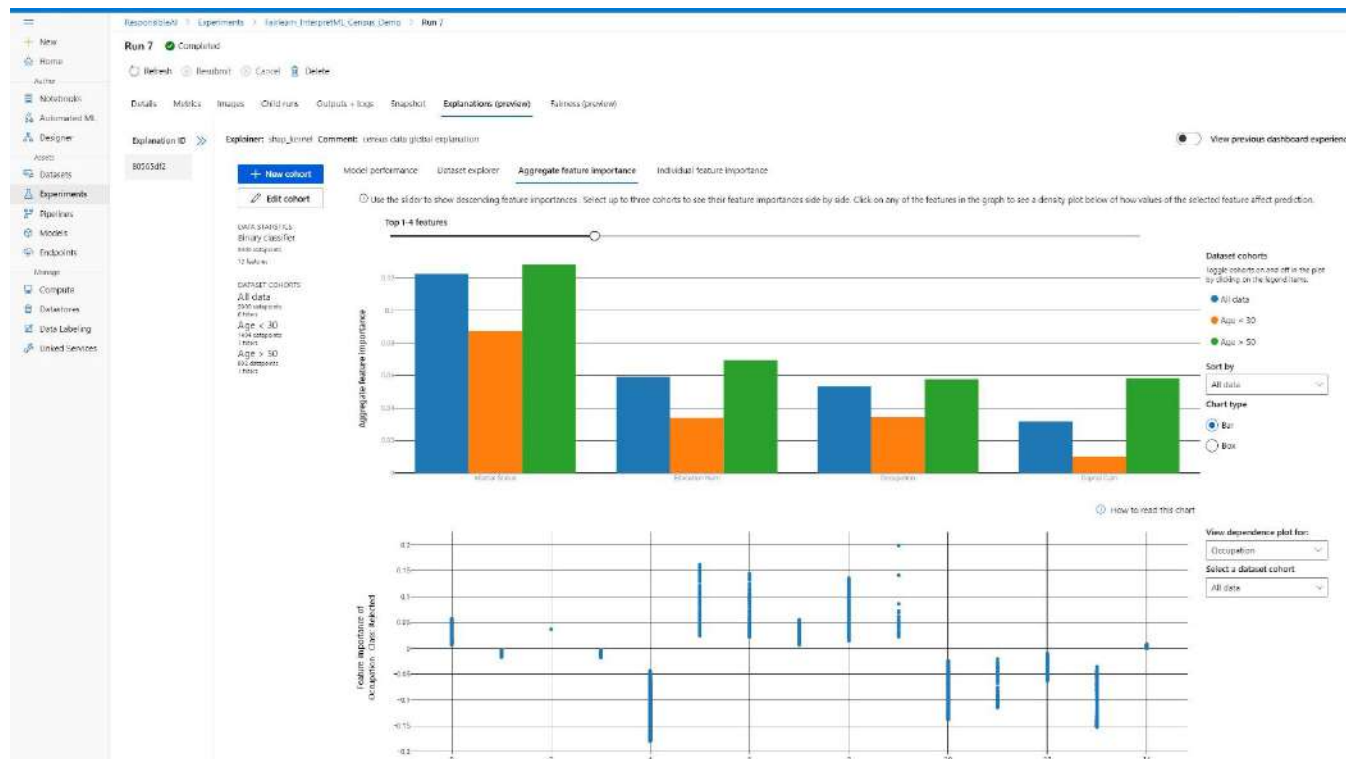
RStudio



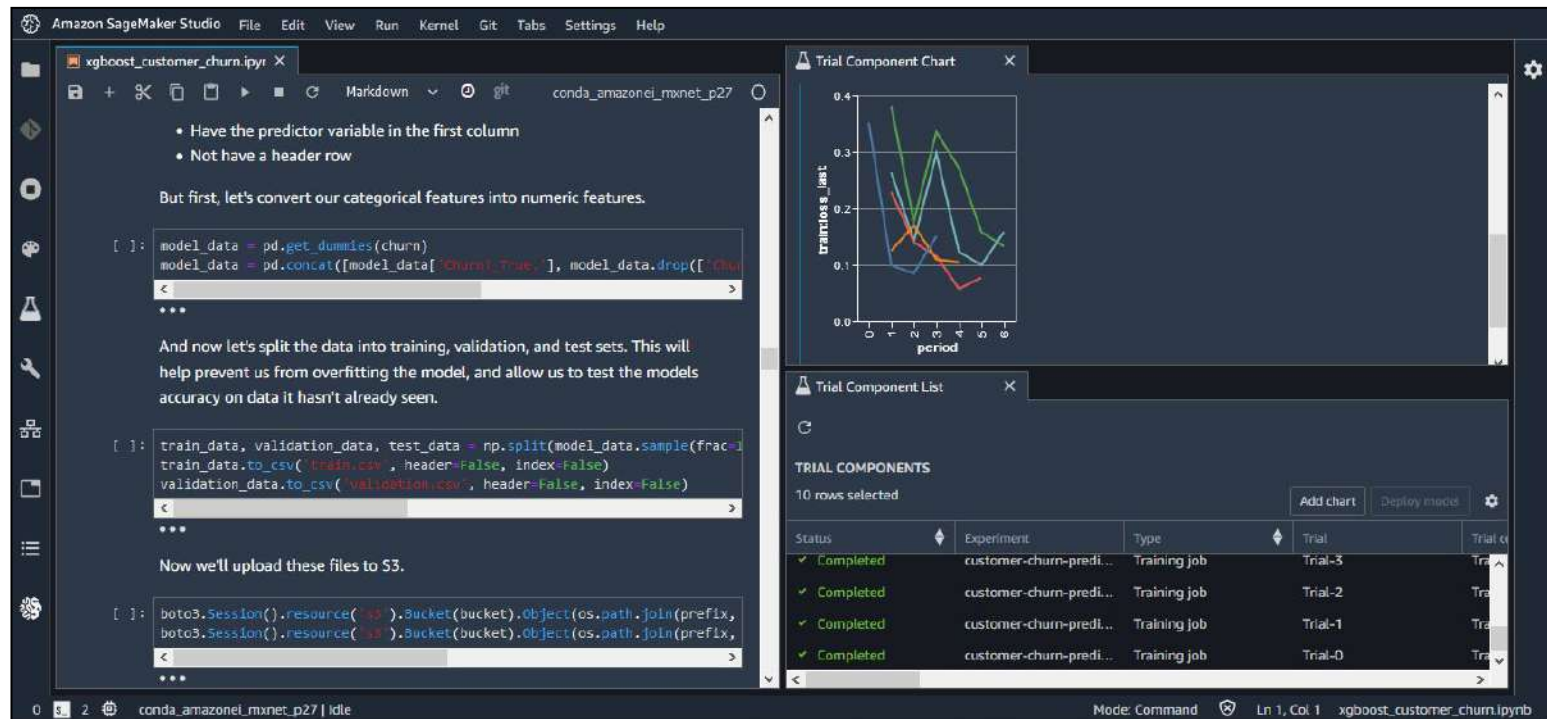
Azure Machine Learning



Azure Machine Learning



100%



Deepnote



Search

Integrations

Settings & Members

Workspace

> Bug reports

> Exploration

Lead scoring

Penguin analysis

> Support Projects

> External

Private

> Private Folder

Private project

> 1:1s

Explore

Documentation & Help

Running 2 cells

Run notebook

Penguin dataset analysis 🐧

An exploration of the Penguin dataset demonstrating the power of Deepnote. You can collaboratively work on Python cells, add comments, visualize all your dataframes or schedule a notebook to run automatically at a certain time.

```
1 # Imports
2 import pandas as pd
3
4 penguin_df = pd.read_csv("./penguins.csv")
5
6 print("Dataframe import successful 🐧")
7
8 def load_and_process_img(path_to_img):
9     img = load_img(path_to_img)
10     img = tf.keras.applications.vgg19.preprocess_input(img)
11     return image
```

Dataframe import successful 🐧

Visualization of penguin_df

Scatterplot

X axis

flipper_length_mm

Y axis

body_mass_g


A scatterplot showing the relationship between flipper length (mm) on the x-axis and body mass (g) on the y-axis for penguins. The data points are colored by species: Adelia (orange), Chinstrap (purple), and Gentoo (green). The plot shows a positive correlation between flipper length and body mass, with Gentoo penguins generally having the highest values for both metrics.

Conociendo nuestros datos: palmerpenguins



Welcome to Palmer Station Antarctica LTER

A member of the Long Term Ecological Research Network

A photograph showing a group of penguins in the foreground, looking towards a zodiac boat in the background. The boat is dark and has several people wearing red life jackets on board. The water is calm, and the sky is overcast.

Palmer Station Antarctica LTER



CHINSTRAP!



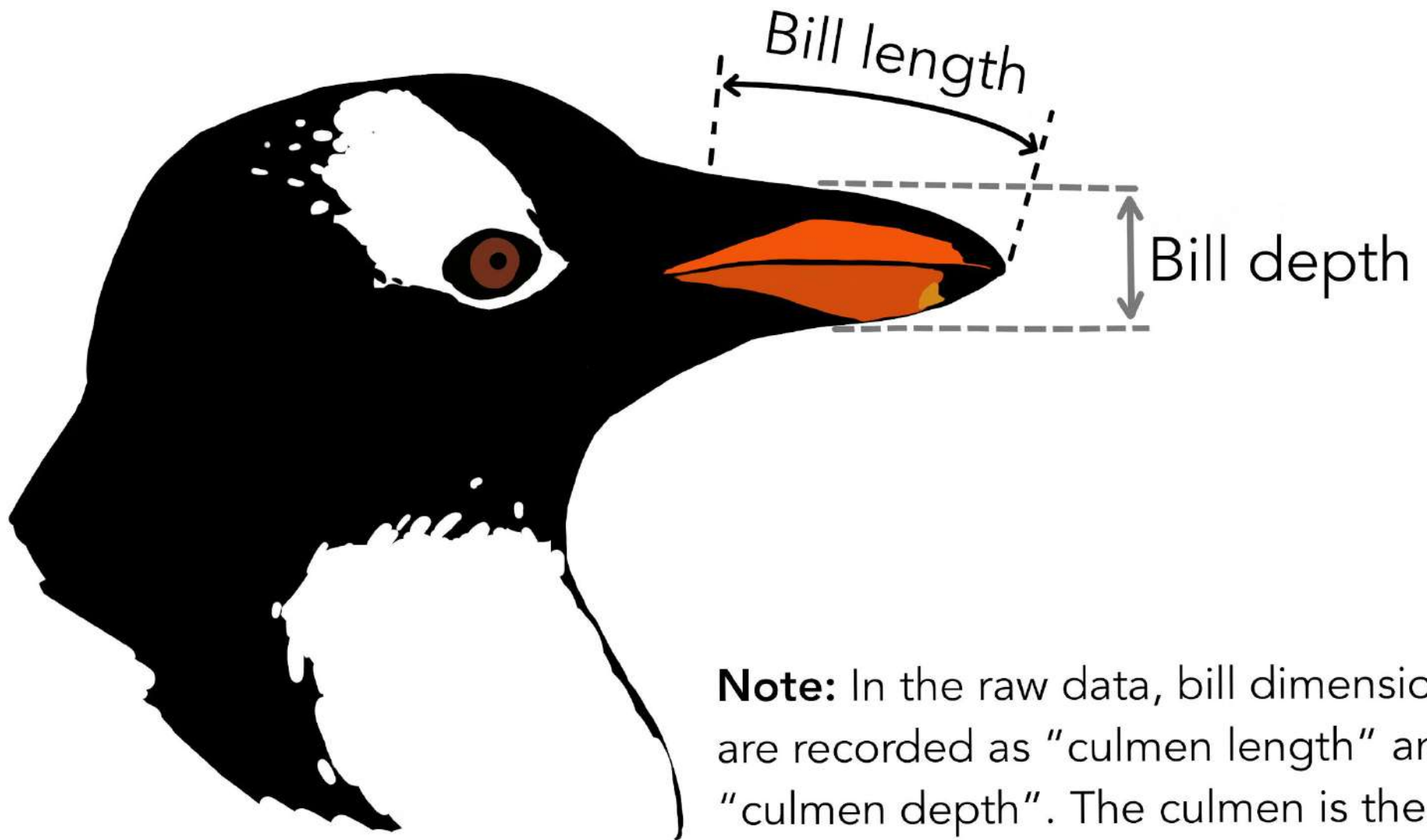
GENTOO!



ADÉLIE!



@allison_horst



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Recolección de datos, limpieza y validación

¿Qué es la recolección de datos?



“

Forma de recolección de información que permite obtener conocimiento de primera mano e ideas originales sobre el problema de investigación.

”

Tipos de recolección de datos

01

Primaria

Datos colectados de primera mano a través de encuestas, entrevistas, experimentos y otros.

02

Secundaria

Datos previamente recolectados por una fuente primaria externa al usuario primario.

Por ejemplo, datos de departamentos de gobierno o empresas similares a la del usuario primario.

03

Terciaria

Son datos que se adquieren de fuentes completamente externas al usuario primario.

Por ejemplo, a través de proveedores de datos.

¿Qué es la validación de datos?



“

**El proceso de asegurar
la consistencia y precisión
dentro de un conjunto de datos.**

”

<https://www.safe.com/what-is/data-validation/>

“

**Si los datos no son precisos
desde el comienzo, los
resultados definitivamente no
serán precisos.**

”

<https://www.safe.com/what-is/data-validation/>

¿Qué se debe validar para asegurar consistencia?

- Modelo de datos.
- Seguimiento de formato estándar de archivos.
- Tipos de datos.
- Rango de variables.
- Unicidad.
- Consistencia de expresiones.
- Valores nulos.

Explorando una variable categórica

Conteos y proporciones



Tabulación

“Contabiliza la frecuencia de aparición de cada valor único de una variable”.

Variable: Specie

Adelie
Gentoo
Chinstrap
Gentoo
Adelie
Adelie
Chinstrap
Adelie
Gentoo
Adelie
Adelie
Gentoo
Chinstrap
Gentoo

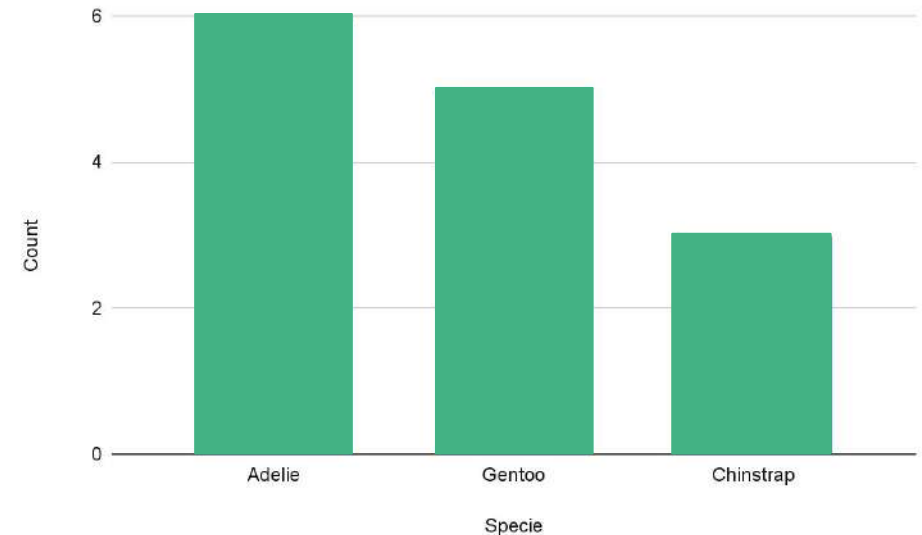
Tabulación

“Contabiliza la frecuencia de aparición de cada valor único de una variable”.

Variable: Specie

Adelie
Adelie
Adelie
Adelie
Adelie
Adelie
Gentoo
Gentoo
Gentoo
Gentoo
Gentoo
Chinstrap
Chinstrap
Chinstrap

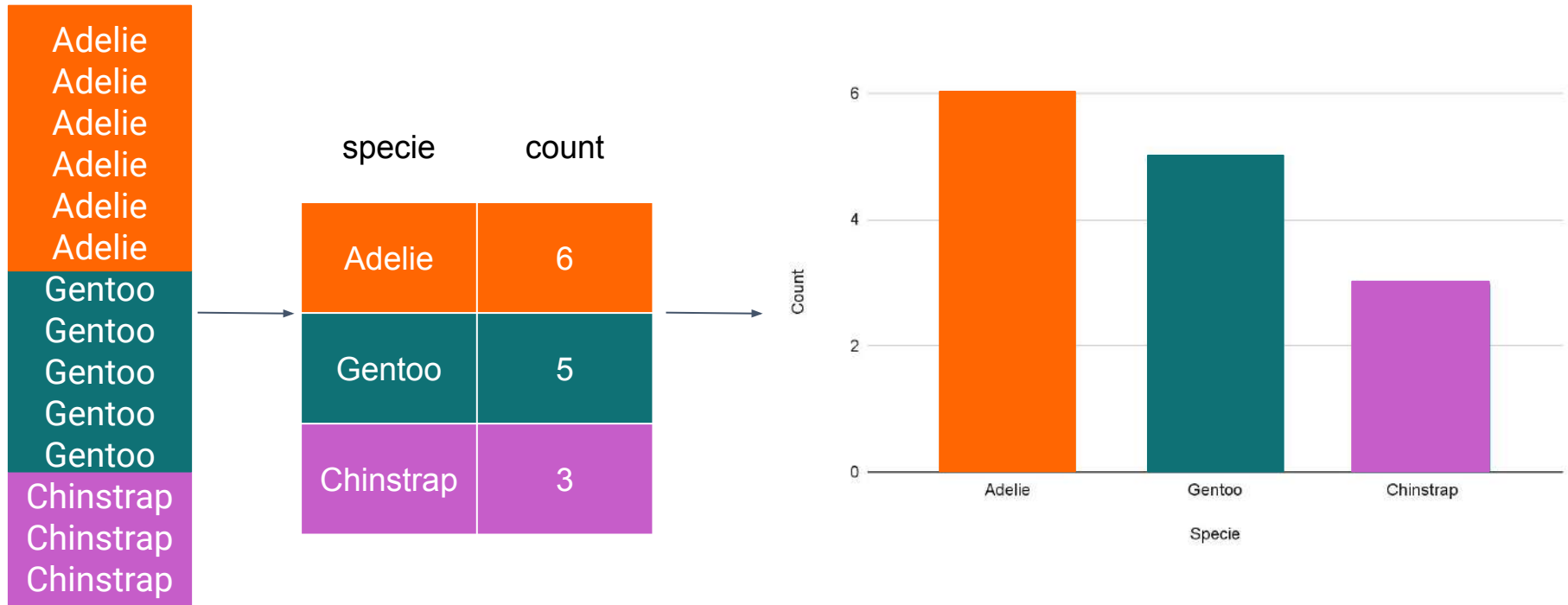
specie	count
Adelie	6
Gentoo	5
Chinstrap	3



Tabulación

“Contabiliza la frecuencia de aparición de cada valor único de una variable”.

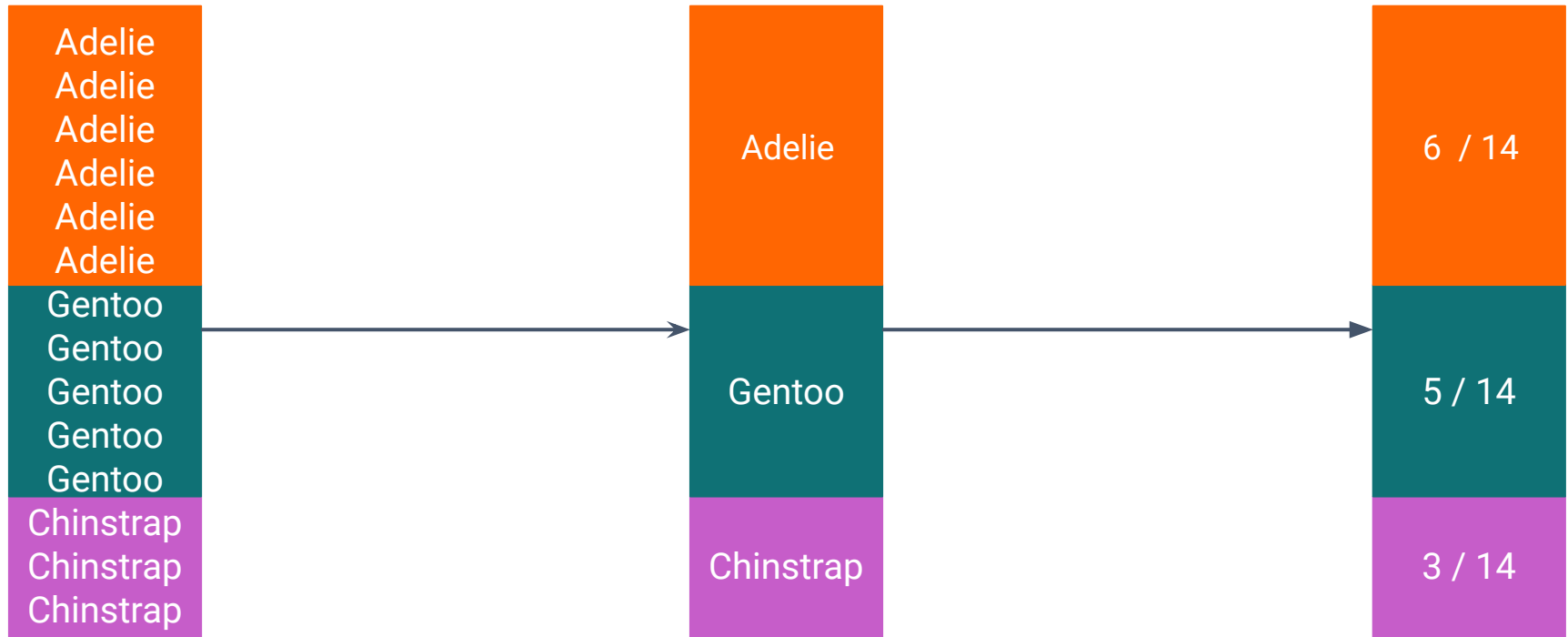
Variable: Specie



Proporciones

“Relación de correspondencia entre las partes y el todo”.

Variable: Specie





Extendiendo la idea de conteo

- Tabulación cruzada o tablas de contingencia.

Estadística descriptiva aplicada

Medidas de tendencia central

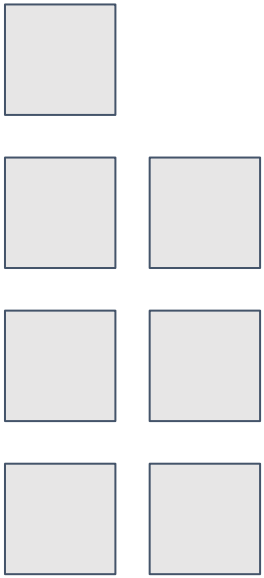


Medidas de tendencia central

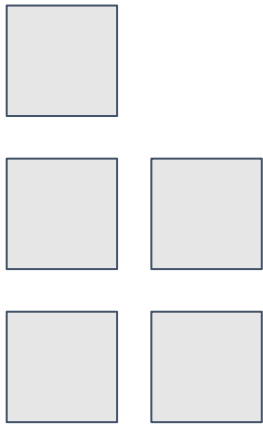
- Media (promedio).
- Mediana (dato central).
- Moda (dato que más se repite).

¿Cómo podrías distribuir equitativamente los cuadros dentro de cada bloque?

7



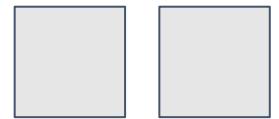
5



6

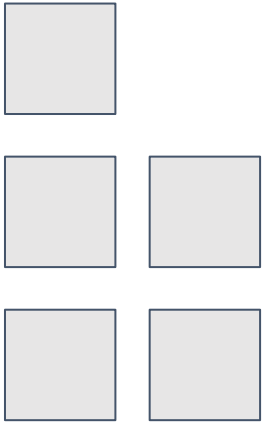


2

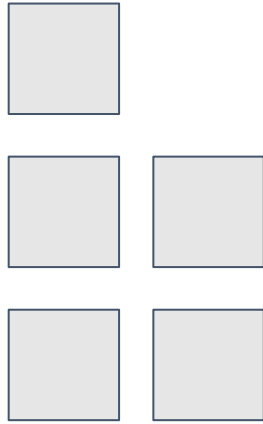


Media (promedio)

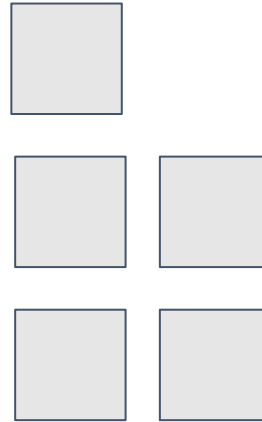
5



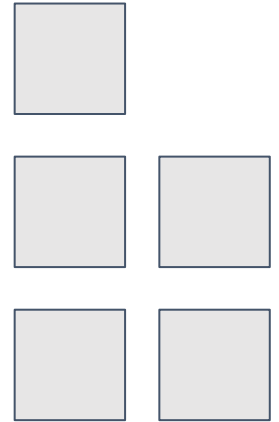
5



5

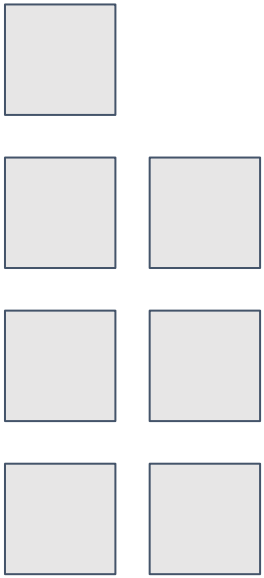


5

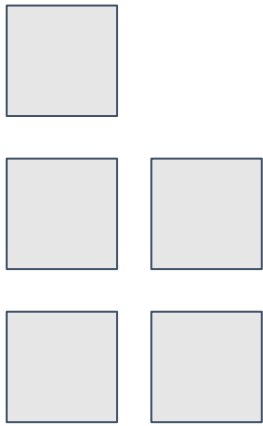


¿Cuál es el valor que divide a los datos?

7



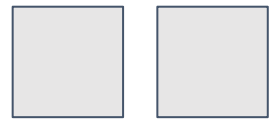
5



6



2





Hay que representarlos y contar

2, 2, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7



¿Cuál es el valor que divide a los datos?

2, 2, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7



50%

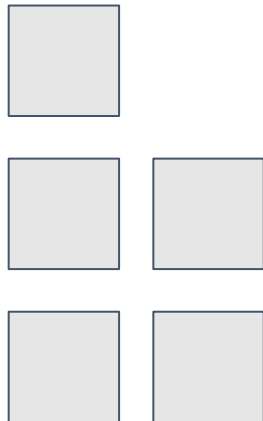
50%

**¿Cuál es el valor que
más se repite en los datos?**

2



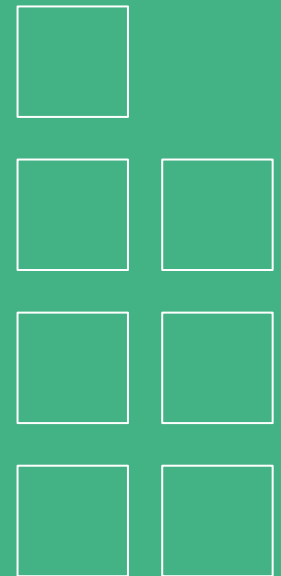
5



6



7





Medidas de tendencia central

- Media (promedio).
- Mediana (dato central).
- Moda (dato que más se repite).
- Media ponderada.
- Media armónica.
- Media geométrica.

Estadística descriptiva aplicada

Medidas de dispersión
y distribuciones

Medidas de dispersión

- **Rango**

La diferencia entre el valor máximo y valor mínimo de los datos.

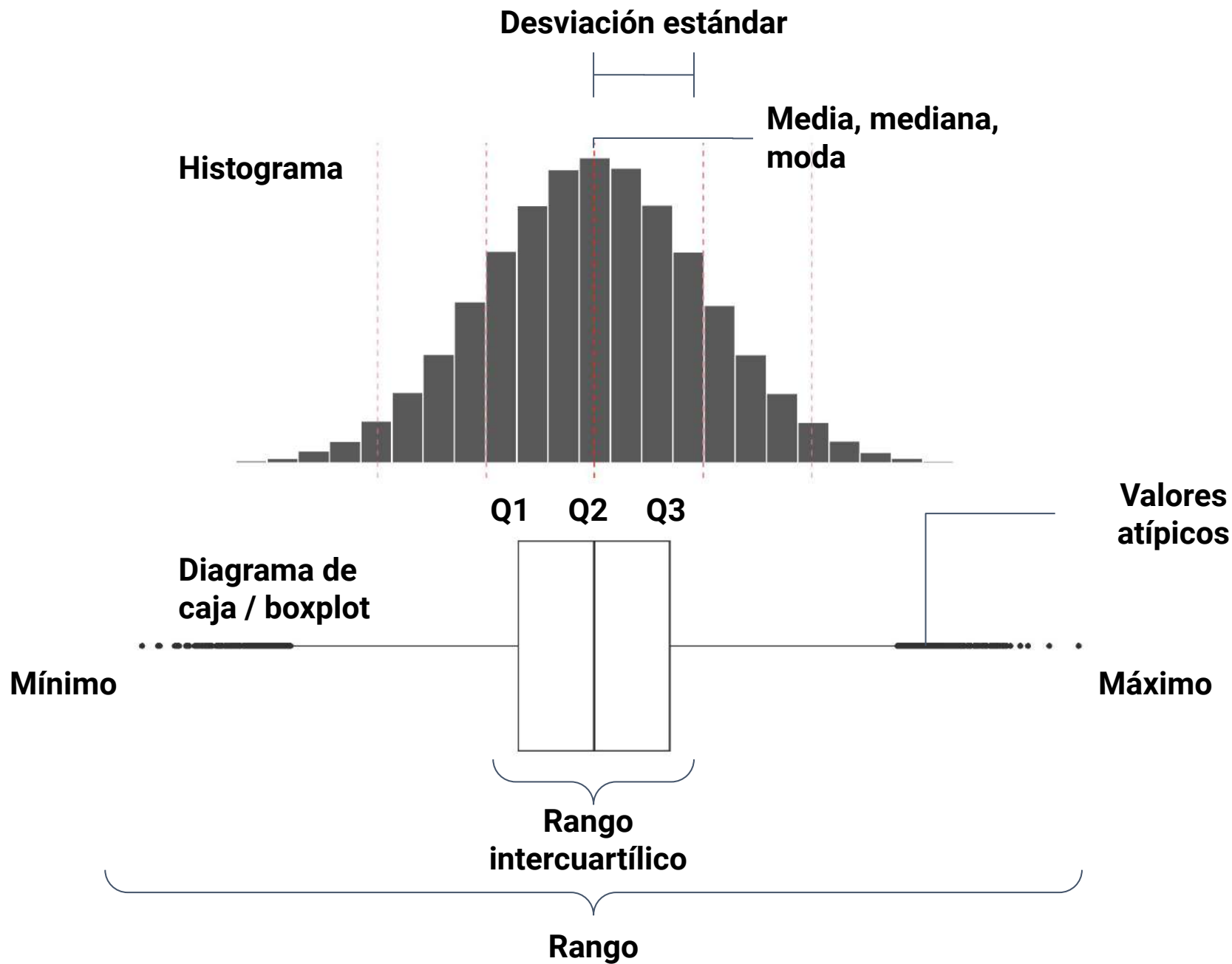
- **Rango intercuartílico**

Comprenden $\pm 25\%$ de los datos respecto a la mediana.



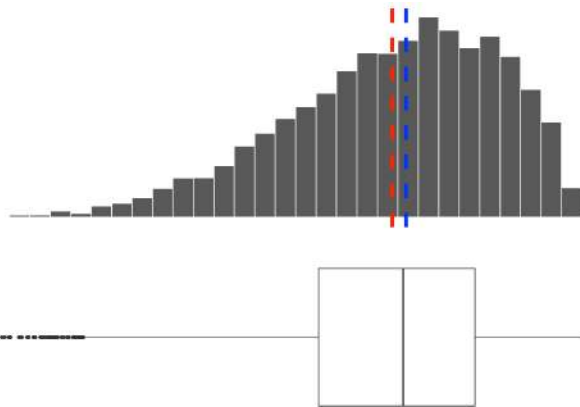
- **Desviación estándar**

Ofrece la dispersión media de una variable.



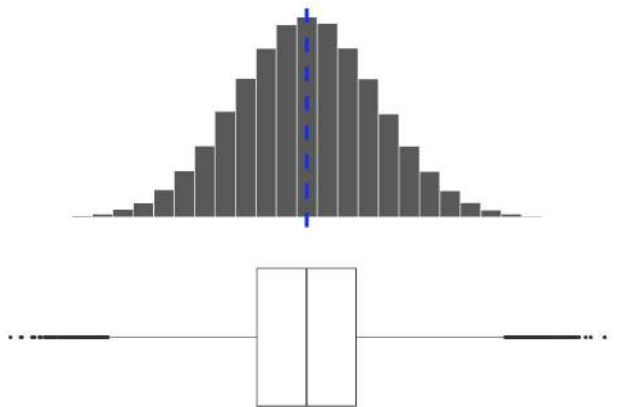
Asimetría estadística

$\text{media} < \text{mediana} < \text{moda}$



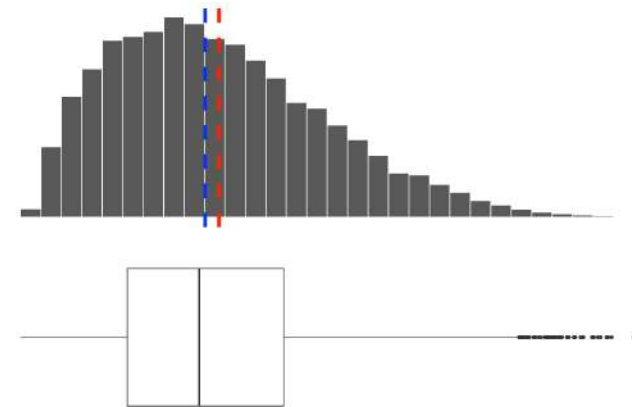
Sesgo negativo

$\text{media} = \text{mediana} = \text{moda}$



Simétrica

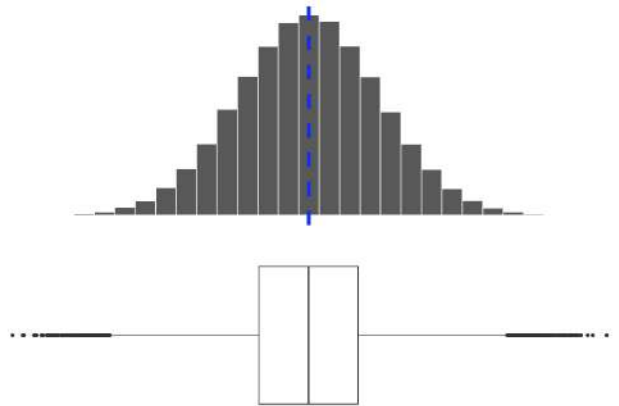
$\text{media} > \text{mediana} > \text{moda}$



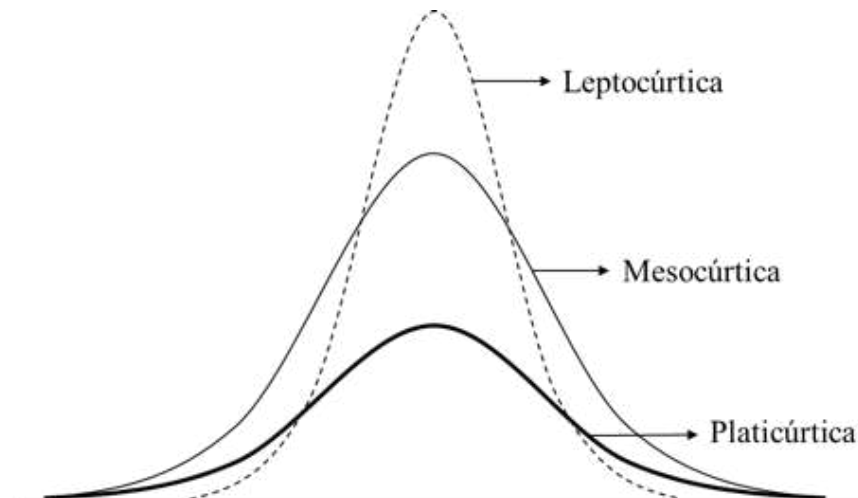
Sesgo positivo

Curtosis

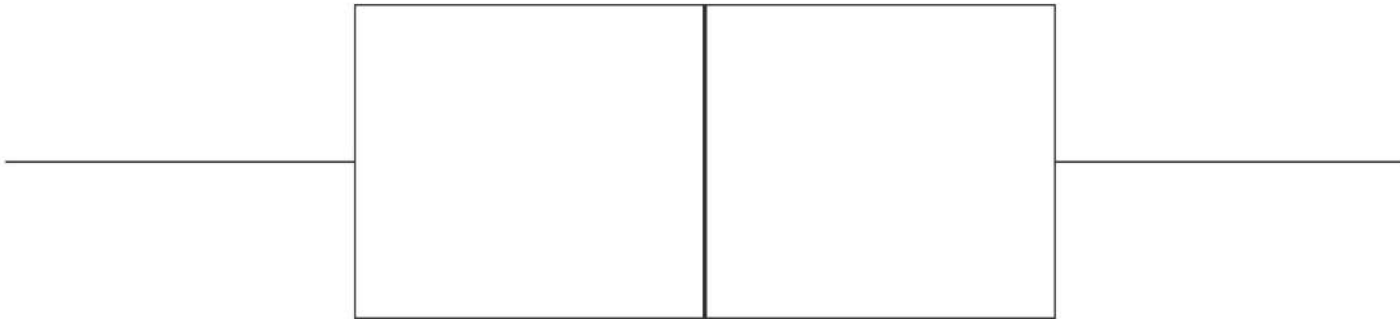
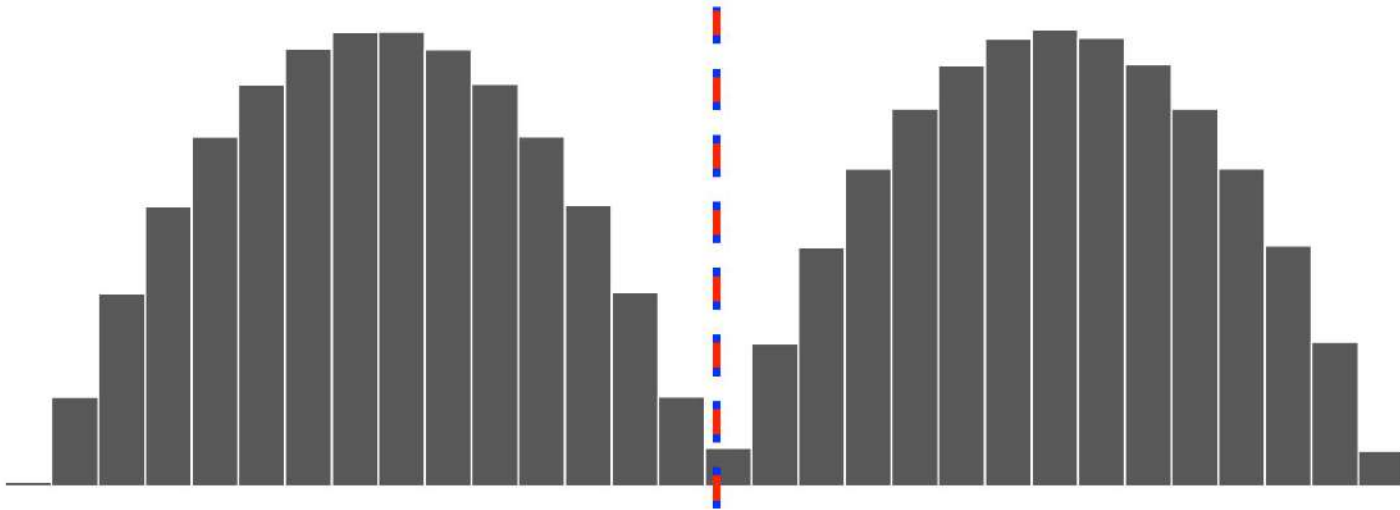
media = mediana = moda



Simétrica



¿Qué observas?




Distribución bimodal

Estadística descriptiva aplicada

Distribuciones

¿Cómo visualizar una distribución?

- Histograma.
- Función de probabilidad de masas (PMFs).
- Función de distribución acumulada (CDFs).
- Función de probabilidad de densidad (PDFs).



Función de probabilidad de masas (PMFs)

Nos dice la probabilidad de que una variable aleatoria discreta tome un valor determinado.



Función de distribución acumulada (CDFs)

Devuelve la probabilidad de que una variable sea igual o menor que un valor determinado.

Función de densidad de probabilidad (PDFs)

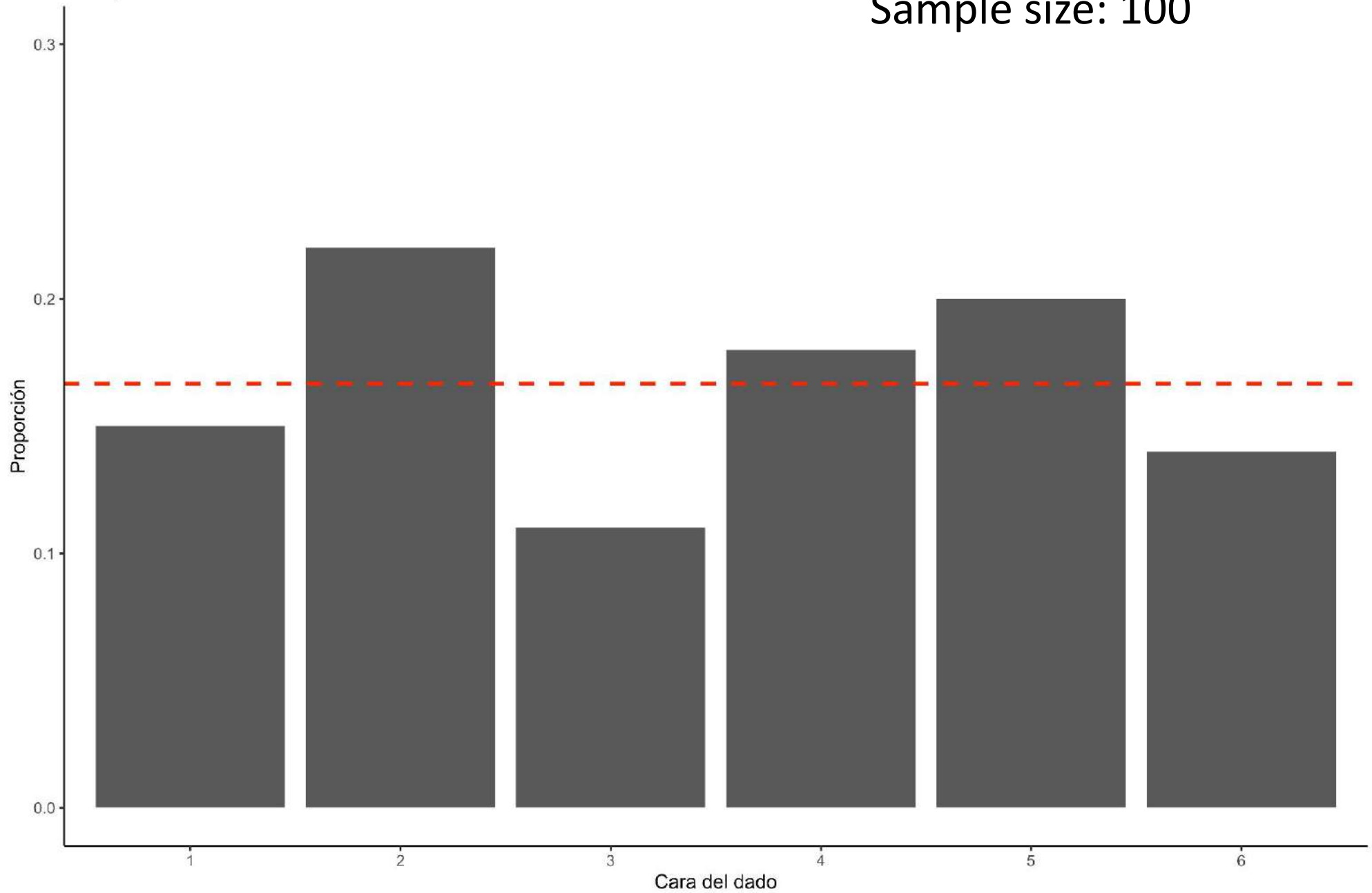
Determina la probabilidad de que una variable continua tome un valor determinado.

Bonus:
**Ley de los Grandes
Números y Teorema
del Límite Central**

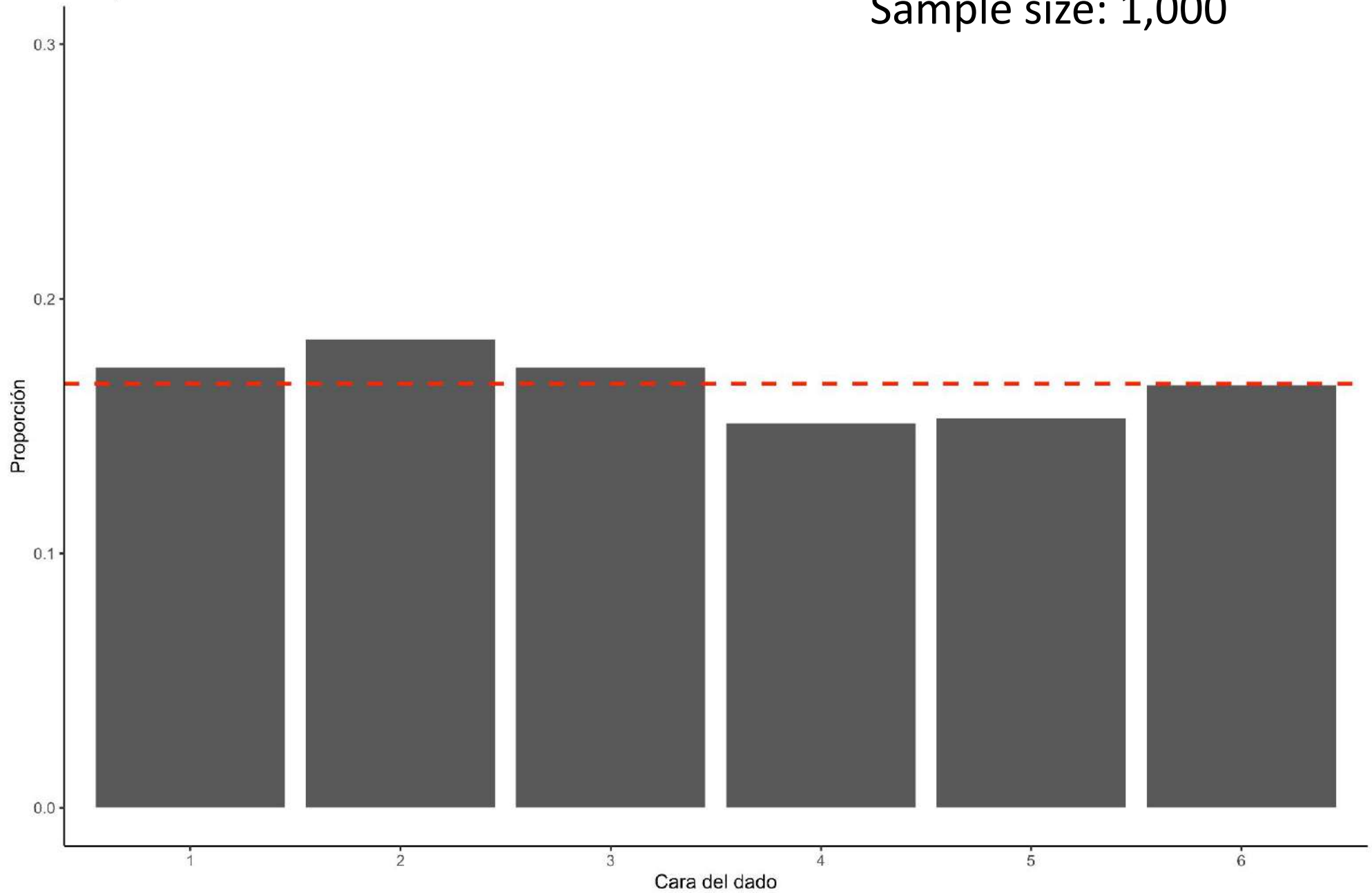
Ley de los Grandes Números



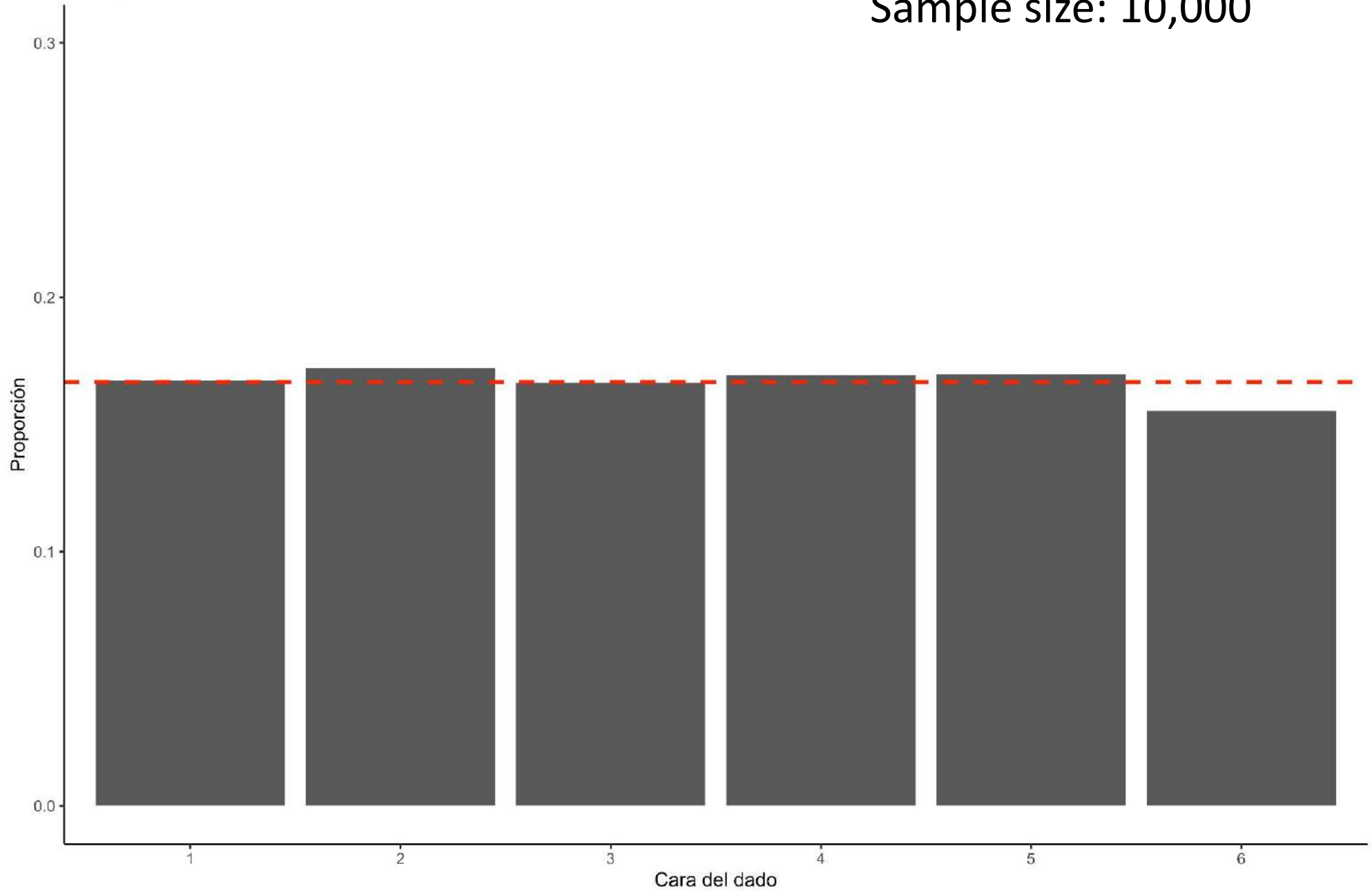
Sample size: 100



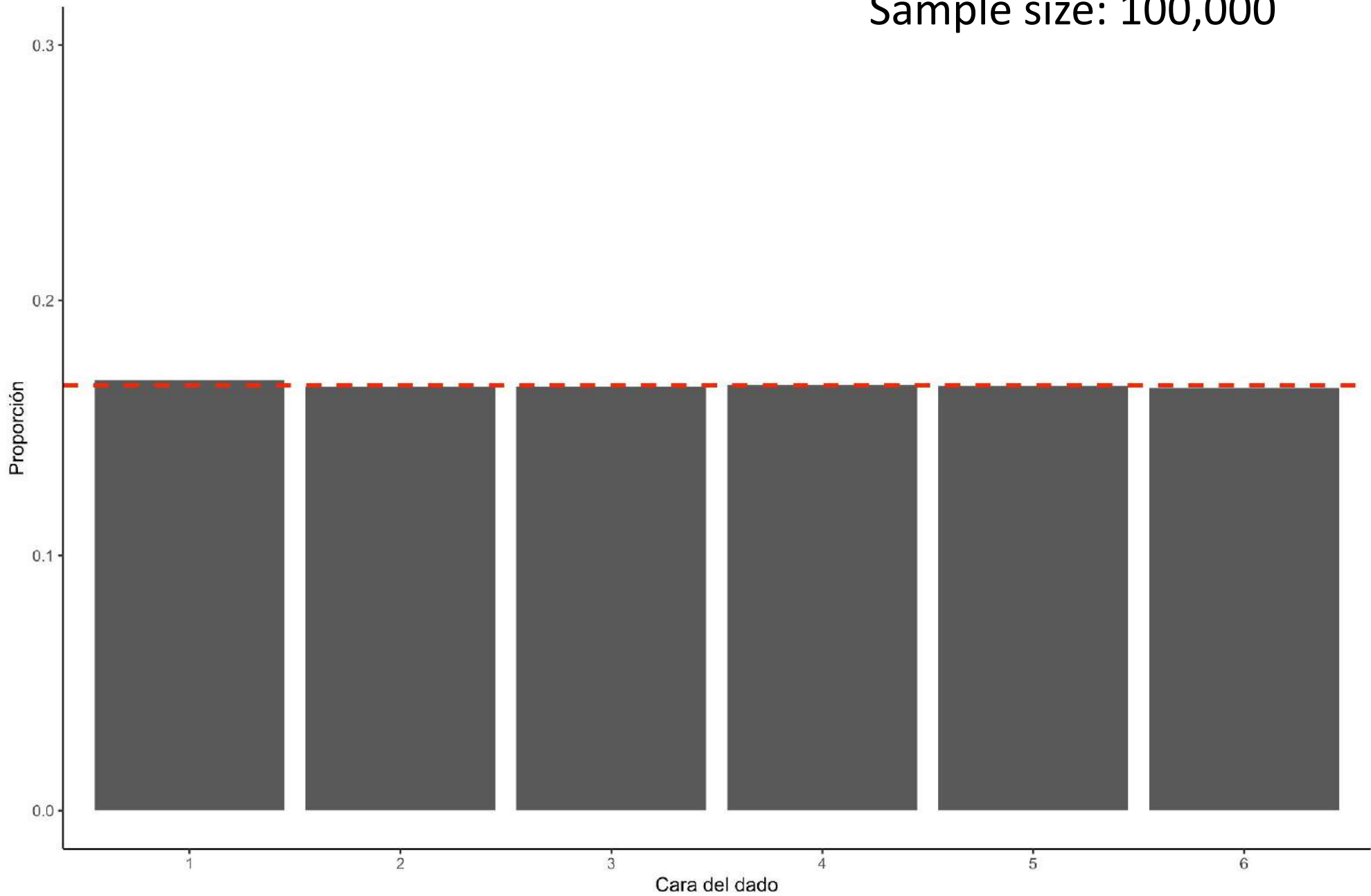
Sample size: 1,000



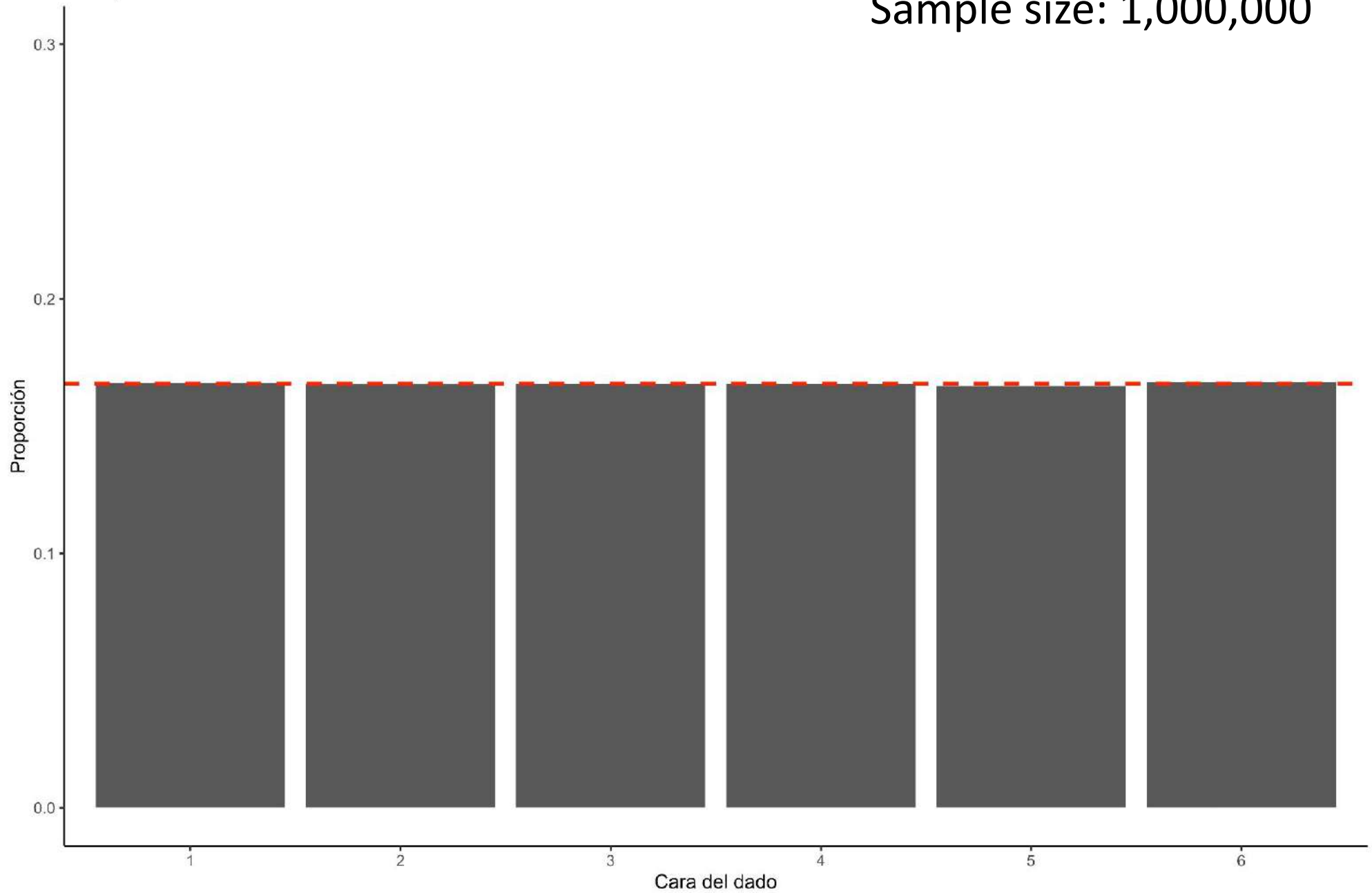
Sample size: 10,000



Sample size: 100,000



Sample size: 1,000,000



“

**La probabilidad experimental
tiende a la probabilidad teórica
a medida que aumenta
el número de repeticiones
del experimento.**

”

Teorema del Límite Central



“

**La media de las muestras
tiende aproximadamente a
una distribución normal.**

”

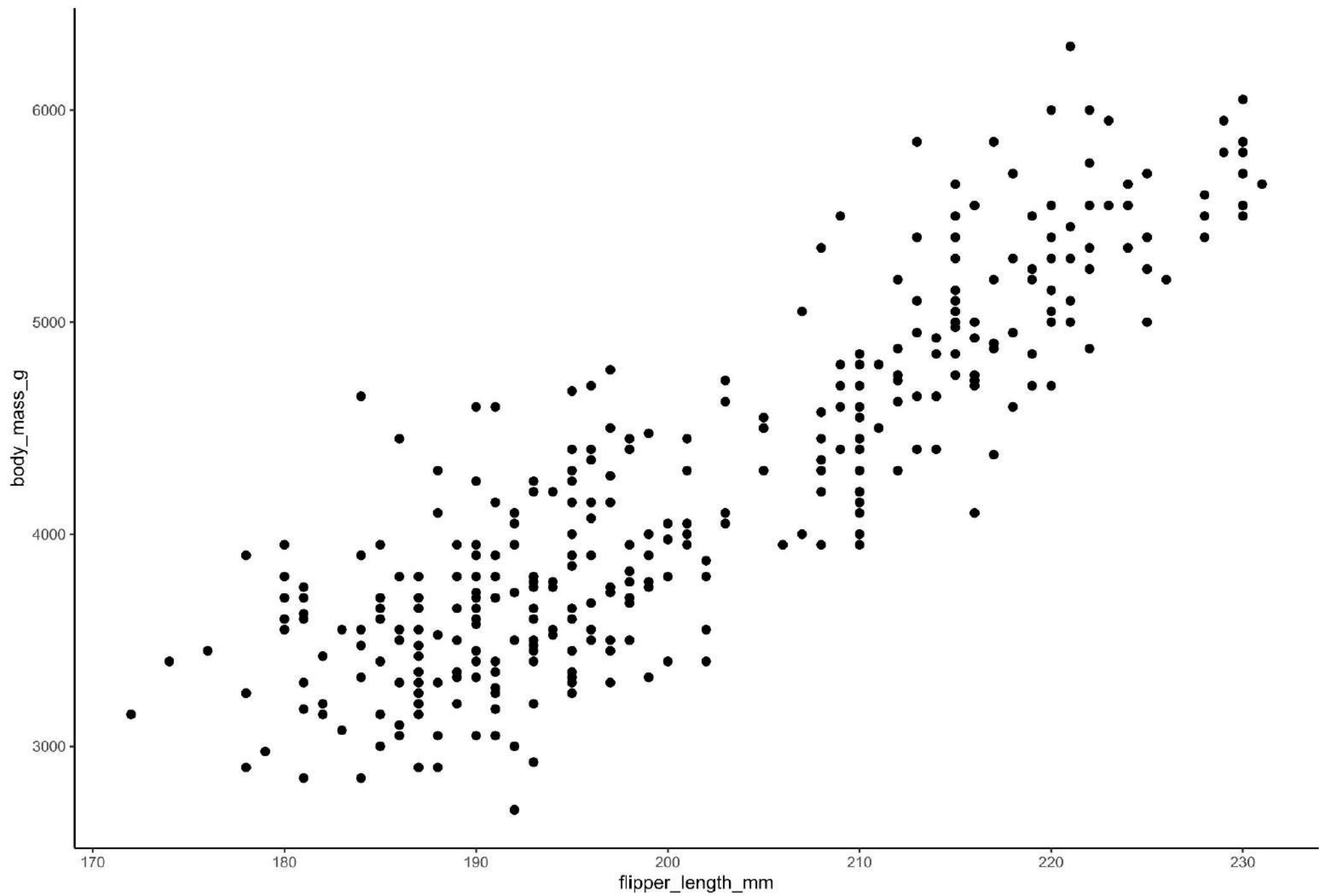
“

**La suma de n variables
aleatorias independientes
con medias y varianzas finitas
converge en distribución
a una variable aleatoria
con distribución normal.**

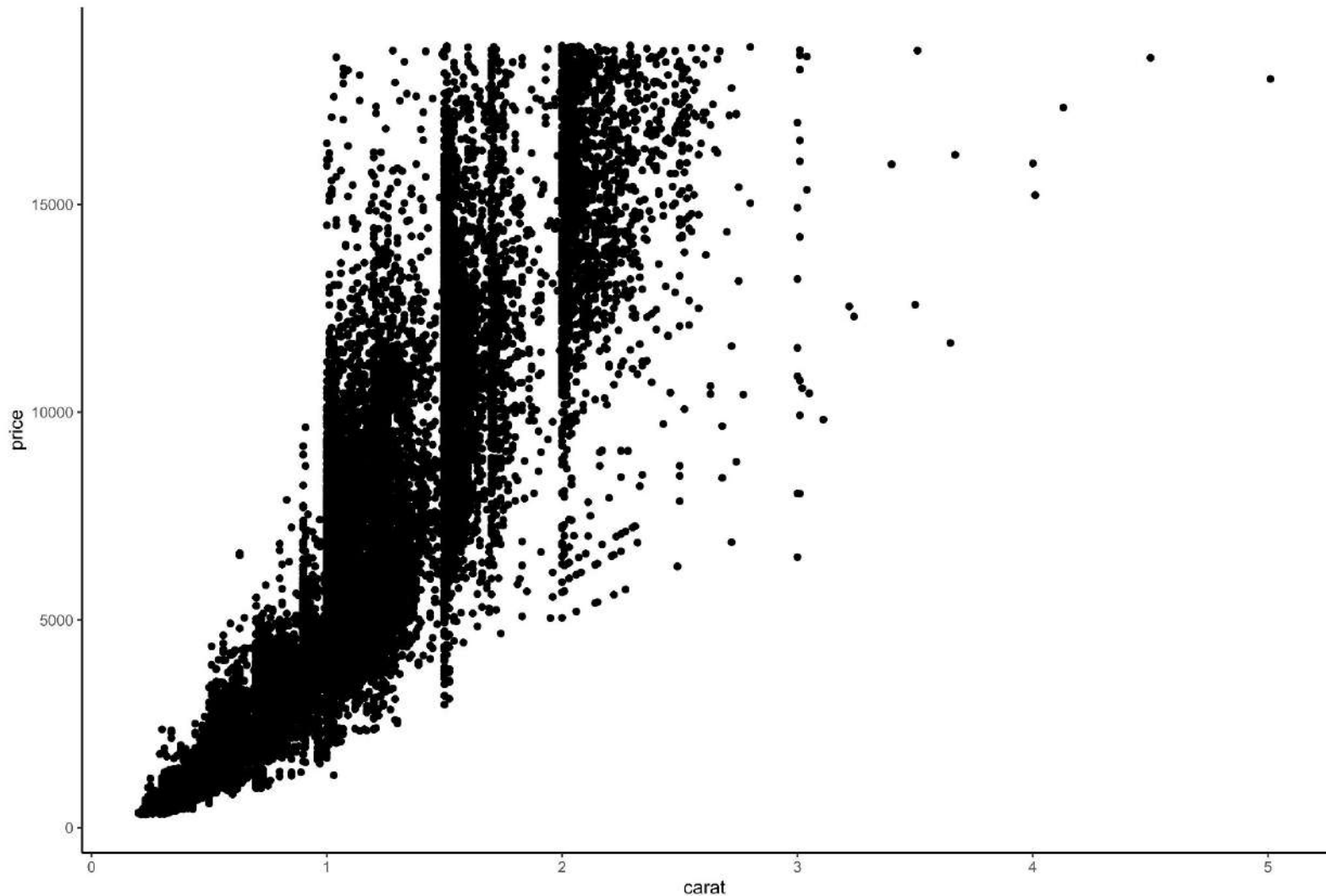
”

Estableciendo relaciones

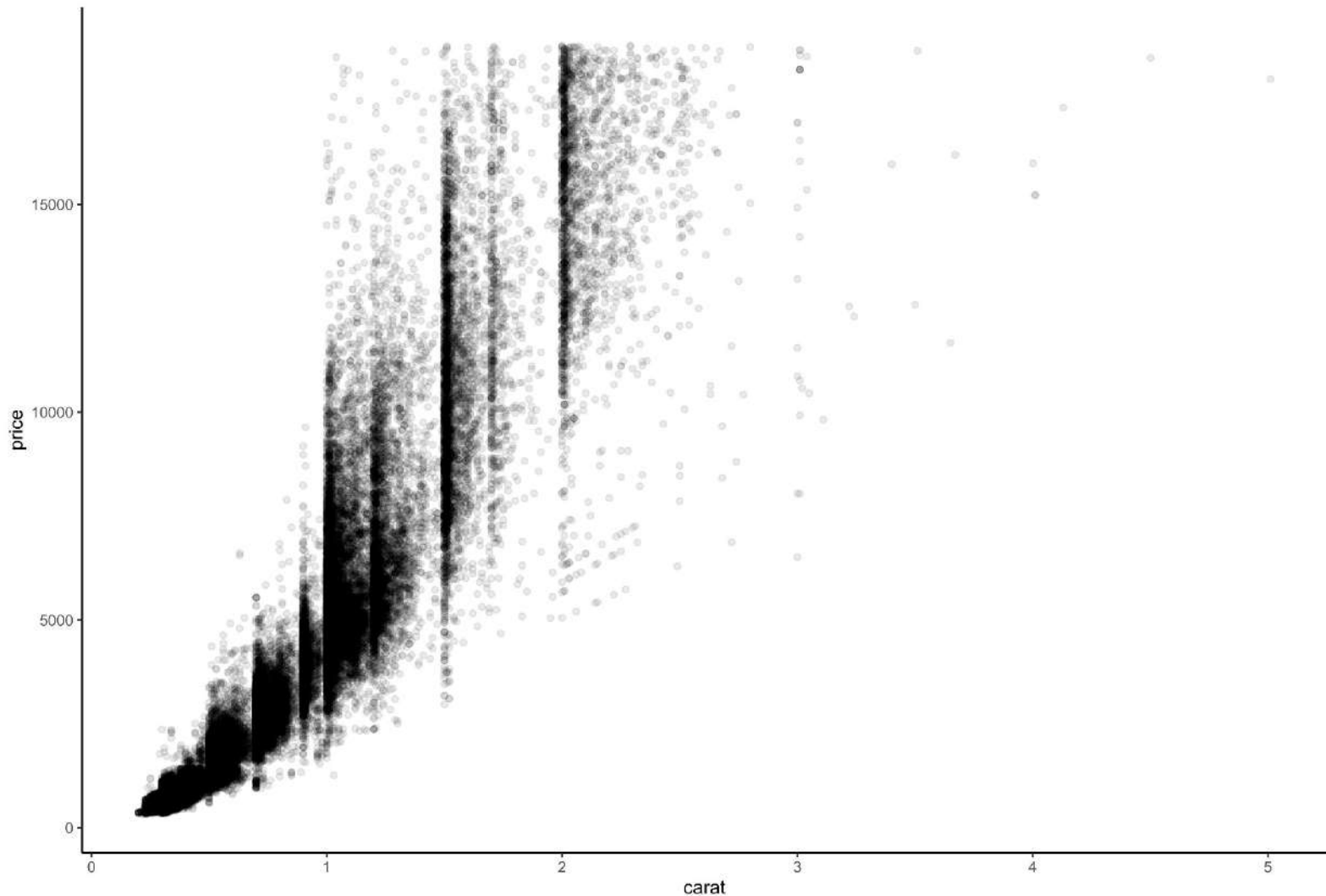
Gráficos de puntos



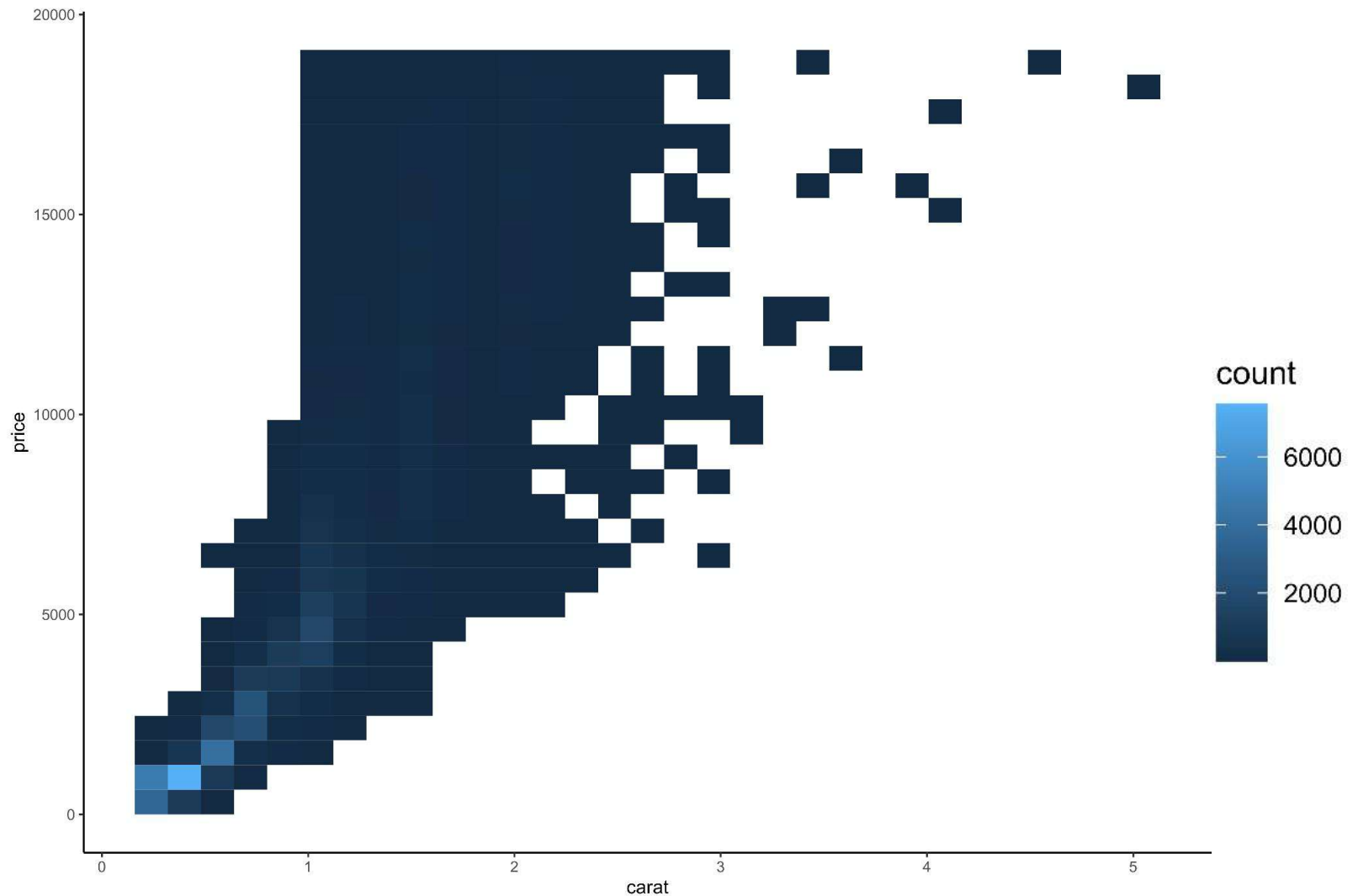
Un caso que no es bueno a primera vista



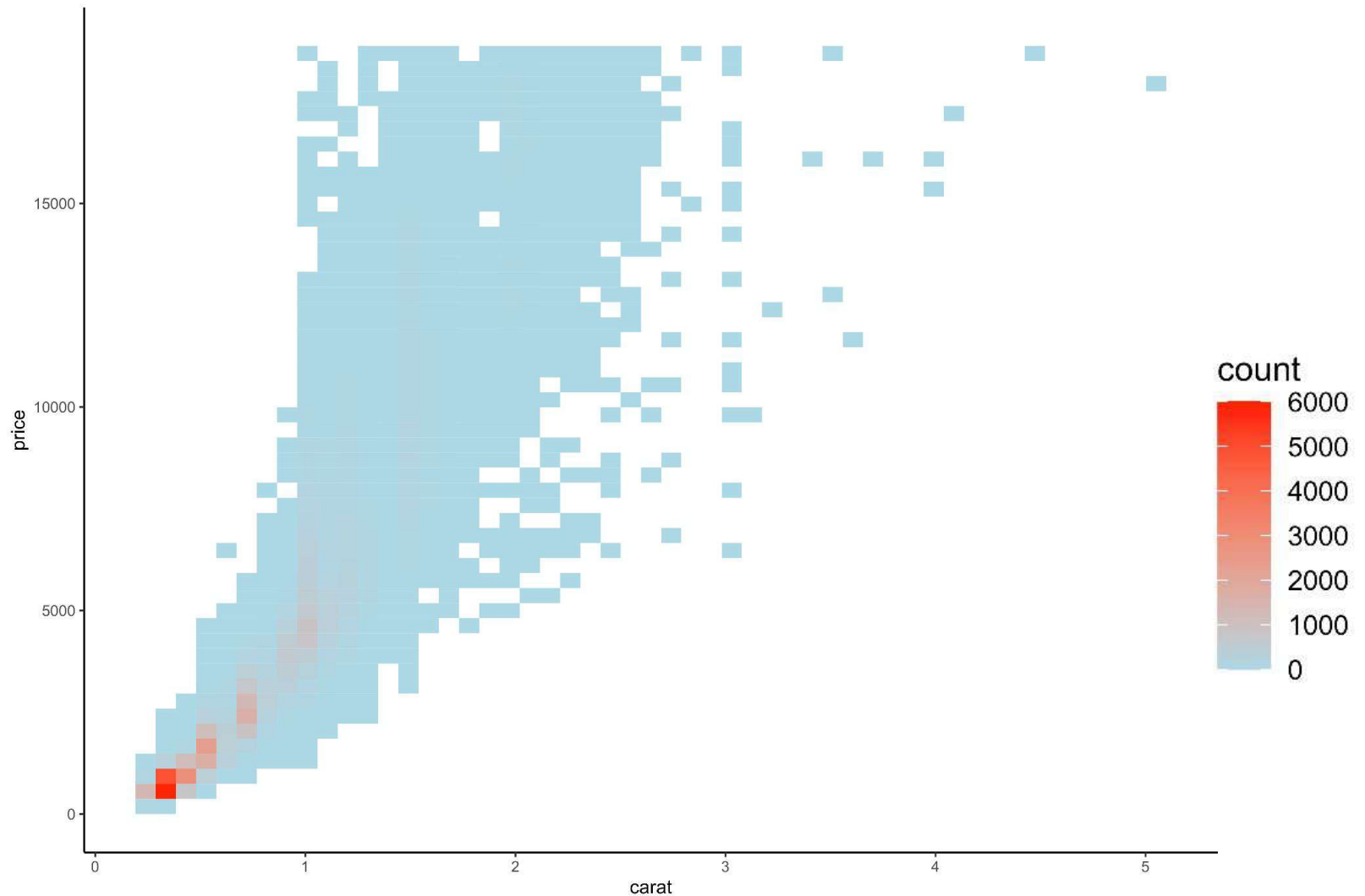
Modificar la transparencia



Histograma de dos dimensiones



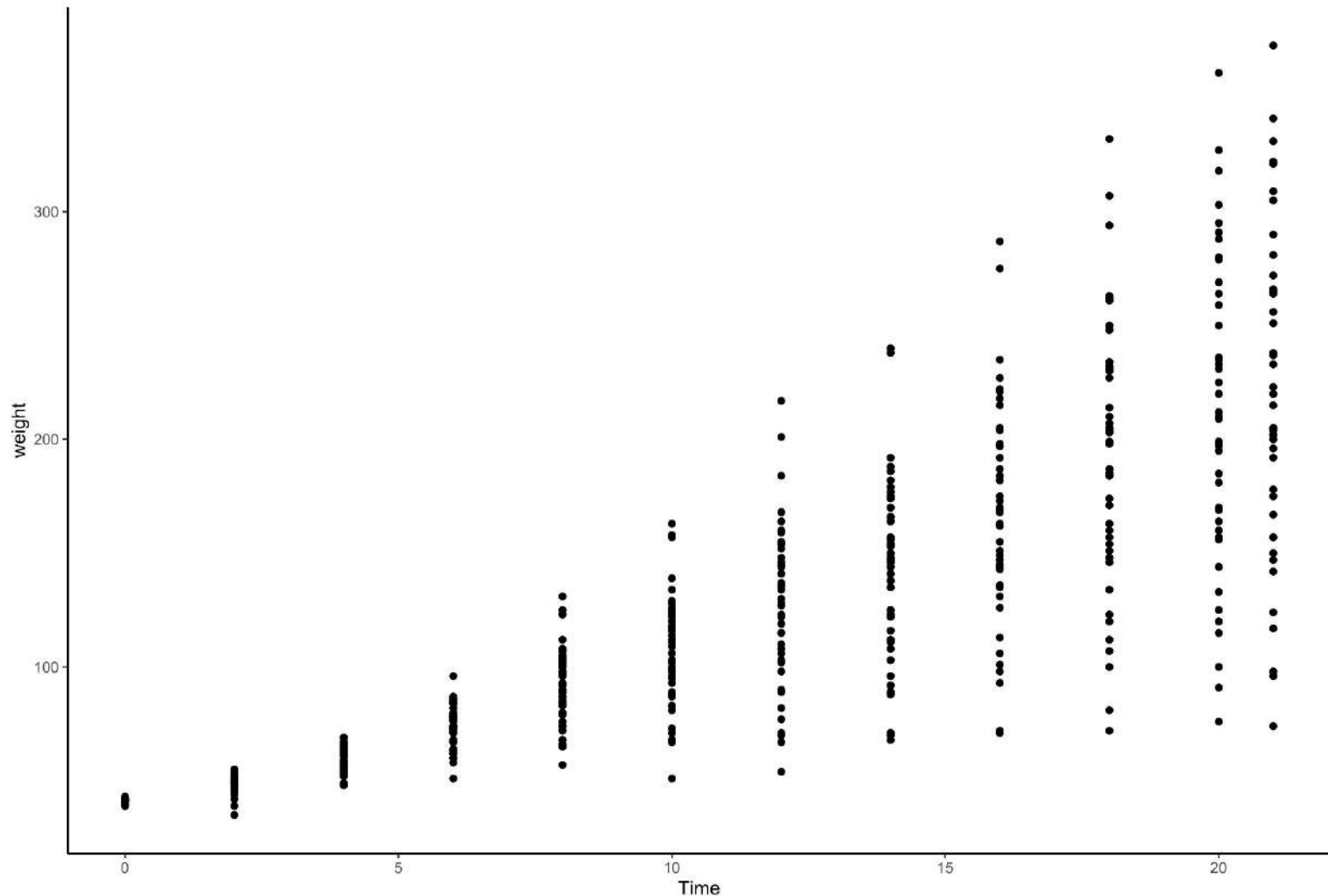
Cambio de color



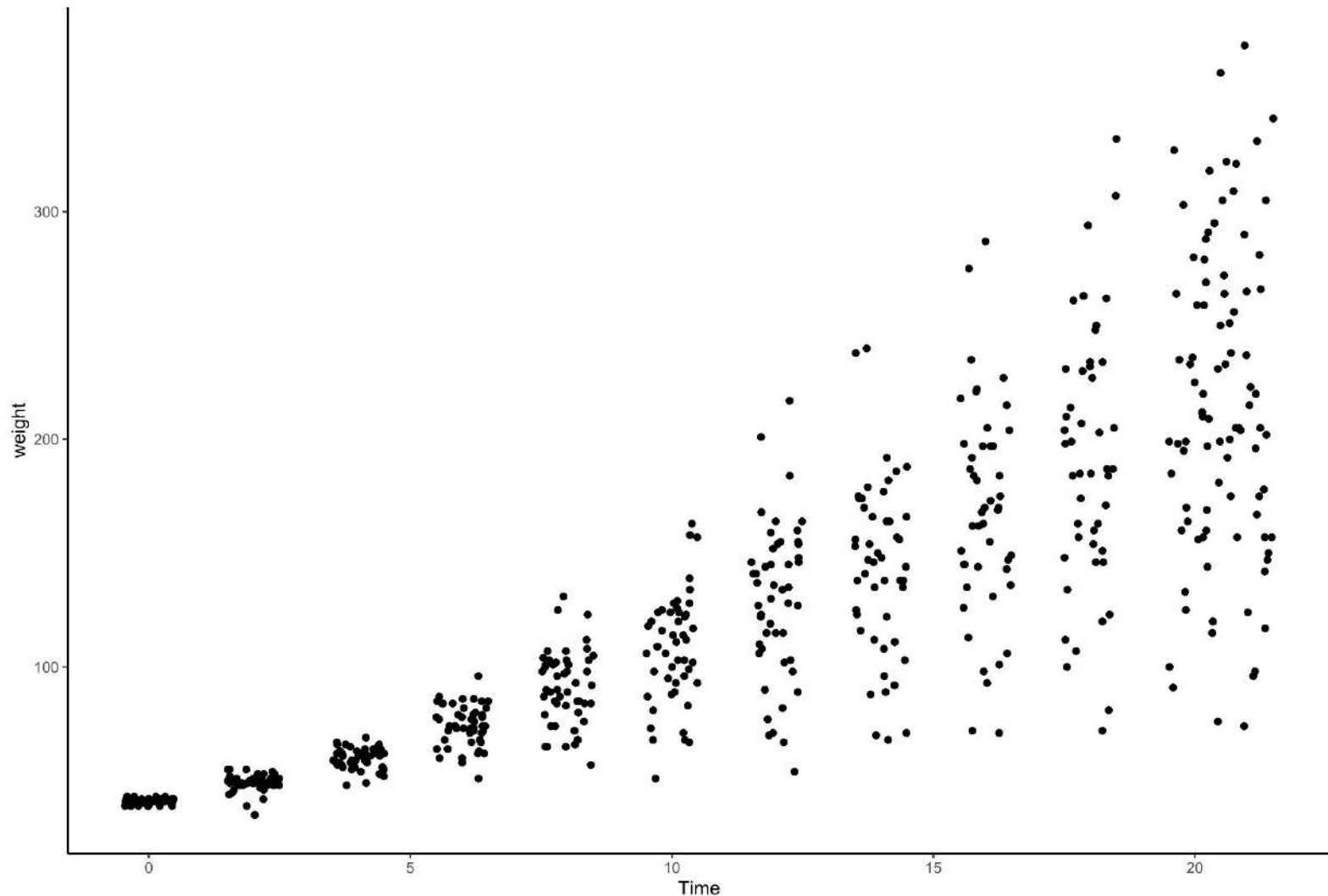
Estableciendo relaciones

Gráficos de violín y boxplots

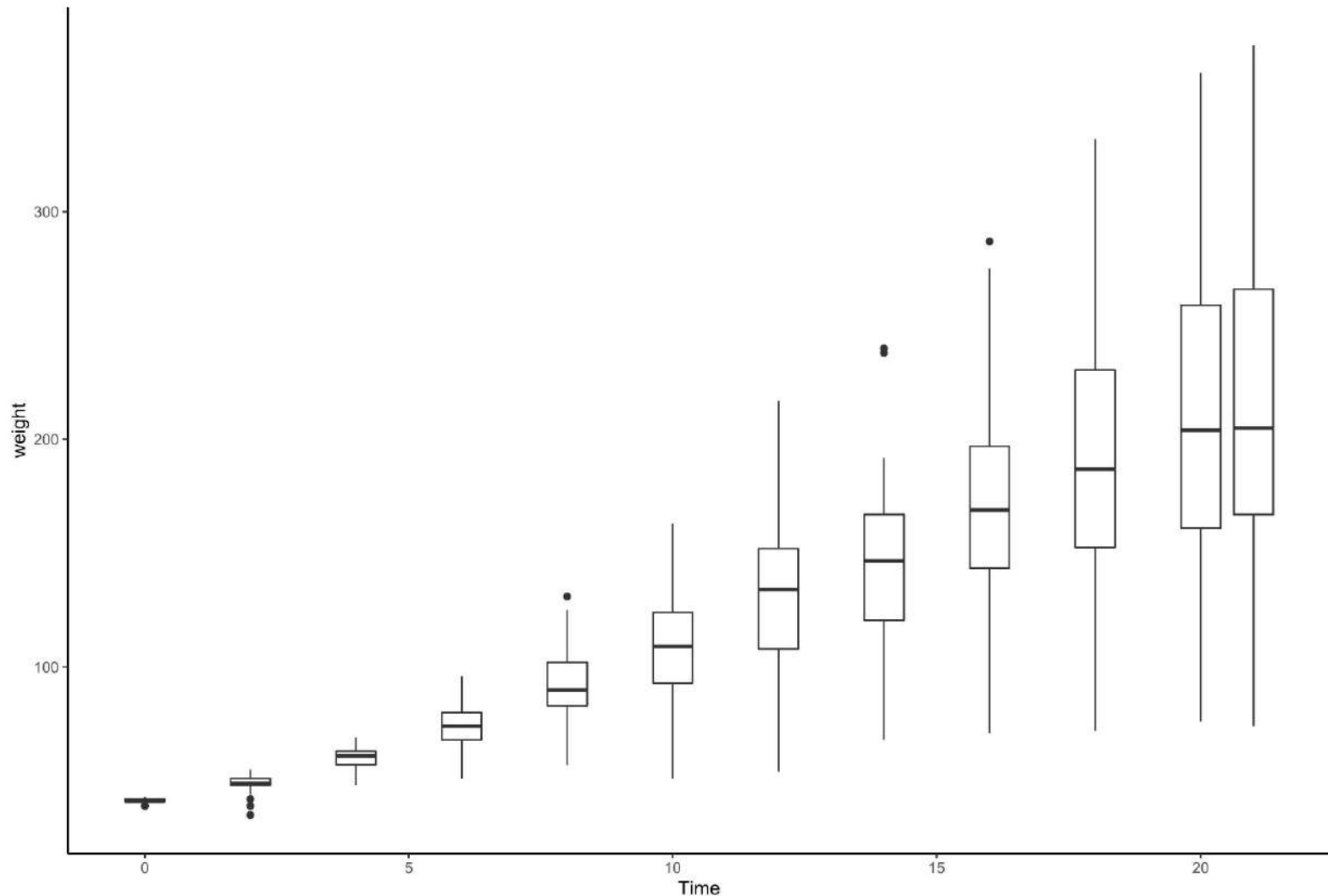
¿Qué pasa si tengo una variable discreta?



¿Qué pasa si tengo una variable discreta?



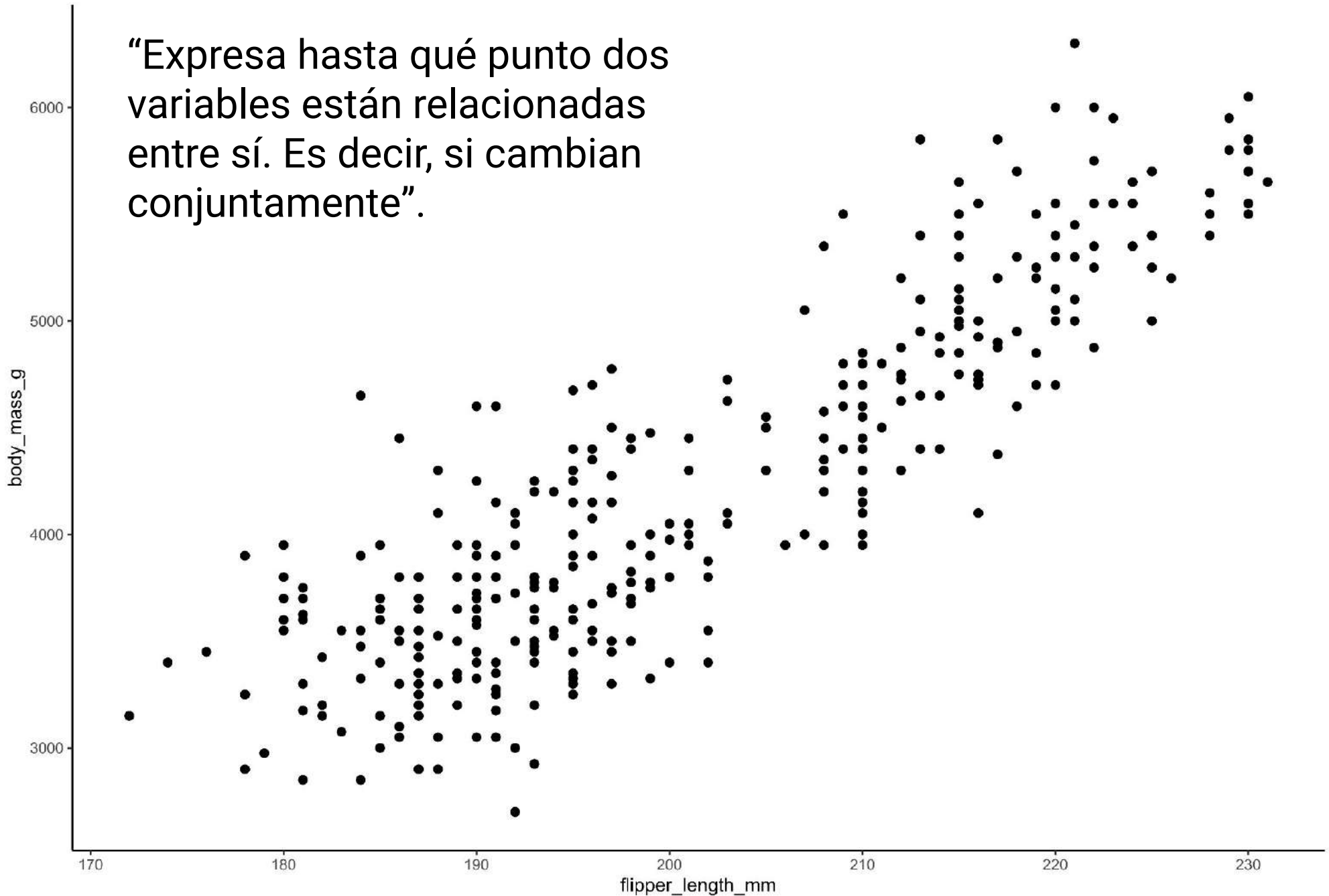
¿Qué pasa si tengo una variable discreta?



Estableciendo relaciones

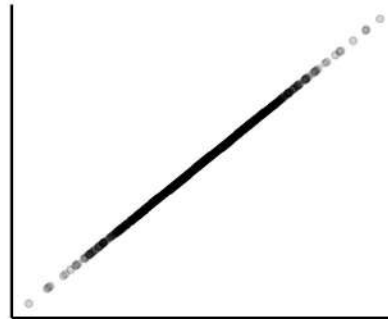
Matrices de correlación

“Expresa hasta qué punto dos variables están relacionadas entre sí. Es decir, si cambian conjuntamente”.



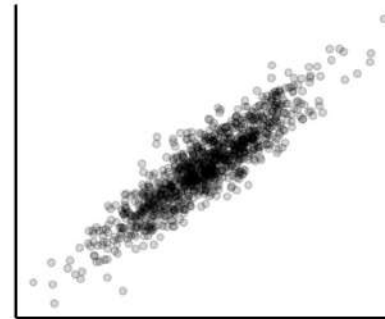
Coeficiente de correlación:
cuantifica la
**intensidad de la
relación lineal**
entre dos
variables en un
análisis de
correlación.

Correlación positiva perfecta



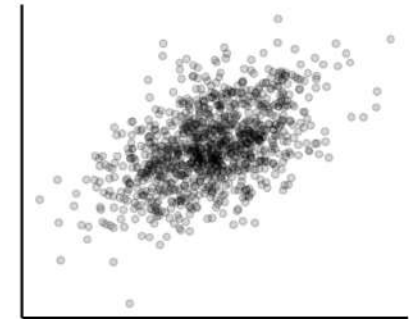
Correlación de Pearson (r): 1.0

Correlación positiva alta



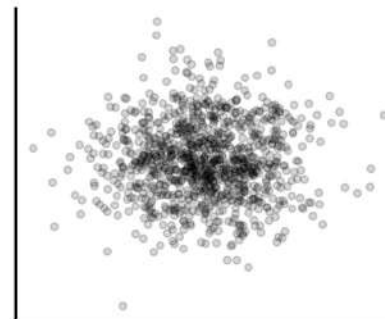
Correlación de Pearson (r): 0.9

Correlación positiva baja



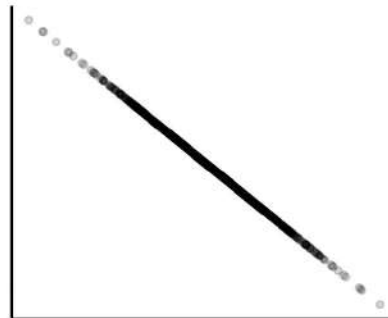
Correlación de Pearson (r): 0.5

Sin correlación



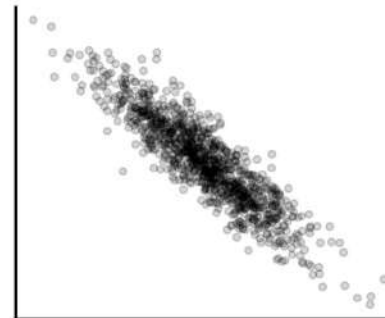
Correlación de Pearson (r): 0

Correlación negativa perfecta



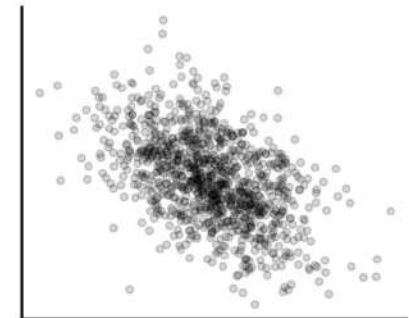
Correlación de Pearson (r): -1.0

Correlación negativa alta



Correlación de Pearson (r): -0.9

Correlación negativa baja



Correlación de Pearson (r): -0.5

Correlación no implica causalidad

Causalidad

Cuando algo (la causa)
genera otra cosa (efecto)



Causa

Efecto

Correlación no implica causalidad

Correlación

Cuando dos o más eventos
aparentan estar relacionados

En verano:



Sube la venta
de helados.



Suben los casos
de quemaduras
de sol.

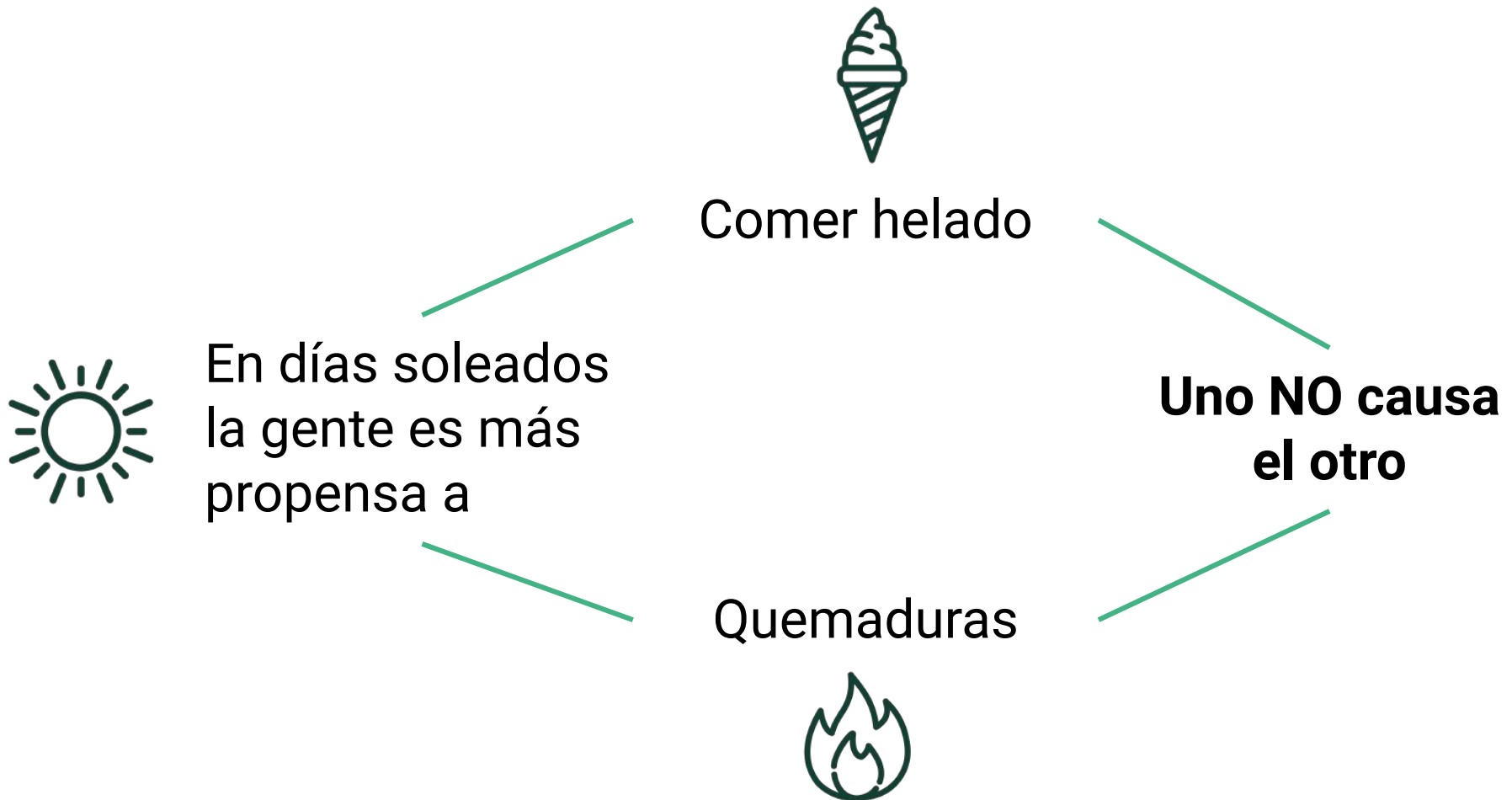


Las ventas
de helado y
los casos de
quemaduras están
correlacionadas.



¿Significa esto
que comer
helado aumenta
el riesgo de
sufrir
quemaduras?

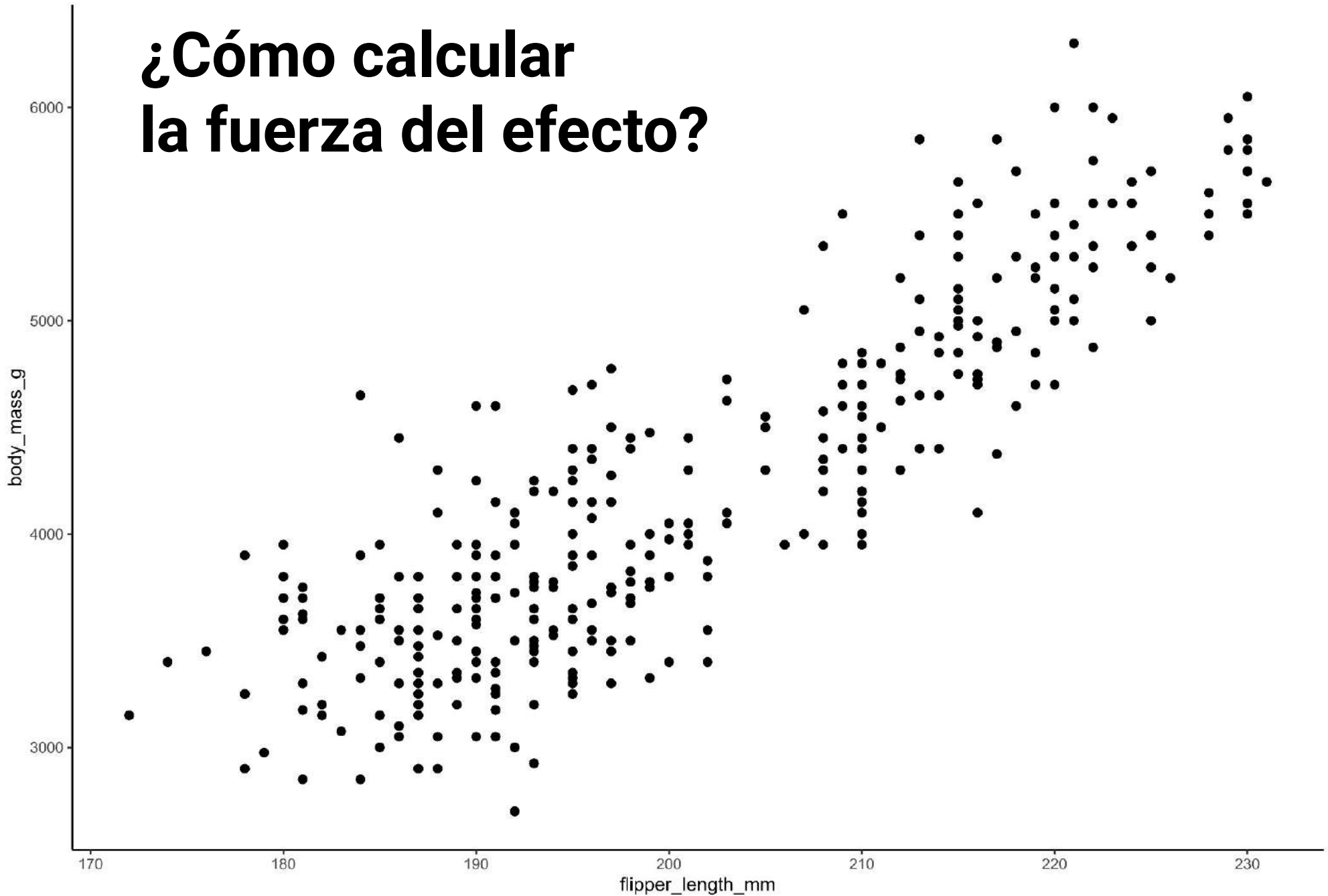
Correlación no implica causalidad



Estableciendo relaciones

Análisis de regresión simple

¿Cómo calcular la fuerza del efecto?



Limitaciones del análisis de regresión simple

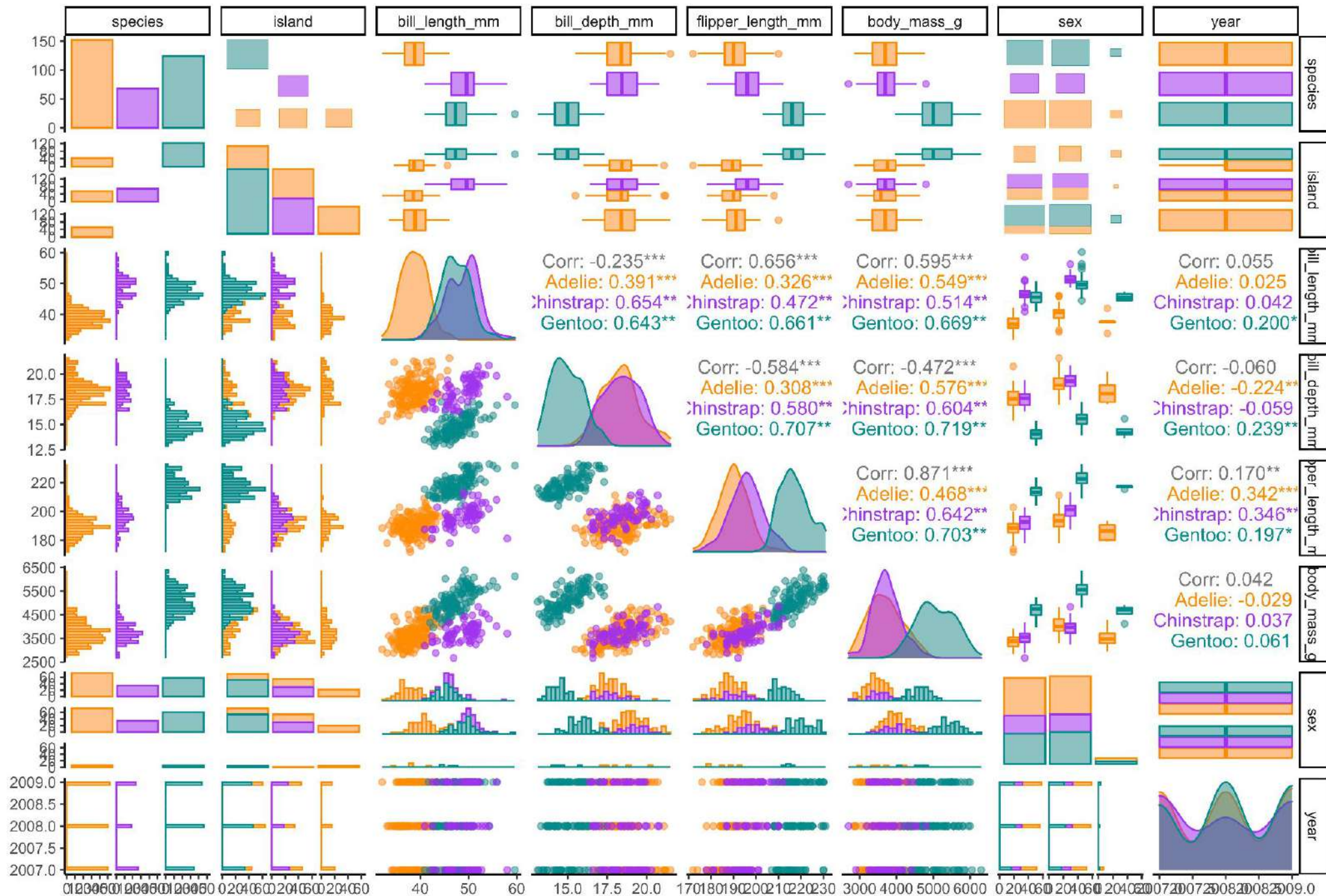
Limitaciones

- La regresión lineal simple de $A - B$ no es lo mismo que la de $B - A$.
- Si dos variables crecen o decrecen siguiendo las mismas pautas, no implica necesariamente que una cause la otra.
- Solo puede manejar relaciones lineales.

Análisis de regresión múltiple

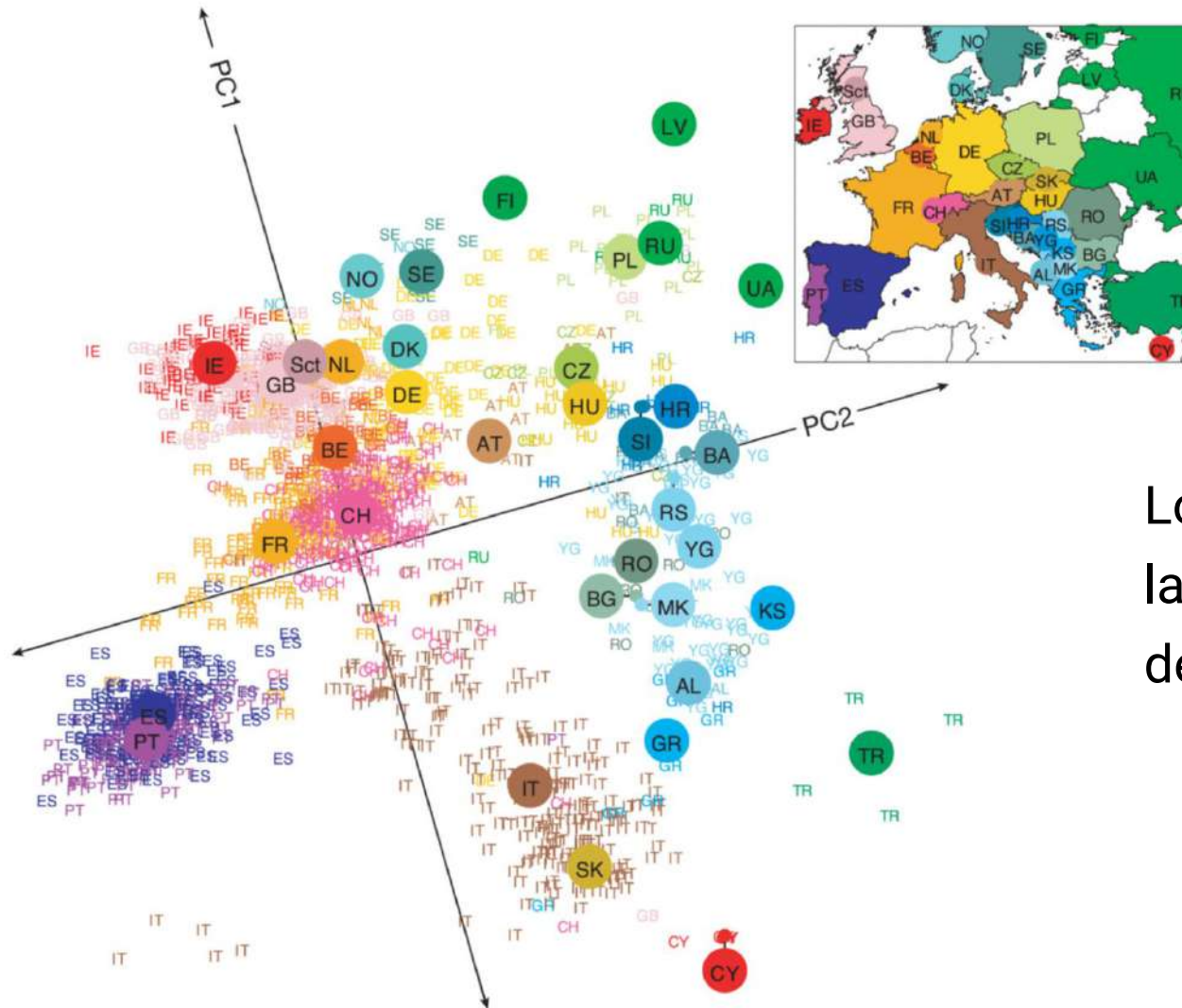
Análisis de regresión logística

**¿Qué hacer
cuando tengo
muchas variables?**



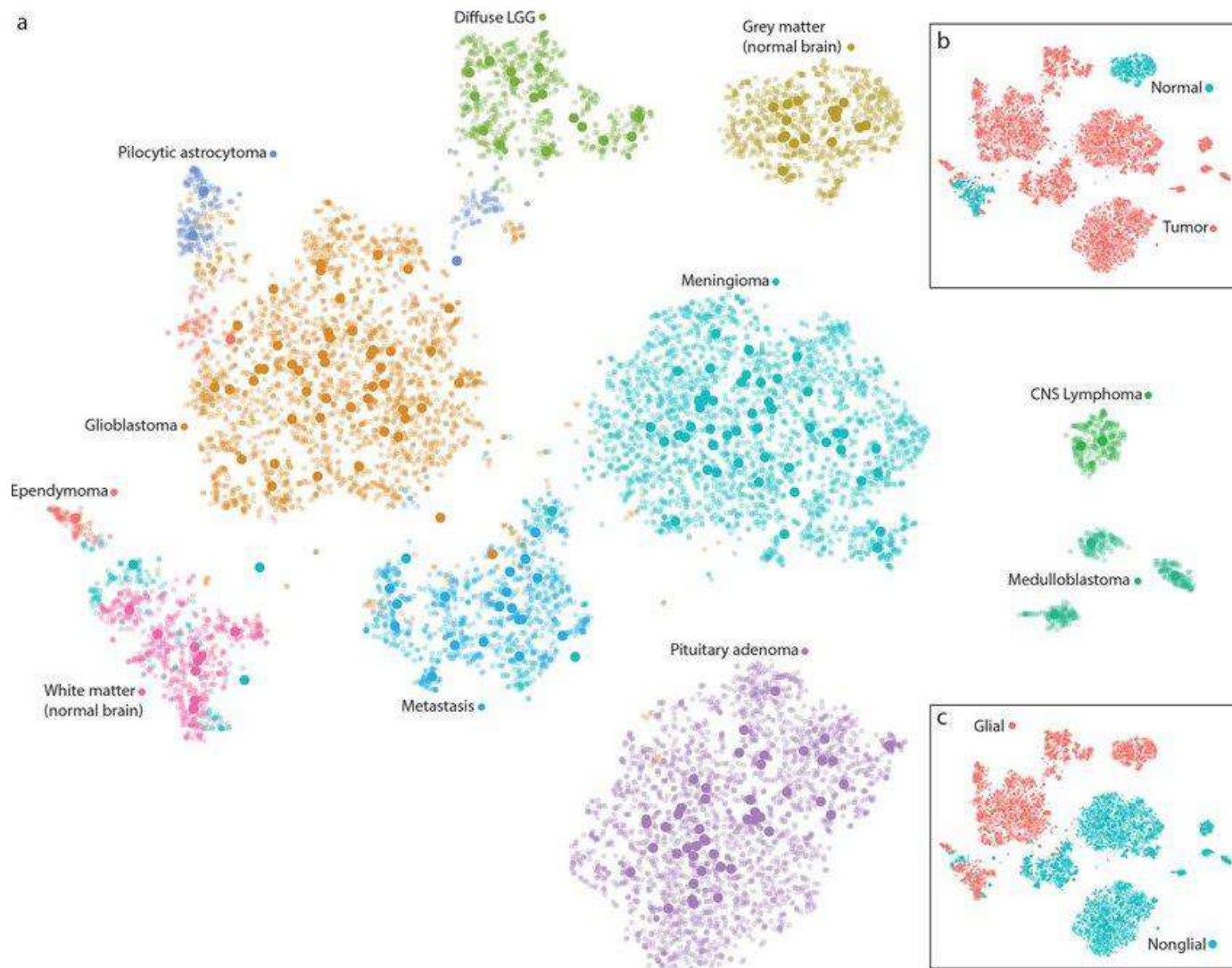
Análisis de Componentes Principales (PCA)

a

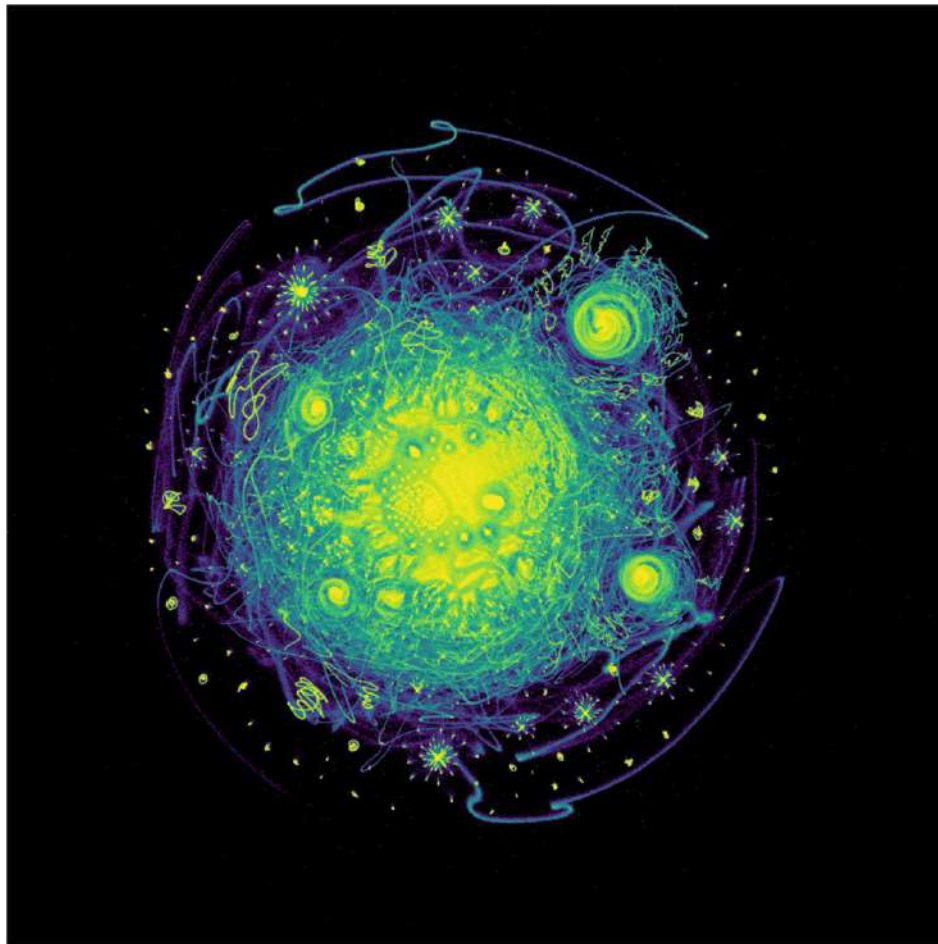


Los genes reflejan la geografía dentro de Europa.

TSNE (T-distributed Stochastic Neighbor Embedding)

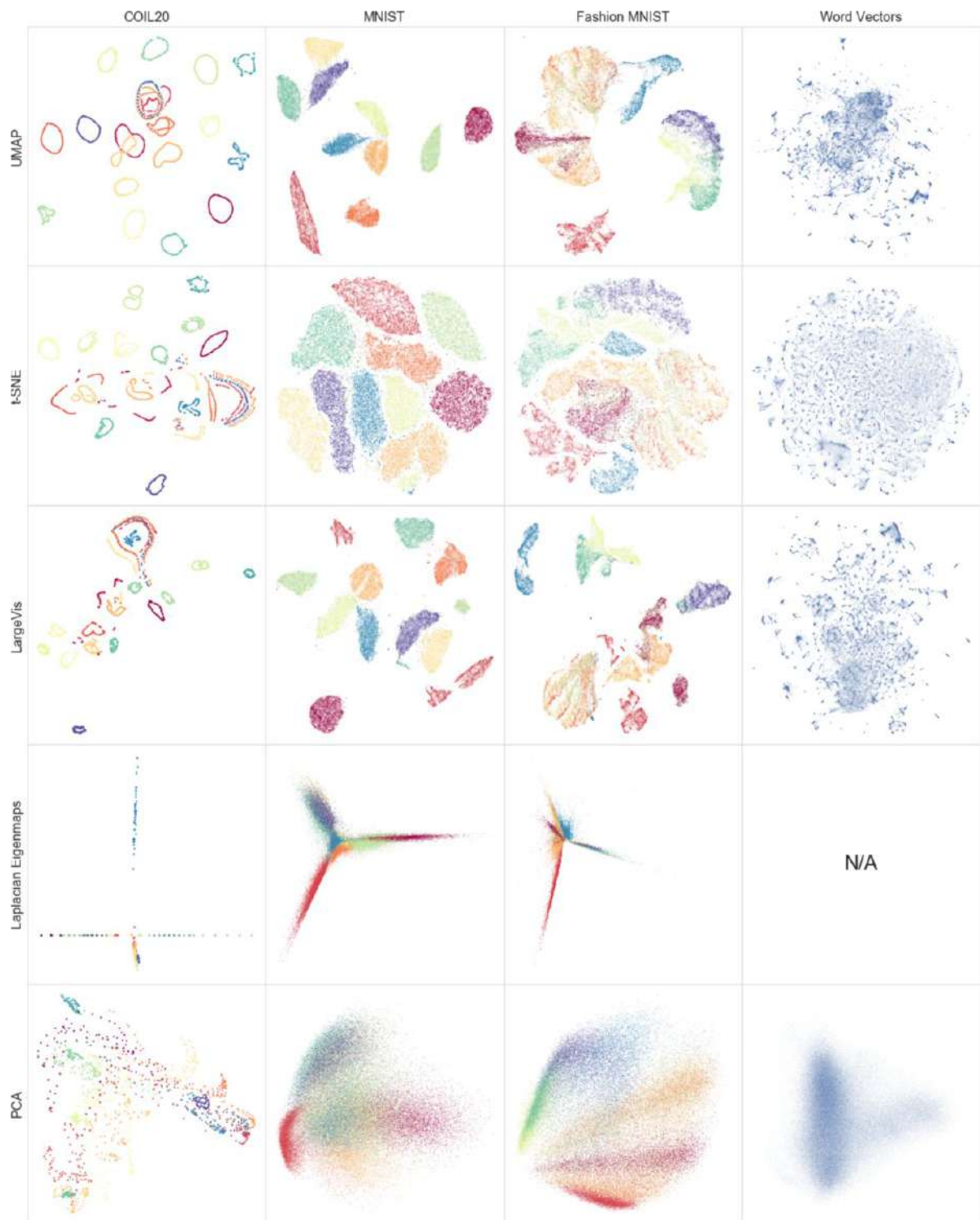


UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction



Visualización de 30,000,000 de números enteros representados por vectores binarios divisibilidad prima, coloreados por densidad de puntos.

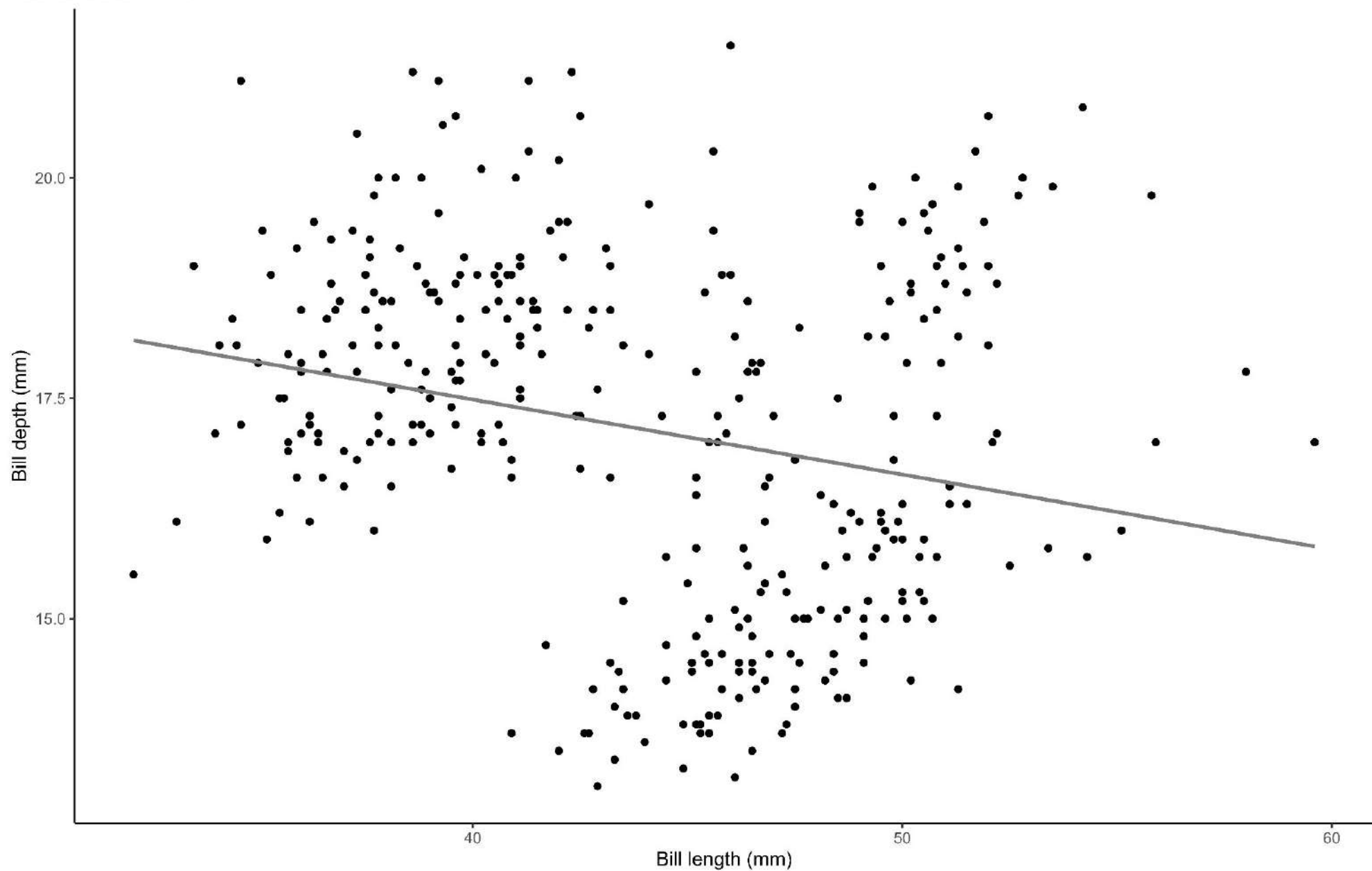
**Comparación:
algoritmo
de reducción
de dimensión vs.
conjunto de datos.**



Paradoja de Simpson

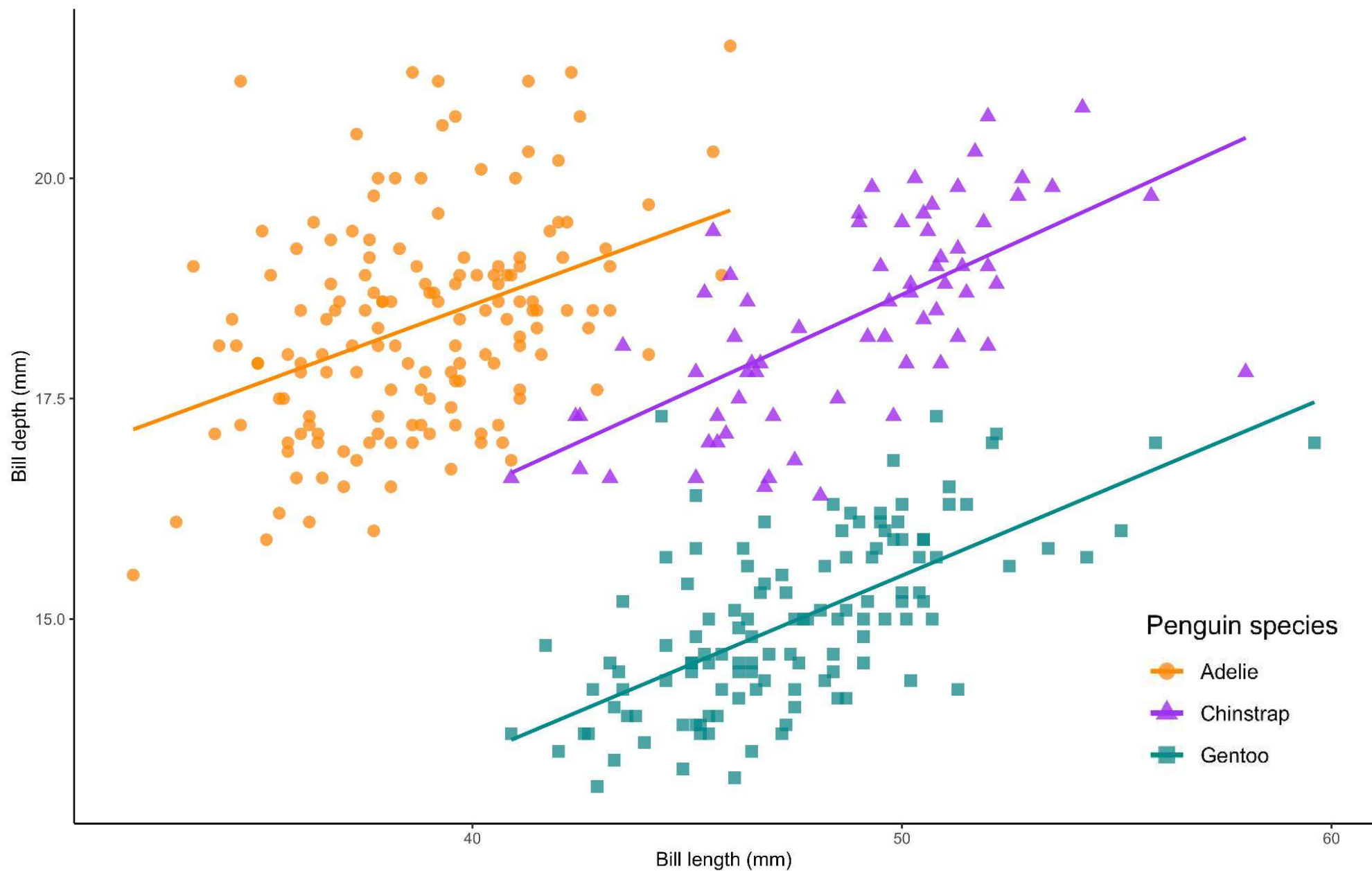
Penguin bill dimensions (omit species)

Palmer Station LTER



Penguin bill dimensions

Bill length and depth for Adelie, Chinstrap and Gentoo Penguins at Palmer Station LTER

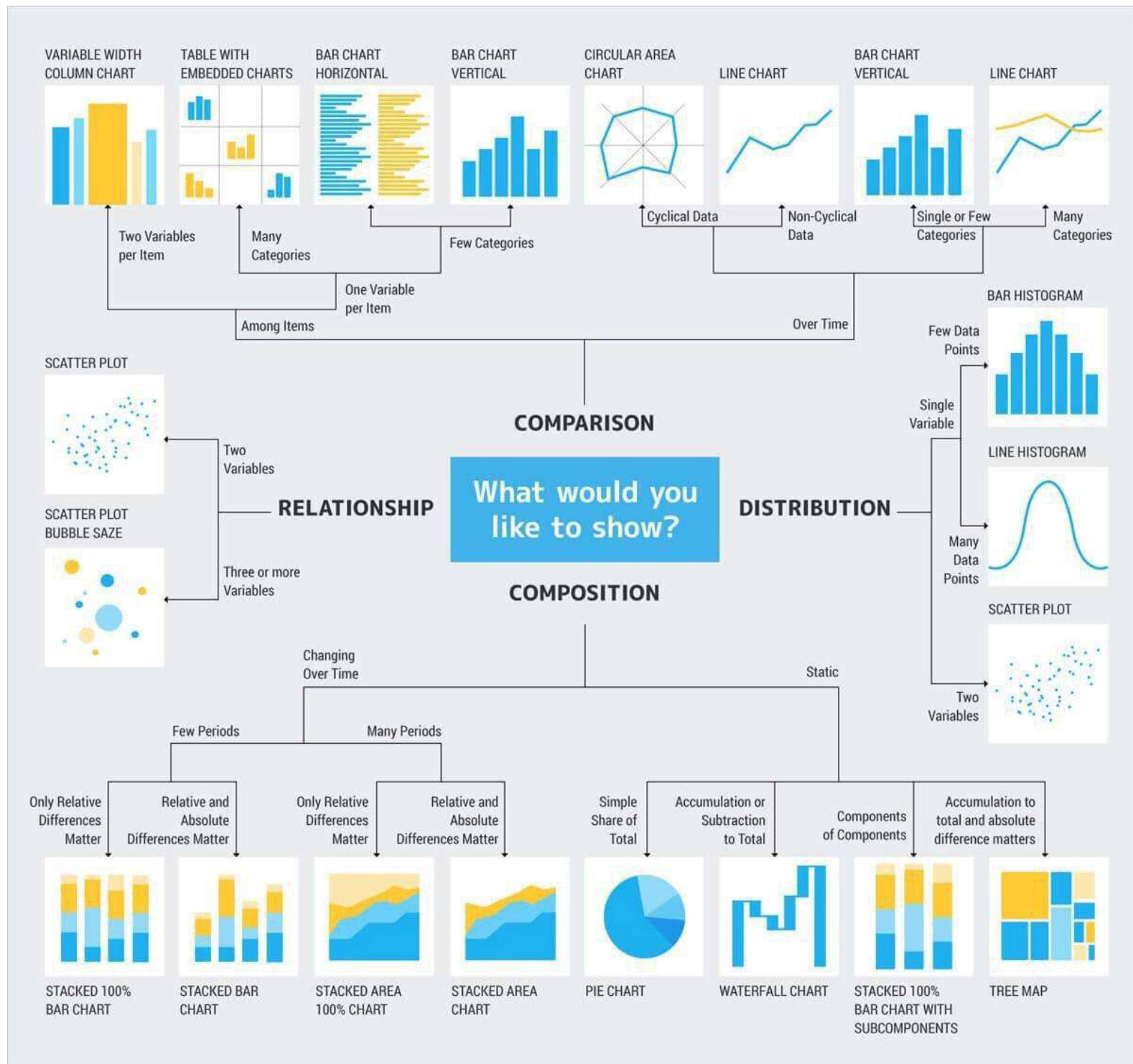


“

Fenómeno en el cual es posible concluir dos cosas totalmente opuestas a partir de los mismos datos, dependiendo el cómo se clasifican estos.

”

Diversidad de gráficas al explorar datos



“

Si bien existen reglas que te ayudarán a encontrar el gráfico apropiado para tu problema, tu imaginación debe estar siempre abierta a crear e iterar sobre el aspecto y comunicación de los gráficos.

”

“

**Los estadísticos y las gráficas
son solo un par de herramientas,
pero la exploración de datos va más
allá y se fundamenta en la resolución
de preguntas.**

”

**¿Cómo continuar
aprendiendo
sobre EDA?**



Conclusiones

- **Las preguntas son la fuente de toda exploración.** Asegúrate de definir qué quieres encontrar y quién necesita consultar los resultados desde un comienzo.
- **Es fundamental identificar el tipo de análisis de datos y variables que se requieren.** Explora las dimensiones de tu conjunto de datos y qué tipos de variables contiene.



Conclusiones

- **Siempre visualiza los estadísticos.** Todos los conjuntos de datos son diferentes, conócelos más allá de sus números de resumen.
- **Visualiza una o varias variables de distintas maneras.** La diversidad de gráficas te permitirá conocer a detalles los matices de los datos.



Continúa explorando:

- **Nuevos datos.**
- **Estadísticos.**
- **Gráficas.**
- **Comunicación de resultados.**

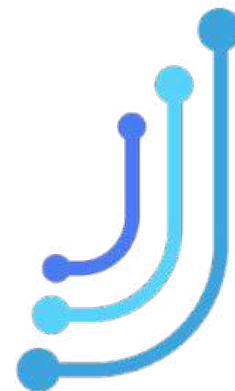
¡Felicidades!



¡Felicidades!



@jvelezmagic



jvelezmagic.com