Clase 12: Visualización de datos

Universidad de los Andes
Facultad de Economía

Contenido

- Elementos comunes
- Tipos
  - Boxplot
  - Líneas
  - Histogramas y densidades
  - Barras (verticales, horizontales, apiladas, por grupos)
  - Diagrama de dispersión (2 variables, 3 variables)
- Ejercicio

Elementos comunes

- Ejes
  - Valores
  - Nombres
  - Saltos
  - Tamaño
  - Color de texto
- Área del gráfico
  - Tipo de gráfico
  - Color del fondo, de la figura
  - Over y By
  - Ancho de la figura
- Leyenda
  - Tamaño
  - Color de texto
  - Series incluidas
  - Tamaño del marcador
- Notas
  - Tamaño
  - Posición

Cuartiles

Tomen los siguientes datos

id	edad
1	15
2	18
3	14
4	19
5	24
6	16
7	99
8	32
9	23
10	40
11	25
12	35

Cuartiles

Si uno calcula el promedio de edad, es:

30

id	edad
1	15
2	18
3	14
4	19
5	24
6	16
7	99
8	32
9	23
10	40
11	25
12	35

Cuartiles

Con estos nuevos datos, el promedio de edad es…

30

id	id	edad
1	1	25
2	2	24
3	3	23
4	4	36
5	5	32
6	6	28
7	7	35
8	8	37
9	9	34
10	10	26
11	11	31
12	12	29

Cuartiles

Con estos nuevos datos, el promedio de edad es…

30

La media no nos dice mucho acerca de cómo se diferencian estos datos… los máximos y mínimos sugieren que son muy diferentes

(entre 23 y 37 vs entre 14 y 99)

id	edad	edad
1	25	15
2	24	18
3	23	14
4	36	19
5	32	24
6	28	16
7	35	99
8	37	32
9	34	23
10	26	40
11	31	25
12	29	35

Cuartiles

Una forma más detallada de aproximarnos a los datos a primer vistazo

Cuartiles

edad
25
24
23
36
32
28
35
37
34
26
31
29

1. Ordenar los datos de menor a mayor

edad
15
18
14
19
24
16
99
32
23
40
25
35

Cuartiles

edad
23
24
25
26
28
29
31
32
34
35
36
37

1. Ordenar los datos de menor a mayor

edad
14
15
16
18
19
23
24
25
32
35
40
99

Cuartiles

edad
23
24
25
26
28
29
31
32
34
35
36
37

1. Ordenar los datos de menor a mayor
2. Separar los datos en 4 grupos más o menos iguales

edad
14
15
16
18
19
23
24
25
32
35
40
99

Cuartiles

edad
23
24
25
26
28
29
31
32
34
35
36
37

1. Ordenar los datos de menor a mayor
2. Separar los datos en 4 grupos más o menos iguales

edad
14
15
16
18
19
23
24
25
32
35
40
99

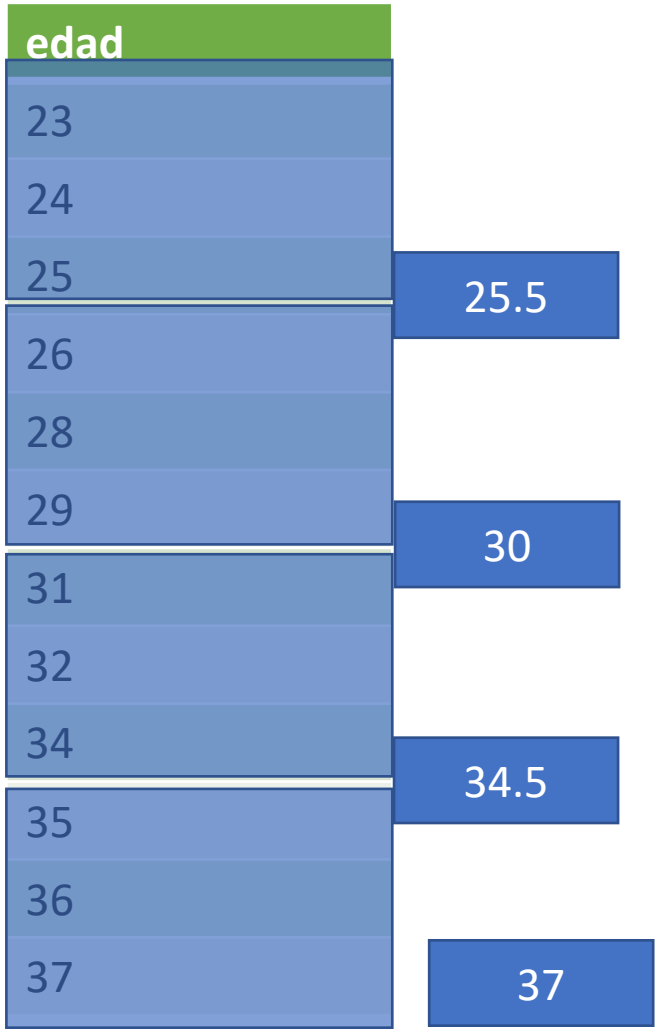
Cuartiles

edad
23
24
25
26
28
29
31
32
34
35
36
37

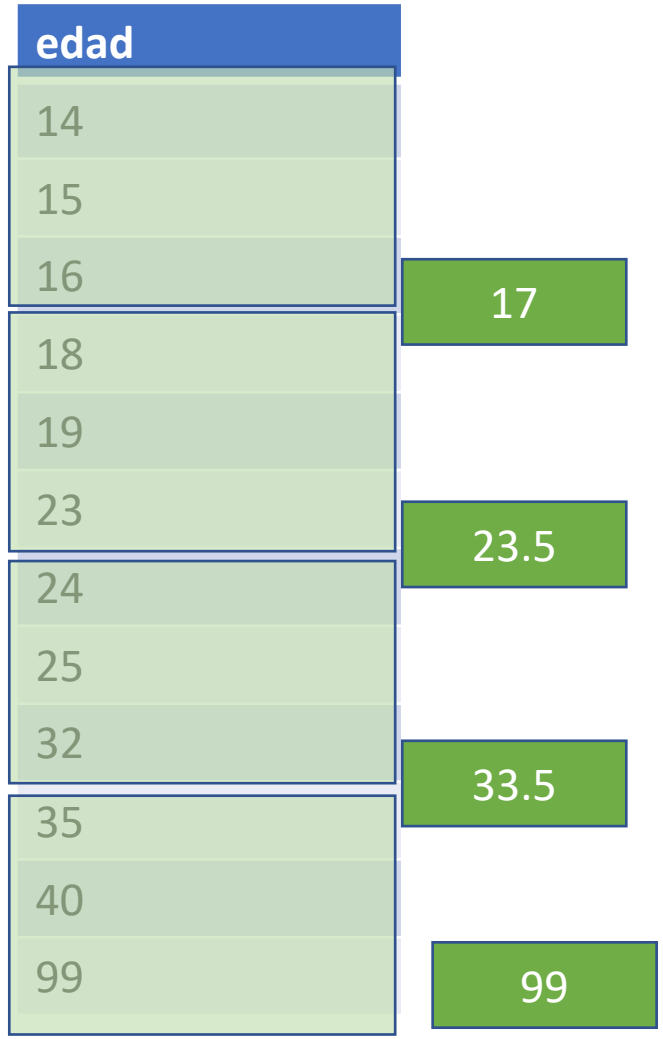
1. Ordenar los datos de menor a mayor
2. Separar los datos en 4 grupos más o menos iguales
3. Identificar los valores entre cada grupo

edad
14
15
16
18
19
23
24
25
32
35
40
99

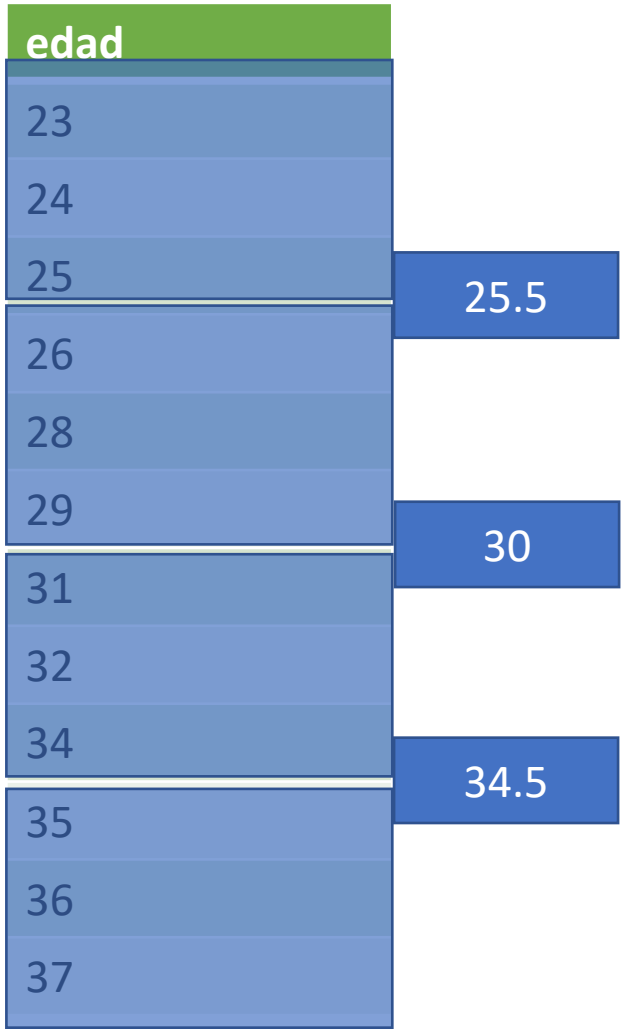
Cuartiles



1. Ordenar los datos de menor a mayor
2. Separar los datos en 4 grupos más o menos iguales
3. Identificar los valores entre cada grupo

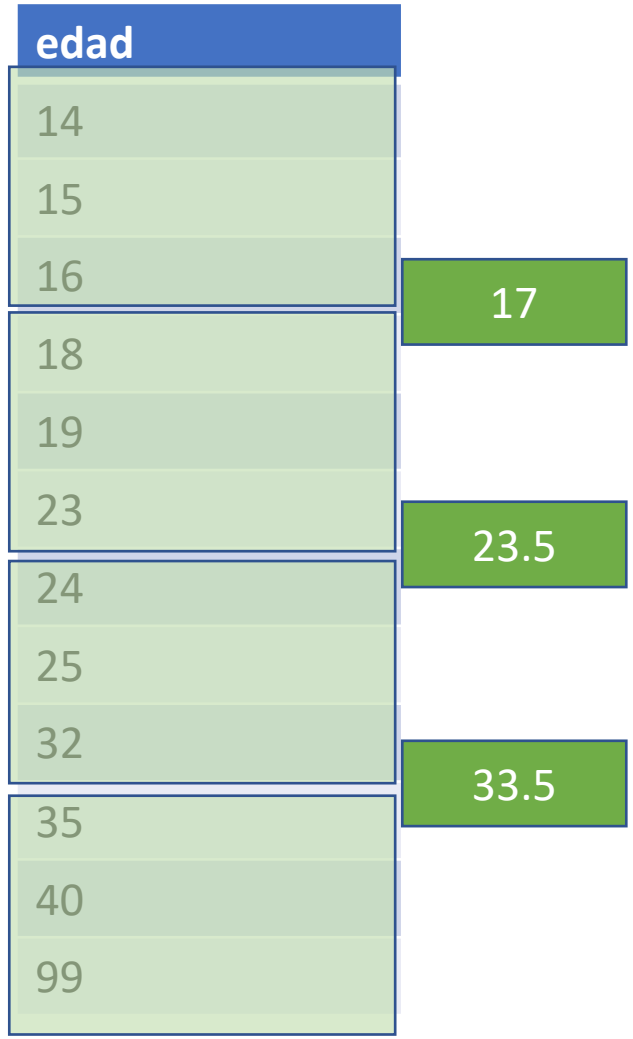


Cuartiles



1. Ordenar los datos de menor a mayor
2. Separar los datos en 4 grupos más o menos iguales
3. Identificar los valores entre cada grupo

Noten: que el 75% de los datos están por debajo de 34.5



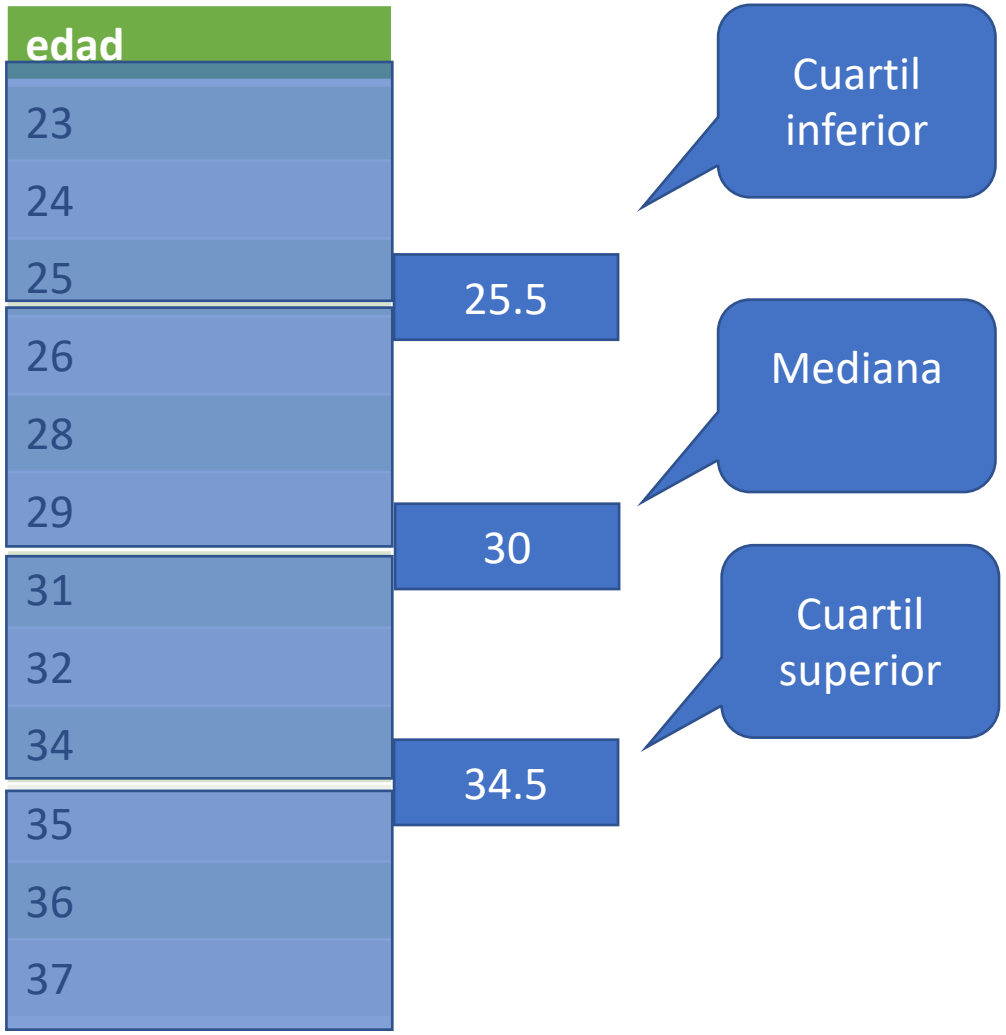
Cuartiles

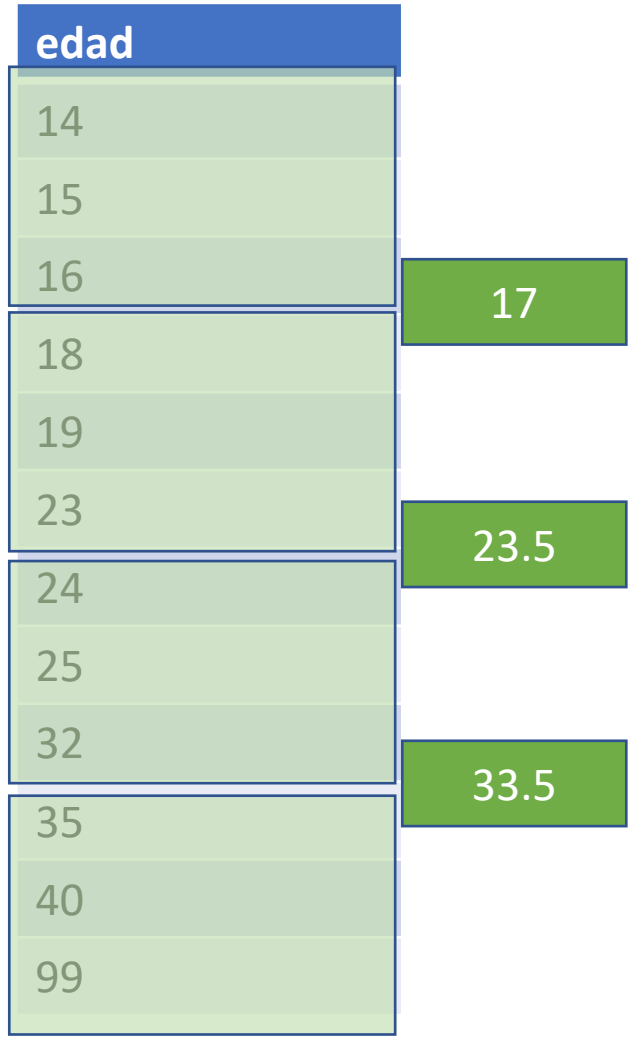
Los cuartiles nos dicen:

- El 25 % de los datos está por debajo de _____
- El 50% de los datos está por debajo de _____
- El 75% de los datos está por debajo de _____

Muy útil para identificar datos atípicos (outliers).

Cuartiles


edad
23
24
25
26
28
29
31
32
34
35
36
37
25.5
30
34.5
Cuartil inferior
Mediana
Cuartil superior


edad
14
15
16
18
19
23
24
25
32
35
40
99
17
23.5
33.5

BoxPlot

edad
23
24
25
26
28
29
31
32
34
35
36
37

25.5

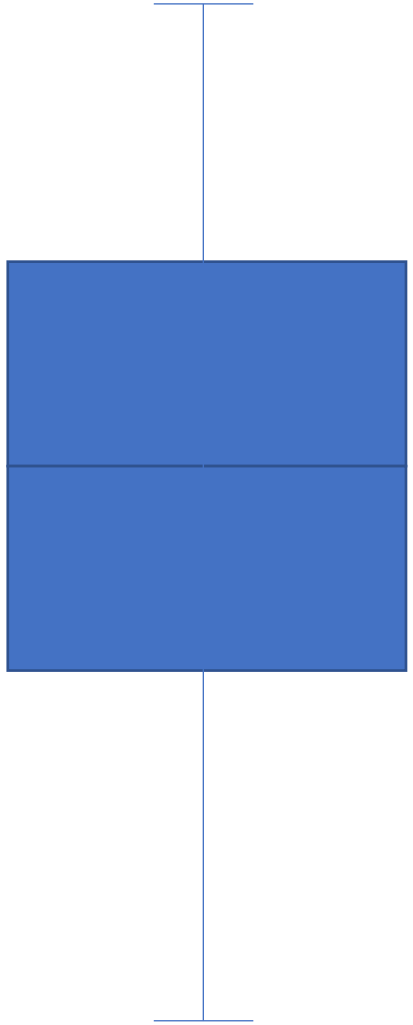
30

34.5

Cuartil inferior

Mediana

Cuartil superior



BoxPlot

edad
23
24
25
26
28
29
31
32
34
35
36
37

25.5

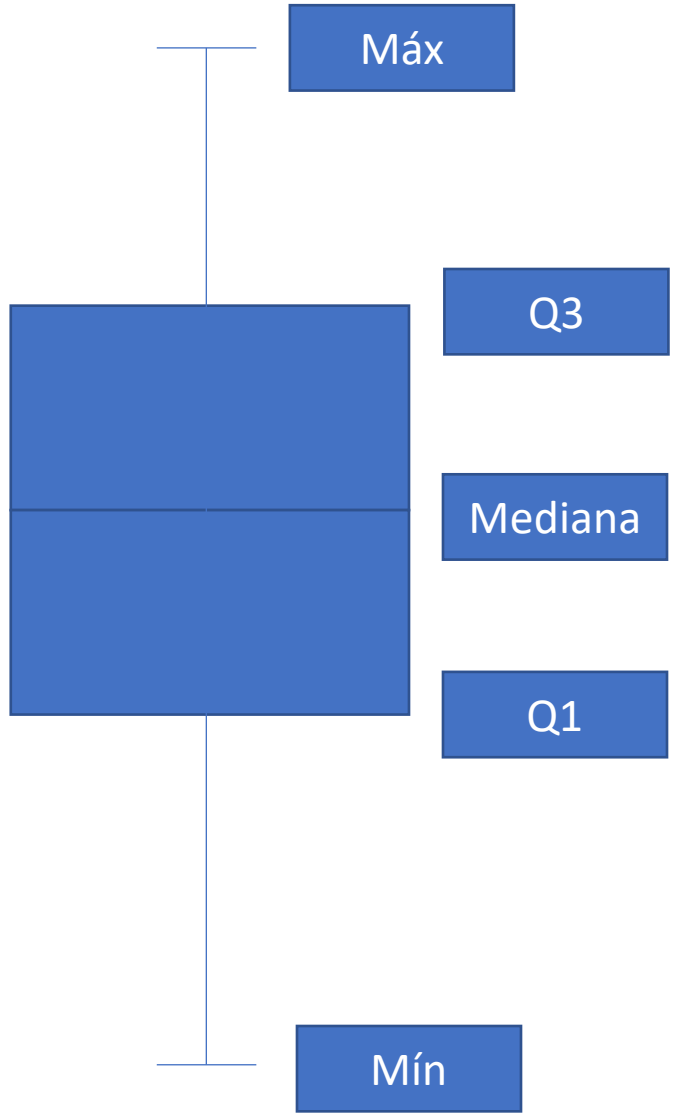
30

34.5

Cuartil inferior

Mediana

Cuartil superior



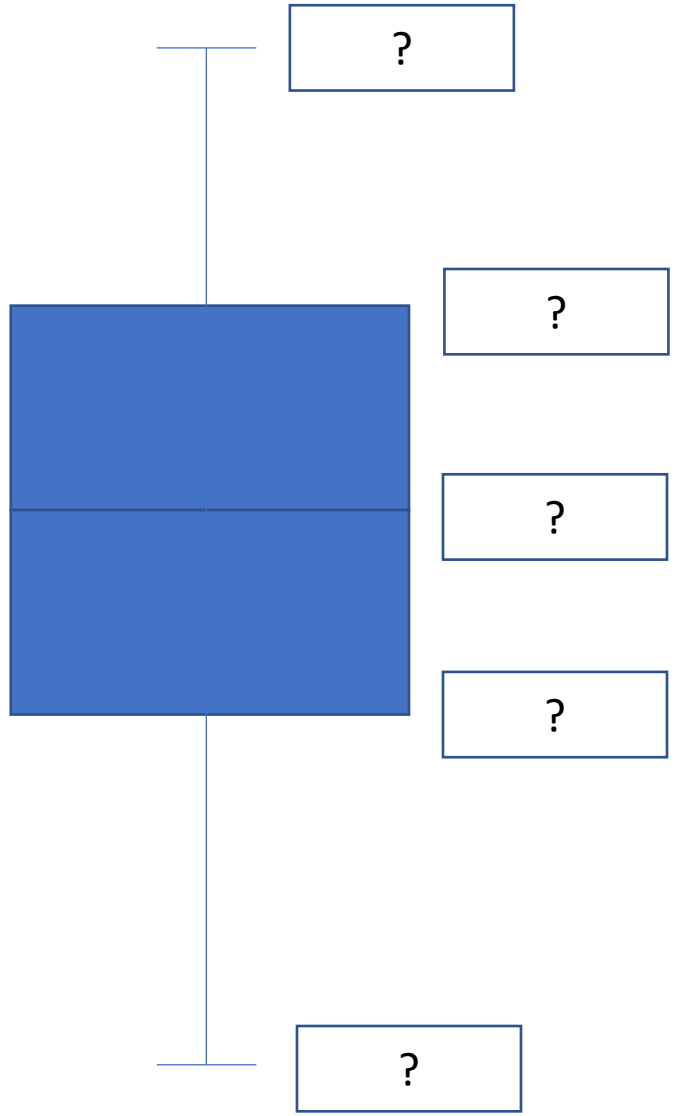
BoxPlot

edad
23
24
25
26
28
29
31
32
34
35
36
37

25.5

30

34.5



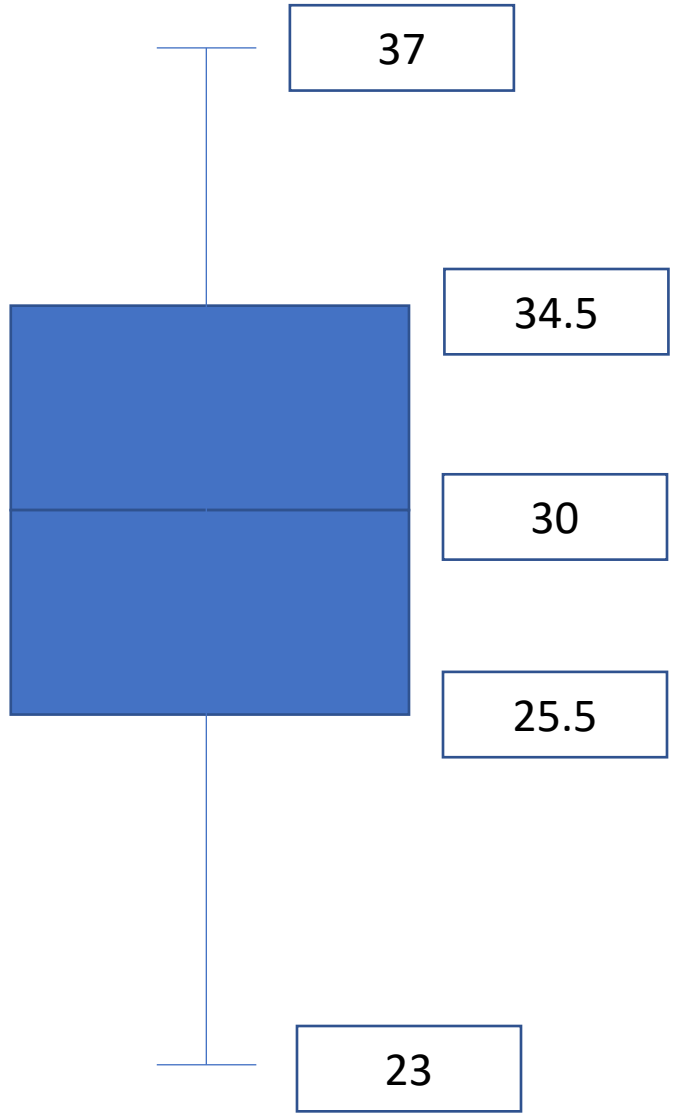
BoxPlot

edad
23
24
25
26
28
29
31
32
34
35
36
37

25.5

30

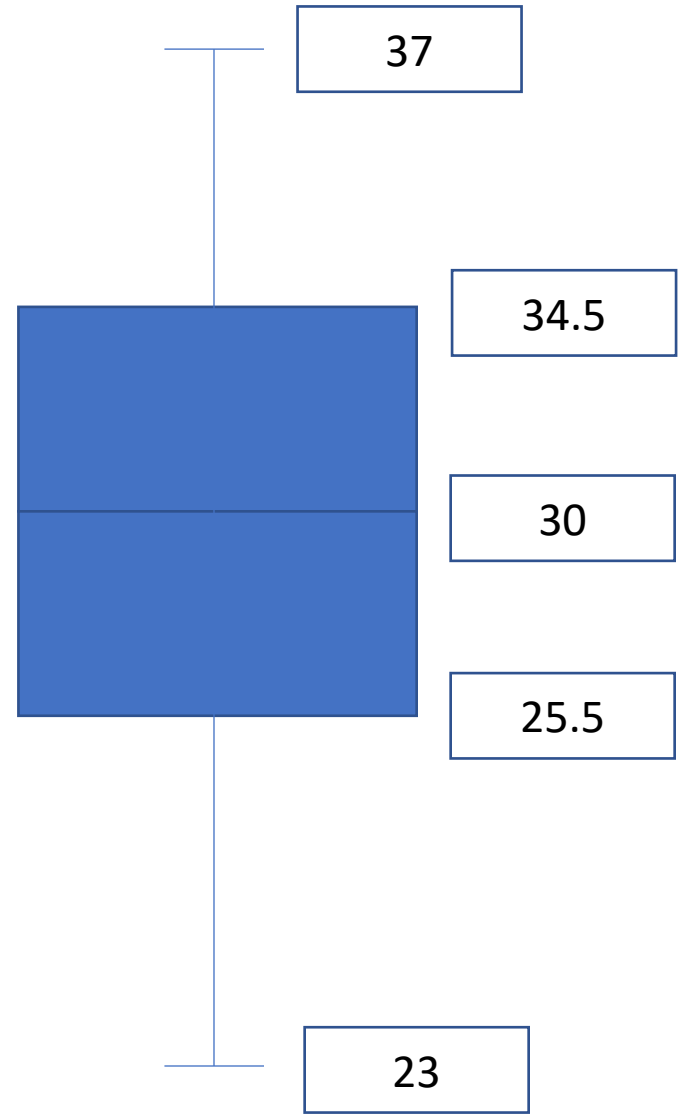
34.5



BoxPlot

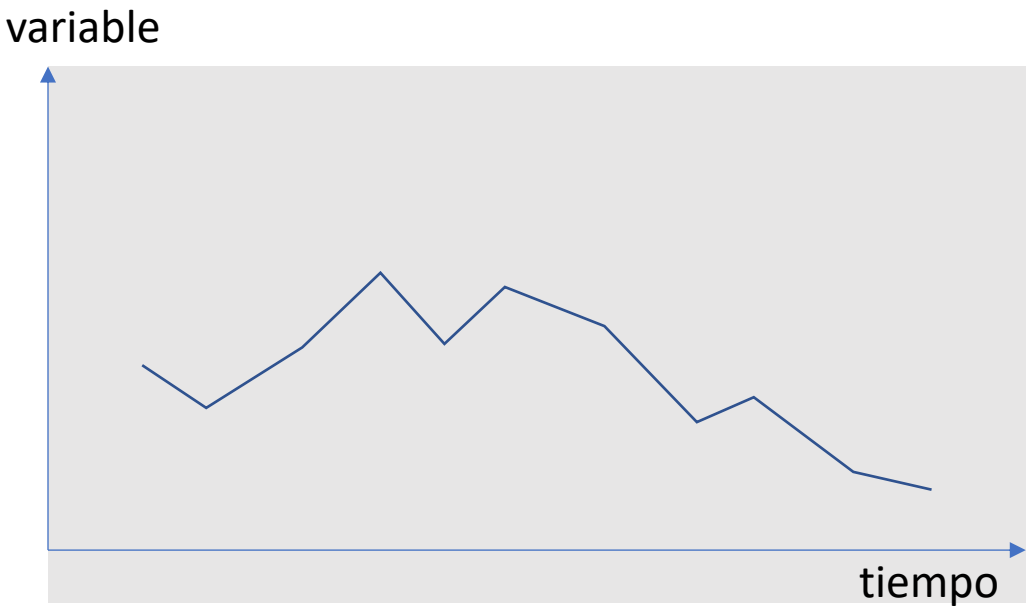
Nos permiten observar la **distribución ordenada** de los datos con mucho detalle.

Son útiles para identificar datos atípicos o outliers.



Líneas

Lineas



Útiles cuando tenemos un dato que hemos podido rastrear a lo largo del tiempo.

Nos permiten observar **la evolución de los procesos** en el tiempo.

También son útiles para comparar cambios en las tendencias.

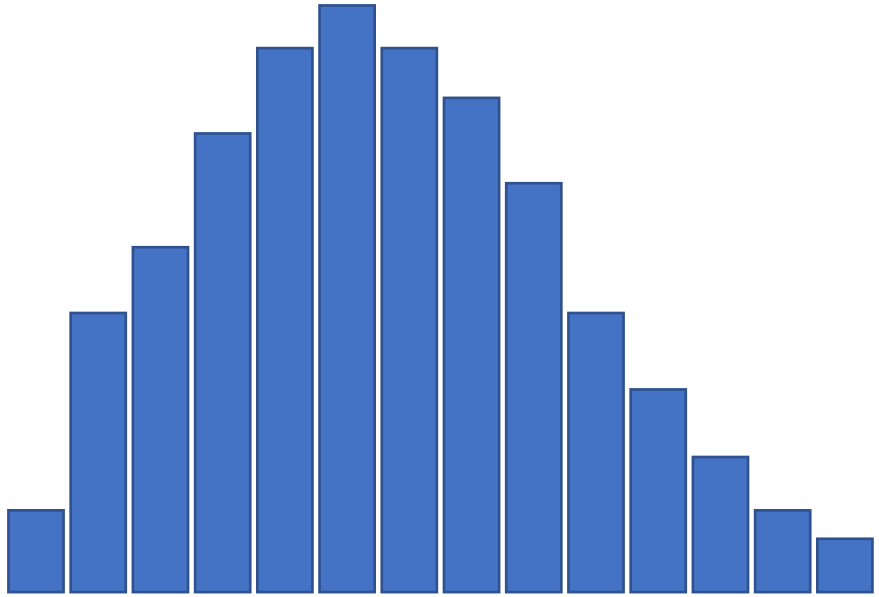
NOTE: Un diagrama de líneas puede requerir modificar la tabla antes de graficarse:

id	año	valor
1	2012	25
2	2012	24
1	2013	23
2	2013	36
1	2014	32
2	2014	28
1	2015	35
2	2015	37
1	2016	34
2	2016	26
1	2017	31
2	2017	29

año	promedio anual
2012	23.5
2013	29.5
2014	30
2015	36
2016	30
2017	30

Histogramas y densidades

Histograma

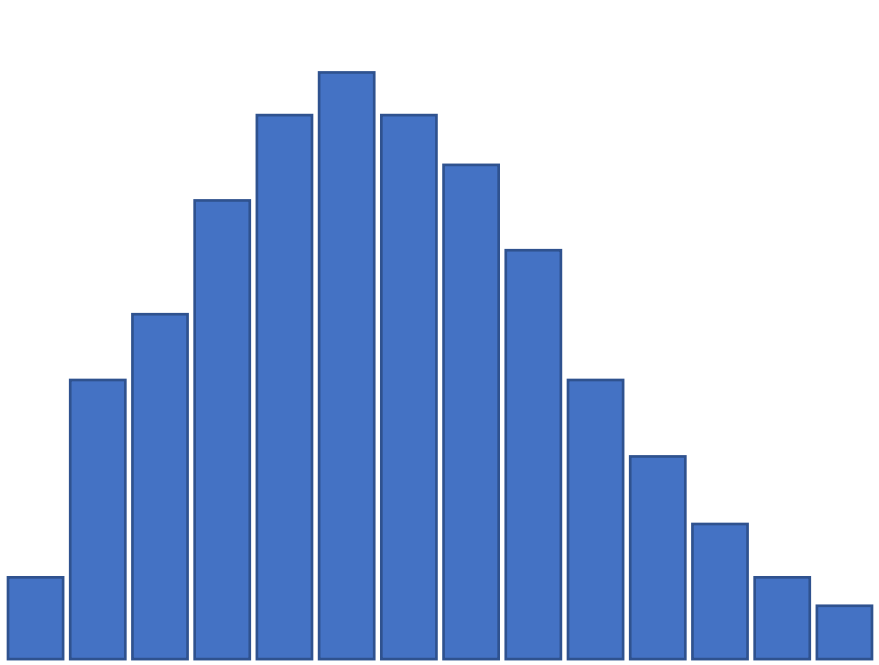


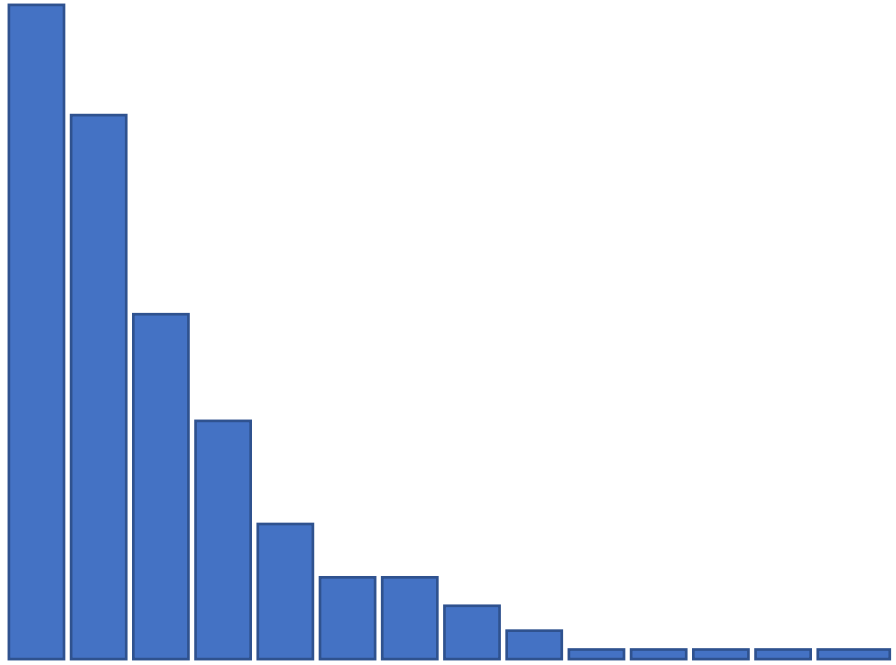
Si una variable toma valores entre ***tanto*** y ***tanto***, ¿cuántos de los registros caen en cada posible valor?

Similar a la pregunta que se hace boxplot.

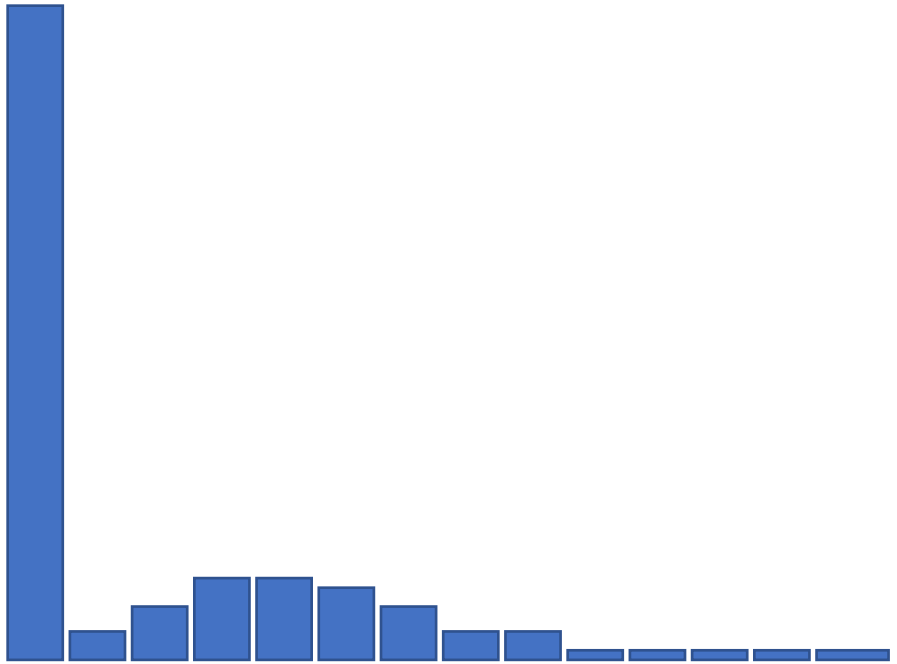
Más útil para ver picos, y dispersión.

¿Qué nos dicen estos histogramas?



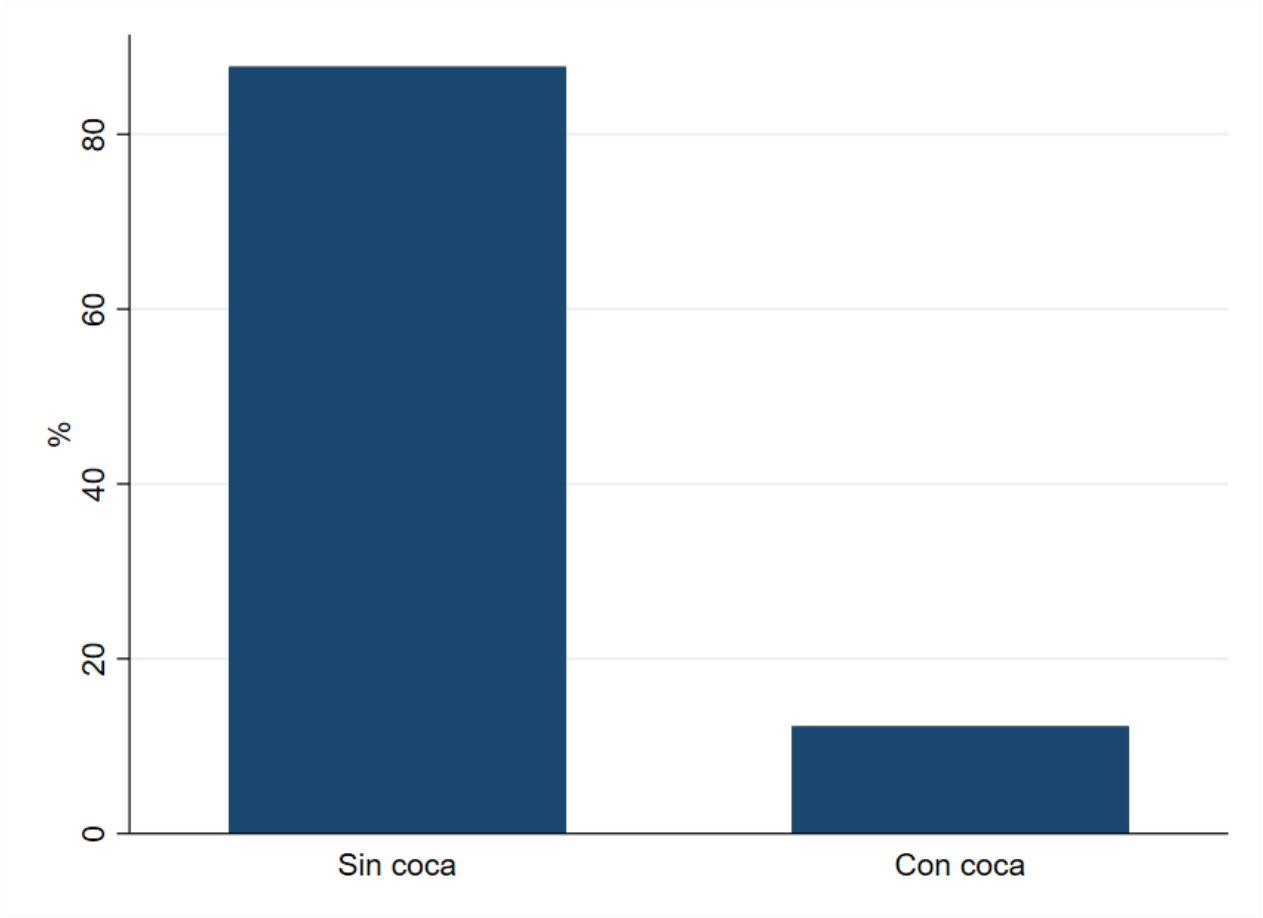


¿Y este?



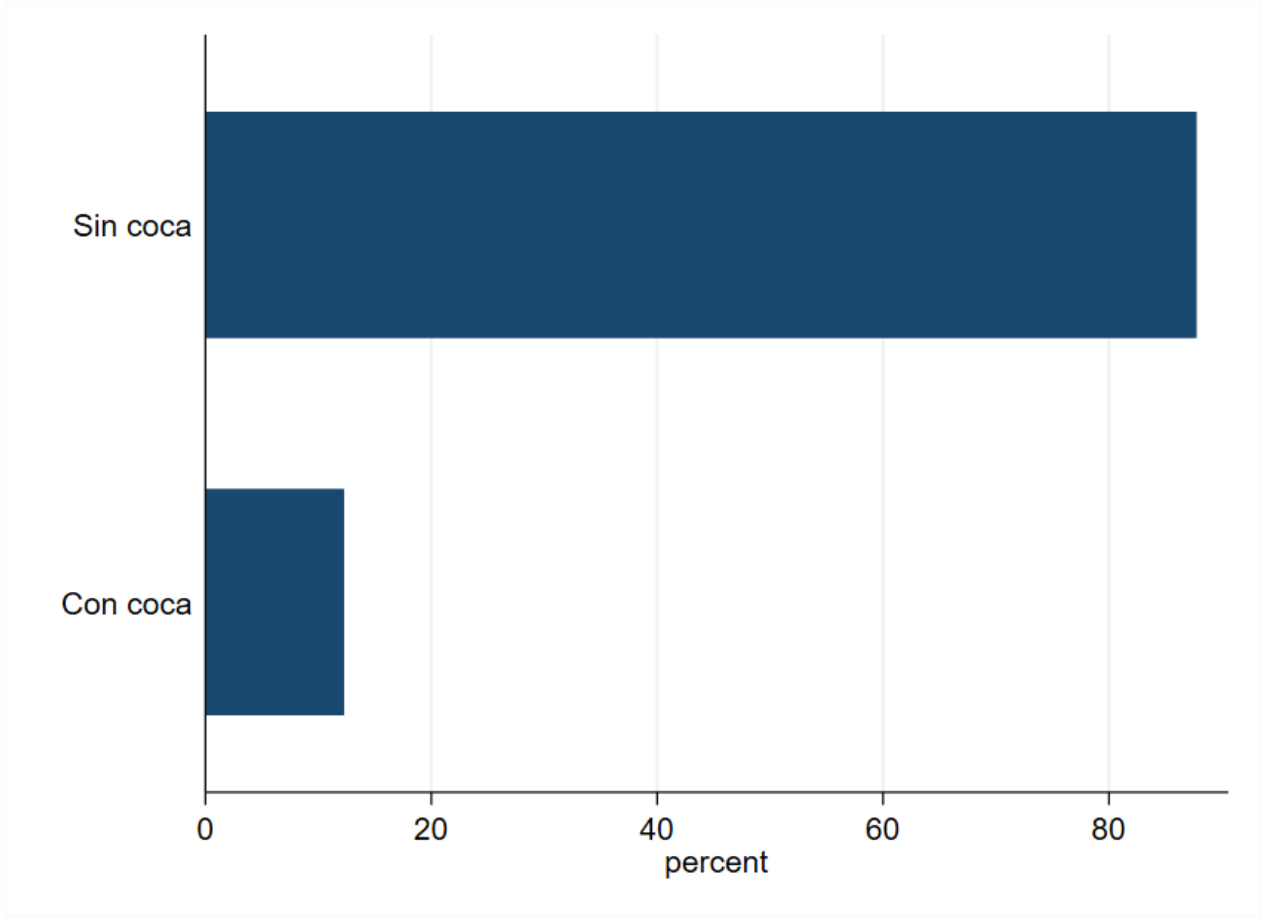
Barras

Barras (1)



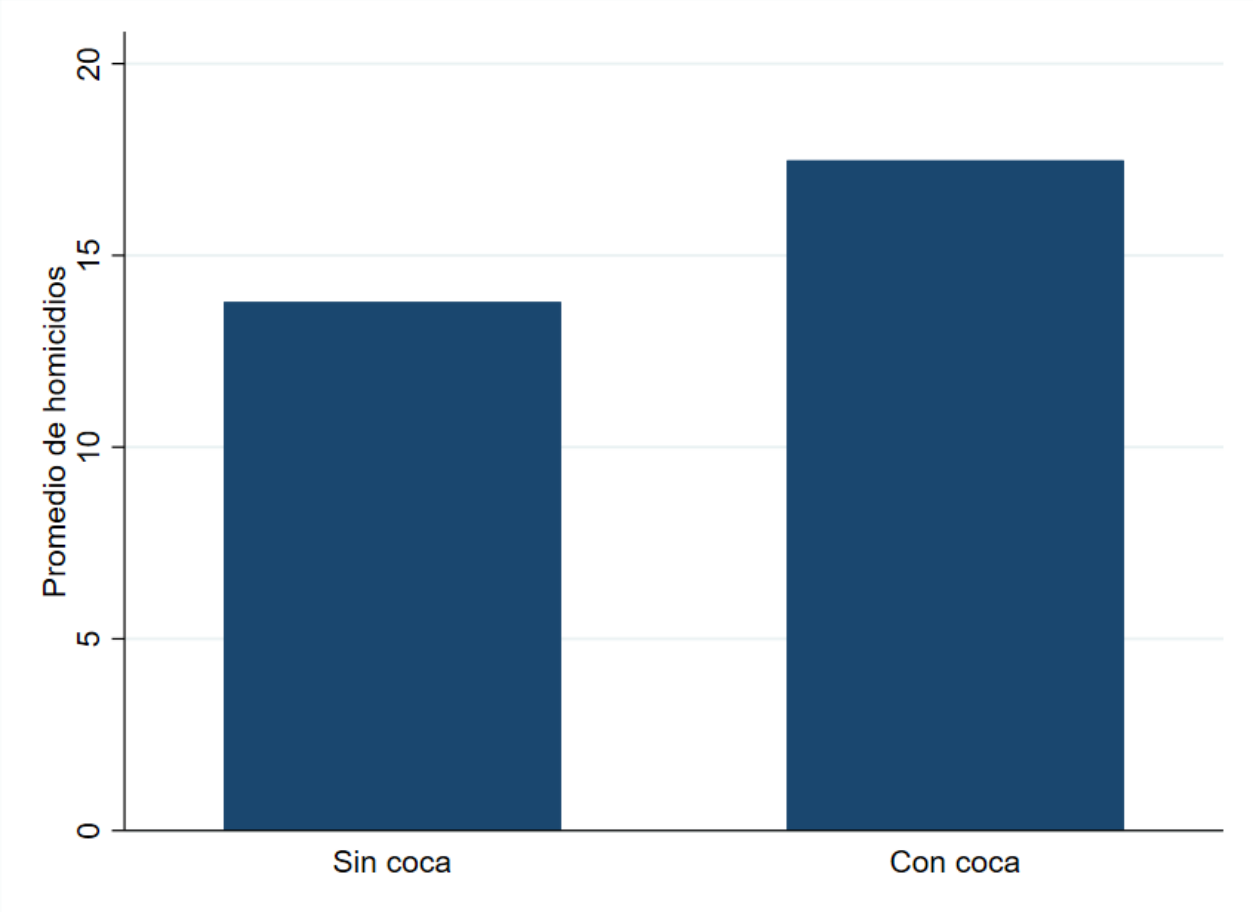
¿Cuándo usar?
Son gráficos que se ajustan a variables de pocas categorías y se requieran hacer conteos absolutos o relativos de las mismas.

Barras (2)



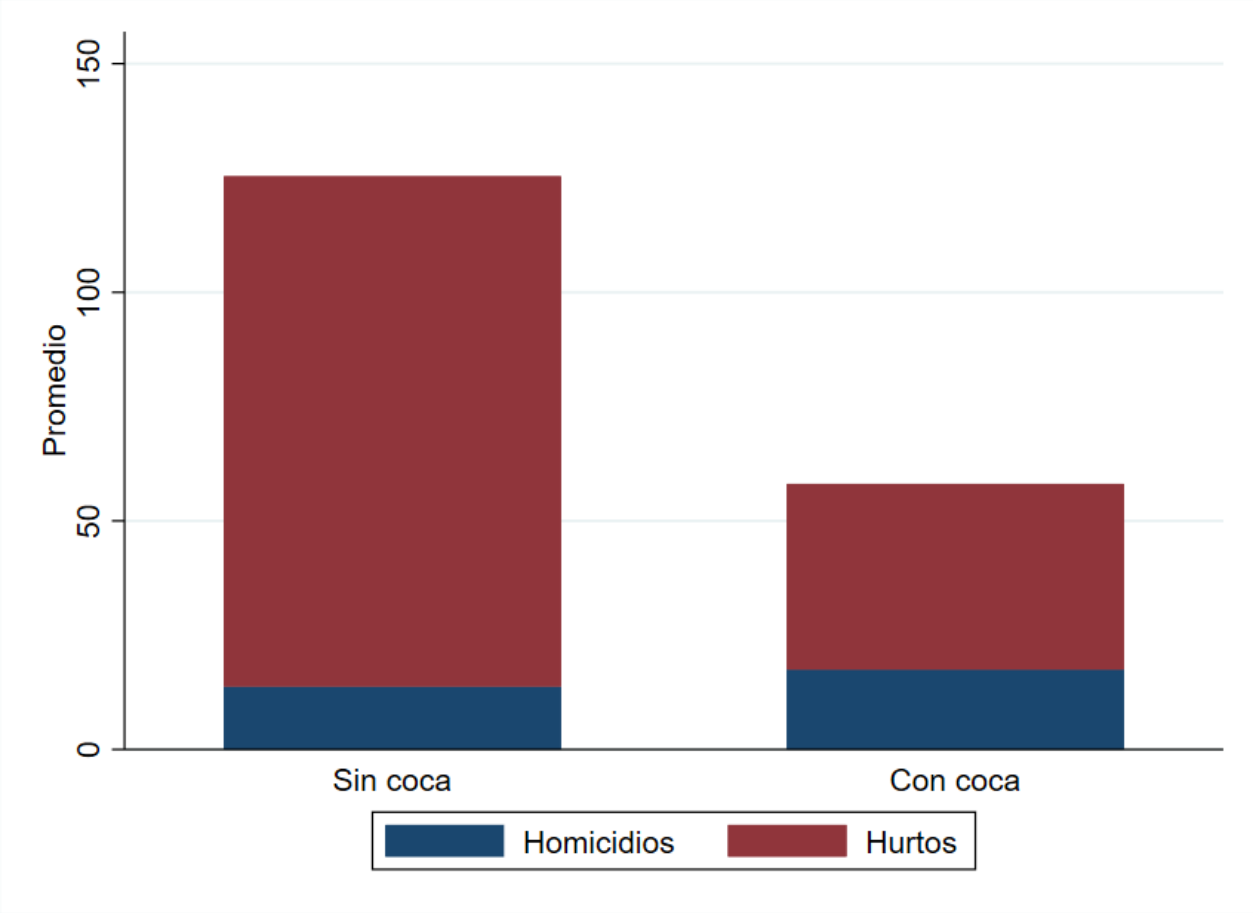
¿Cuándo usar?
En el mismo caso que el anterior pero cuando los nombres de la categoría sean muy extensaas

Barras (3)



¿Cuándo usar?
Son gráficos útiles para analizar una variable continua comparando las diferencias respecto a categorías de otra variable.

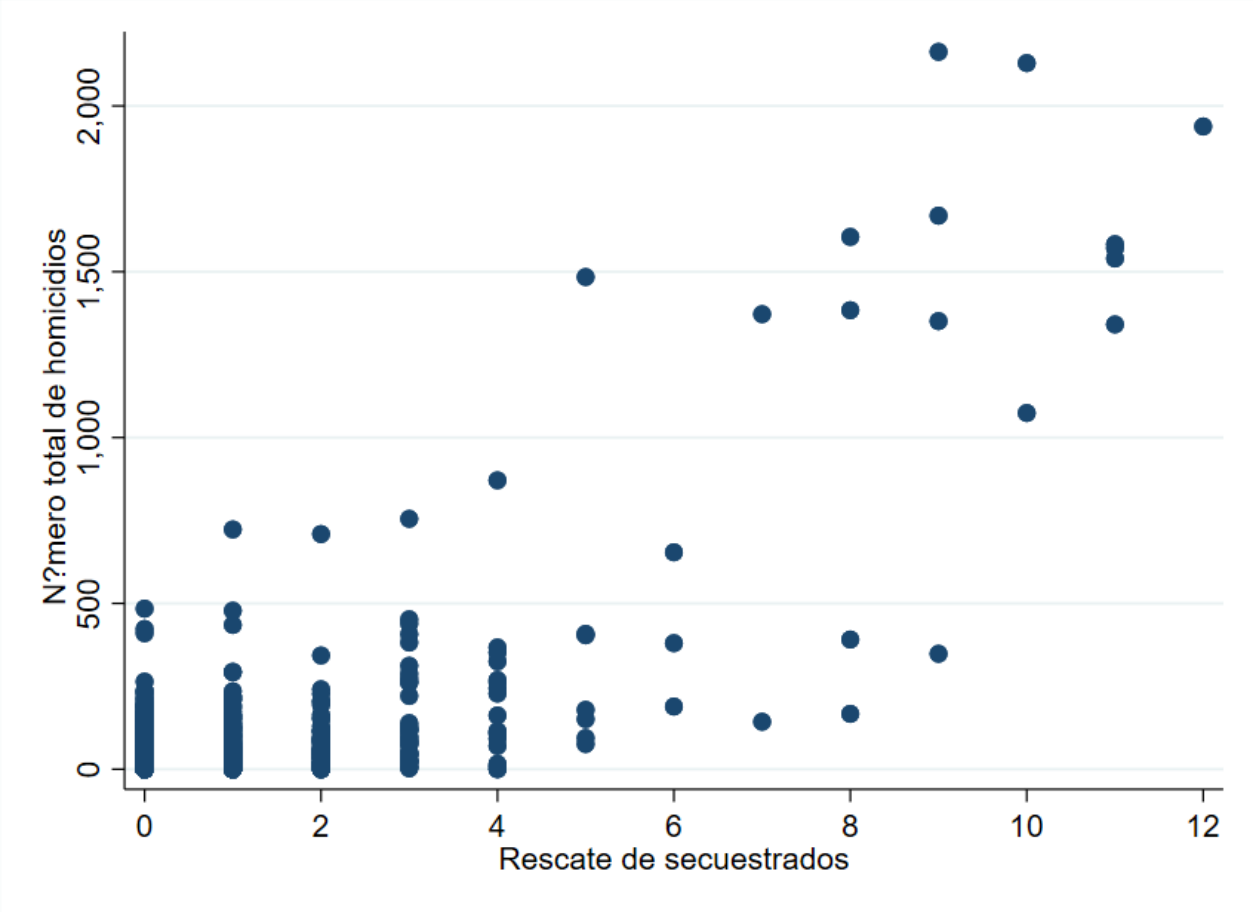
Barras (4)



¿Cuándo usar?
Son útiles cuando sea necesario sumar más de una variable continua para analizar el efecto conjunto de todas ellas. Pueden complementarse con la comparación de categorías como en el ejemplo.

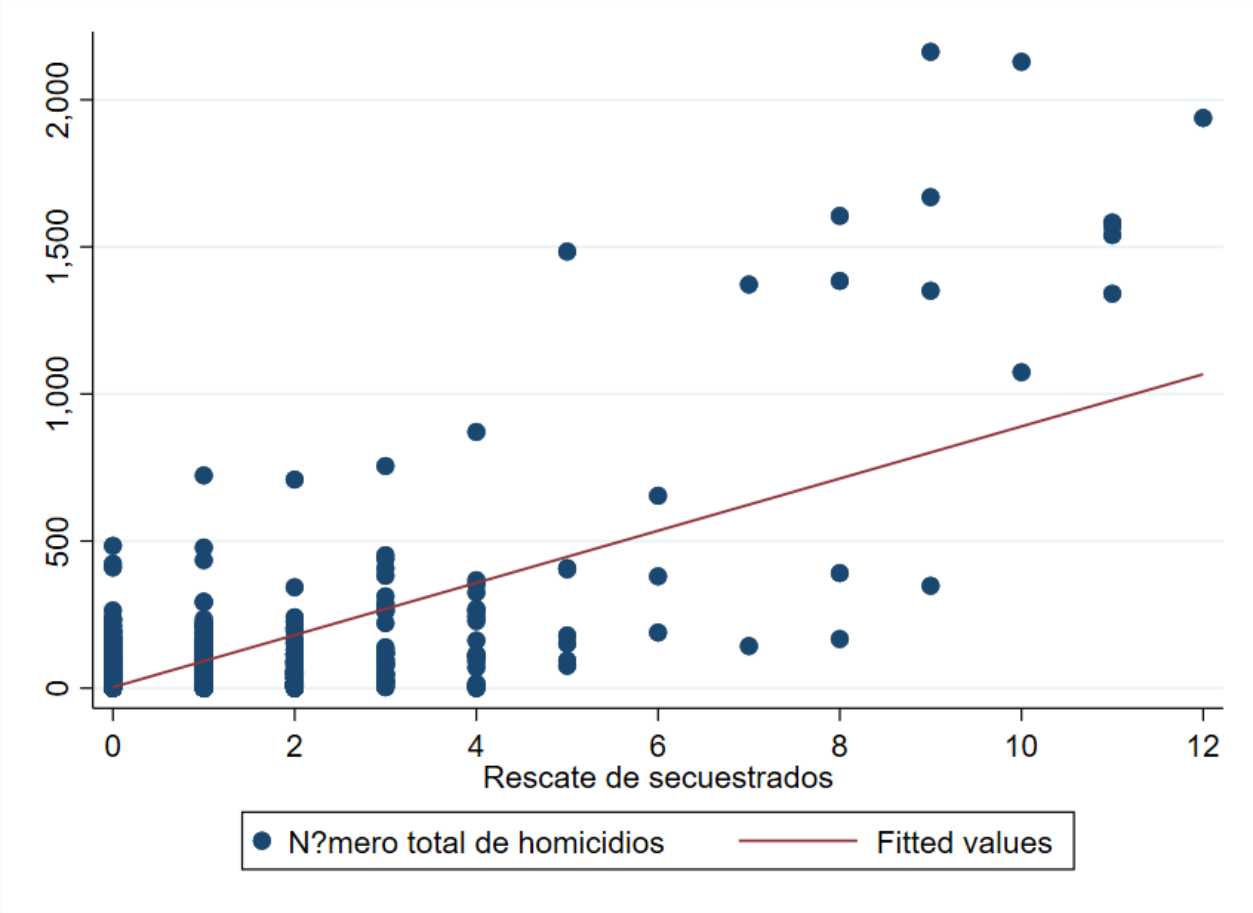
Diagramas de dispersión

Diagrama de dispersión (1)



¿Cuándo usar?
Son gráficos especiales para analizar relación entre variables continuas. Para variables discretas los ajustes en ocasiones se quedan cortos.

Diagrama de dispersión (2)



¿Cuándo usar?
Son gráficos especiales para analizar relación entre variables continuas y se quiera estar seguro del sentido del efecto. El ajuste lineal es la línea de regresión muestral.

Visualización (vs solo graficar)

Tomado (casi literalmente) de una charla de Cole Nussbaumer Knaflic en Google

https://youtu.be/8EMW7io4rSI

¿Cuántos 8 hay?

3 3 8 3 3 3 3 3 3 3 3

3 3 3 3 3 3 3 8 3 3 3

3 3 3 3 8 3 3 3 3 3 3

3 3 3 3 3 3 8 3 3 8 3

3 3 8 3 3 3 3 3 3 3 3

3 3 3 3 3 3 3 8 3 3 3

3 3 3 3 8 3 3 3 3 3 3

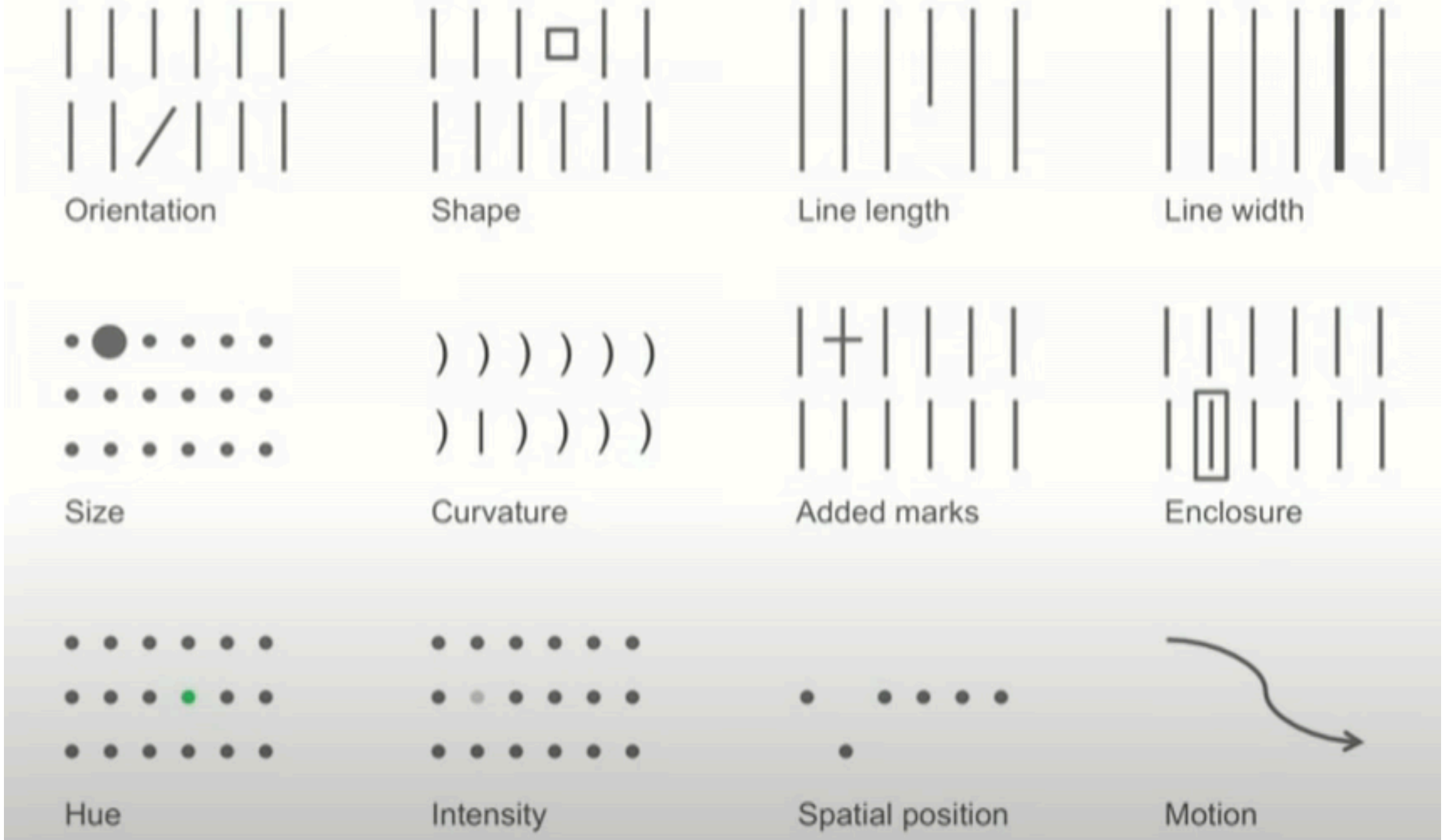
3 3 3 3 3 3 8 3 3 8 3

Orientar la atención

Pistas visuales

Nuestro cerebro está “cableado” para poner atención a cierto tipo de información visual.

Si ponemos atención, podemos utilizar estas características estratégicamente para comunicar una idea.



source: Knaflic, Cole. *Storytelling With Data: A Data Visualization Guide for Business Professionals*, Wiley, © 2015.

storytelling with data

Note

Algunas características tienen una noción de cantidad asociada, mientras otras no:

- Una línea más larga se suele percibir como una mayor cantidad
- Un punto más azul, no

- Algunos atributos nos sirven como **representadores de cantidad**, otros como **diferenciadores de categorías**.

Hacer de los datos **un punto pivotal** en una historia

No solo mostrarlos

Una idea por ~~párrafo~~ gráfico

- Los gráficos nos dejan **aprendizajes**
- Nos traen **sorpresas**
- Crean **misterio** sobre las posibles causas
- Inicio, nudo desenlace:
  - Qué contexto es necesario para entender el gráfico
  - Qué es interesante de los datos
  - Cómo queremos que la audiencia reaccione: acción

Esos son componentes asociados a una historia:

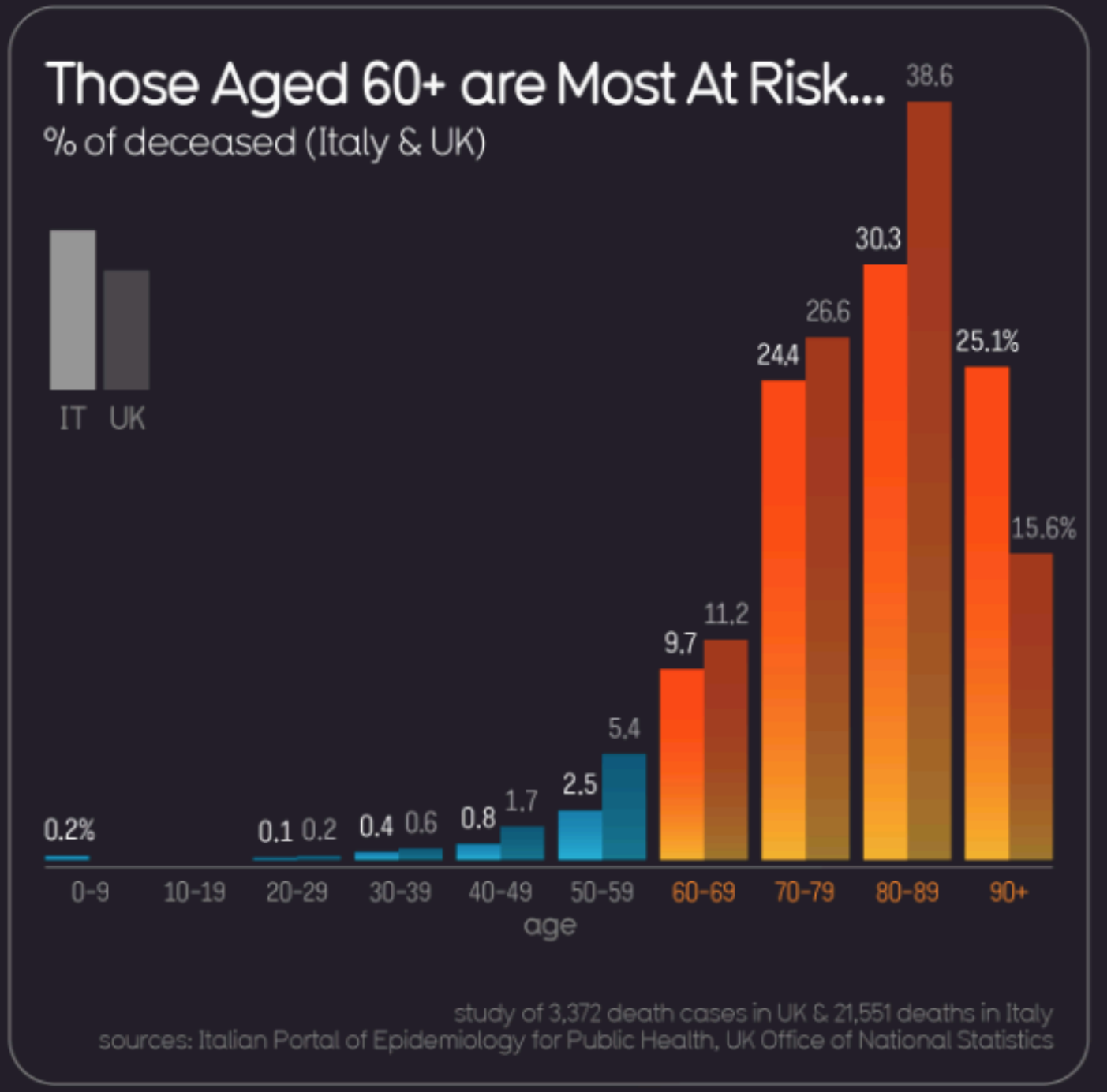
Es la historia de un viaje de descubrimiento…

…

¡Es la historia de un evento real!

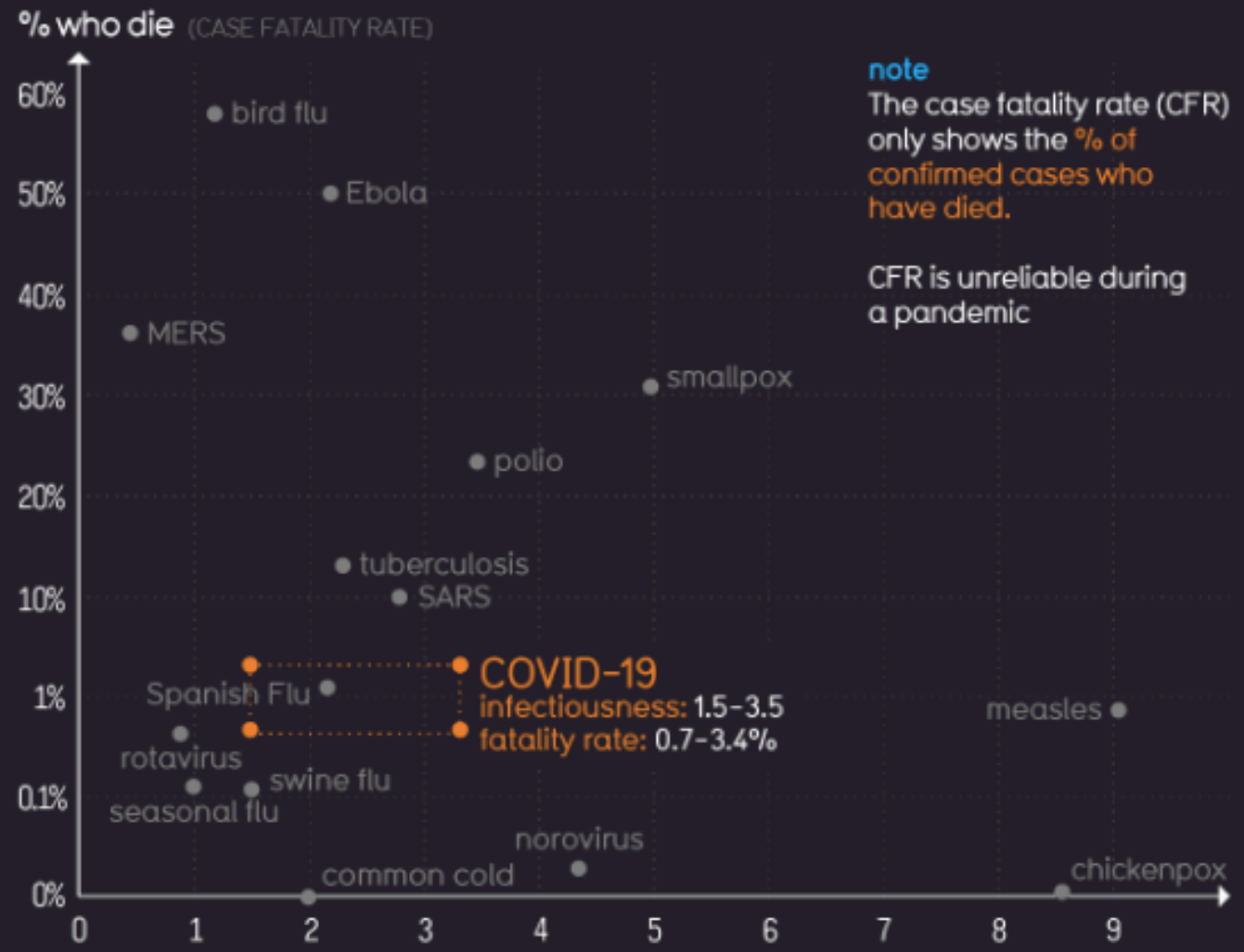
Ejemplos por

https://informationisbeautiful.net/


Those Aged 60+ are Most At Risk...
% of deceased (Italy & UK)
IT
UK
0.2%
0.1
0.2
0.4
0.6
0.8
1.7
2.5
5.4
9.7
11.2
24.4
26.6
30.3
38.6
25.1%
15.6%
0-9
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90+
age
study of 3,372 death cases in UK & 21,551 deaths in Italy
sources: Italian Portal of Epidemiology for Public Health, UK Office of National Statistics

How Contagious & Deadly is It?

We don't fully know yet but it's in this range



note

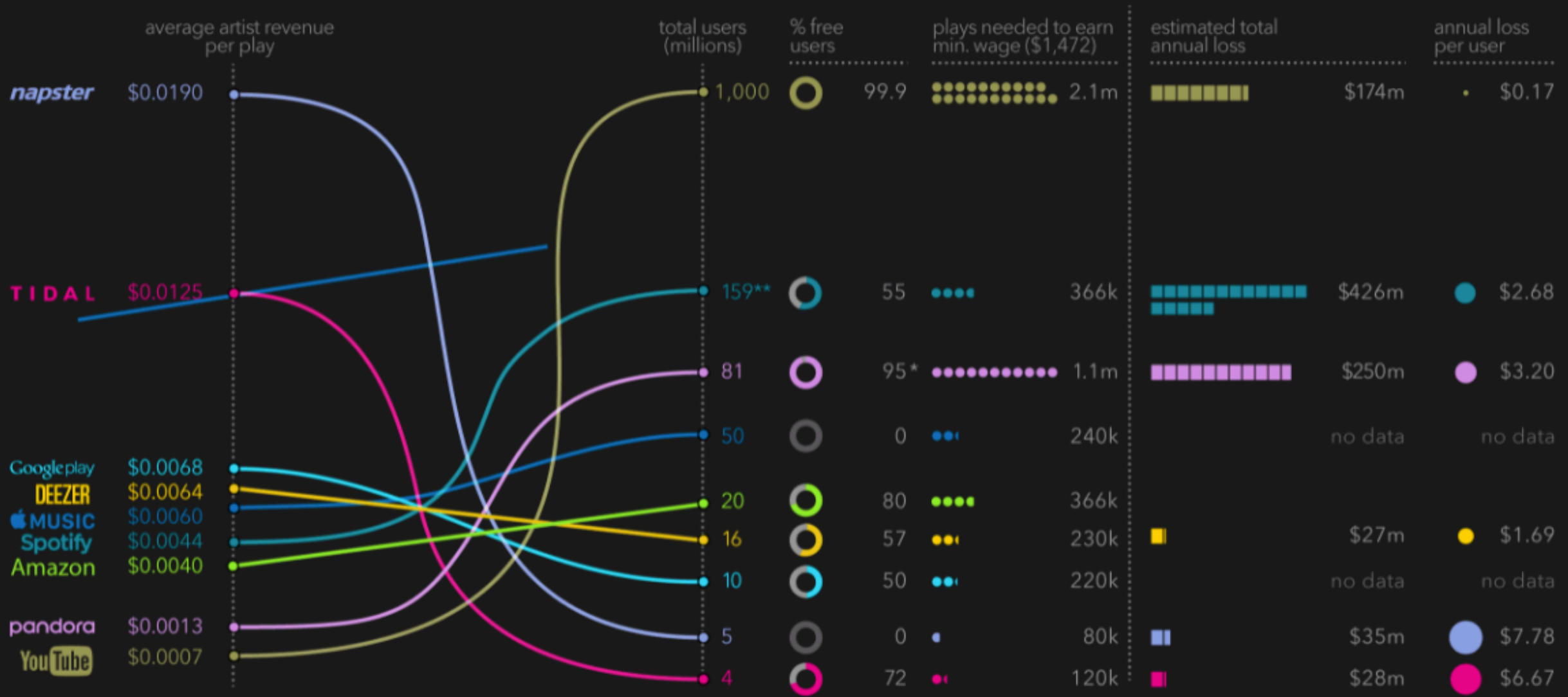
The case fatality rate (CFR) only shows the % of confirmed cases who have died.

CFR is unreliable during a pandemic

sources: US Centers for Disease Control & Prevention, WHO, New York Times

Money Too Tight to Mention?

Major music streaming services compared



Last Update: 3rd Mar 2018 * based on 5% of standard Pandora users taking up their new on-demand service

data: bit.ly/KIB_stream

informationisbeautiful.net

Armemos una historia para

- https://informationisbeautiful.net/visualizations/gender-pay-gap/

Ejercicio

Ejercicio: contexto

- Suponga que trabaja para el Banco Interamericano de Desarrollo (BID) y le piden complementar el informe final con algunos gráficos para apoyar las ideas que allí están plasmadas.

- Cuenta con una base de datos de 935 observaciones a nivel de individuos y 17 variables que se describen rápidamente a continuación: ingresos mensuales, horas trabajadas a la semana, coeficiente intelectual, años de educación, años de experiencia laboral, antigüedad, edad, estar casado (=1), ser de raza negra (=1), vivir en zonas urbanas (=1), número de hermanos, orden de nacimiento, educación de los padres.

Ejercicio: apartes del informe

1. …la educación acumulada de los padres (madre y padre) es diferente al comparar individuos que viven en zonas urbanas y zonas rurales…

2. …existe una brecha salarial en relación a la raza, pues al analizar la distribución de los ingresos es evidente que las personas de raza negra tiene una mediana menor…

3. …la distribución de salarios presenta un efecto inesperado. Los ingresos de los solteros son mayores en relación a los casados para salarios por debajo de los 750 dólares (aproximadamente). Sin embargo, después de dicho valor las personas solteras ganan consistentes menos que sus pares casados…

Ejercicio: apartes del informe

4. …de acuerdo a la teoría de la economía laboral, los individuos con más educación tienden a percibir ingresos mensuales más altos…

5. …el estudio no encuentra diferencias significativas entre el promedio de las edades de las personas que viven en zonas urbanas y aquellos que habitan las zonas rurales…