

Tema 3: Exploración de datos.



Universidad
Francisco de Vitoria
UFV Madrid

Alberto Nogales
alberto.nogales@ufv.es
Curso 2022-2023

Índice

- Diagrama de barras
- Histograma
- Pie chart
- Diagrama de dispersión
- Diagrama de burbujas
- Diagrama de tallo y hojas
- Diagrama de caja
- Otras formas de visualización.
- Estadística Descriptiva: frecuencias, percentiles, medidas de centralidad, dispersión y asimetría.
- Principales distribuciones de probabilidad.

Visualización de datos

- **Definición:** convertir los **datos** en formato tabular o visual de modo que las **características** de los **datos** y sus **relaciones** con sus **atributos** sean más intuitivas.
 - El ser humano puede analizar una inmensa cantidad de datos visualmente
 - Es capaz de detectar patrones y tendencias
 - Puede detectar anomalías que se salen de la normalidad.

Visualización de datos

La visualización se puede usar en dos momentos diferentes:

- Visualización previa: se utiliza para entender mejor los datos y sugerir posibles patrones o que tipo herramienta de KDD utilizar.
- Visualización posterior al proceso de minería de datos: se utiliza para mostrar los patrones y entenderlos.

Visualización de datos

Dependiendo de la fase, hay dos tipos de usuarios:

- Visualización previa se utiliza frecuentemente por dataset curators, para ver la calidad de los datos, tendencias o filones que investigar.
- Visualización posterior se utiliza normalmente para validar y mostrar a los expertos/clientes los resultados.

Visualización de datos

Diferentes métodos según el tipo de dato.

- Univariante: medida de variable cuantitativa simple. Mide la distribución. Ej: pie chart, histogramas ...
- Bivariante: parejas de individuos de dos variables cuantitativas. Las variables están relacionadas. Ej: diagramas de dispersión, de líneas ...
- Multivariable: representación multidimensional de datos multivariable. Ej: pixels, coordenadas geográficas.

Diagrama de barras

- Este tipo de gráfico se utiliza generalmente para representar la frecuencia de las categorías de una variable del mismo tipo. Si es cuantitativa, hacer una transformación.



Diagrama de barras ejemplo

Ejemplo: la puntuación de las reseñas de la UFV.

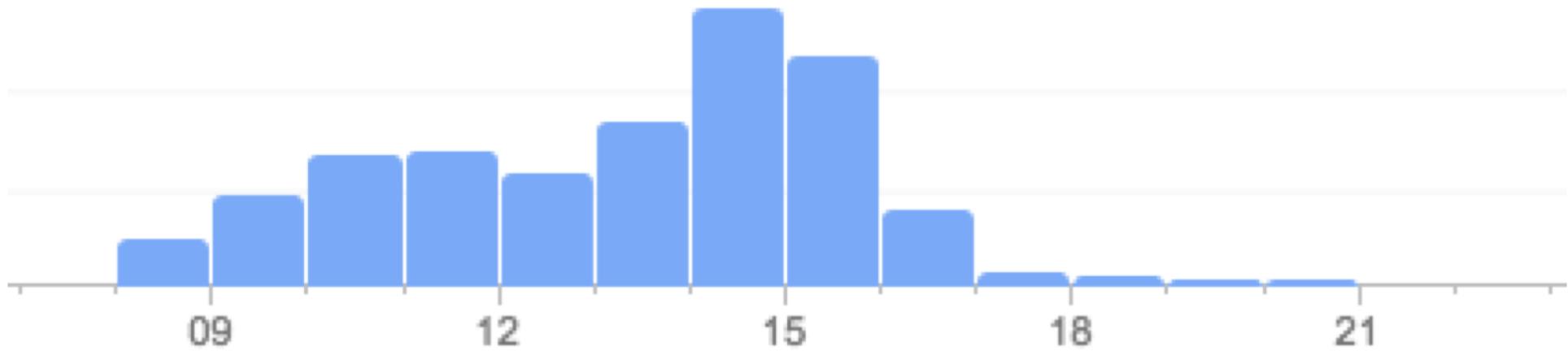
Resumen de reseñas



4,1
★★★★★
126 reseñas

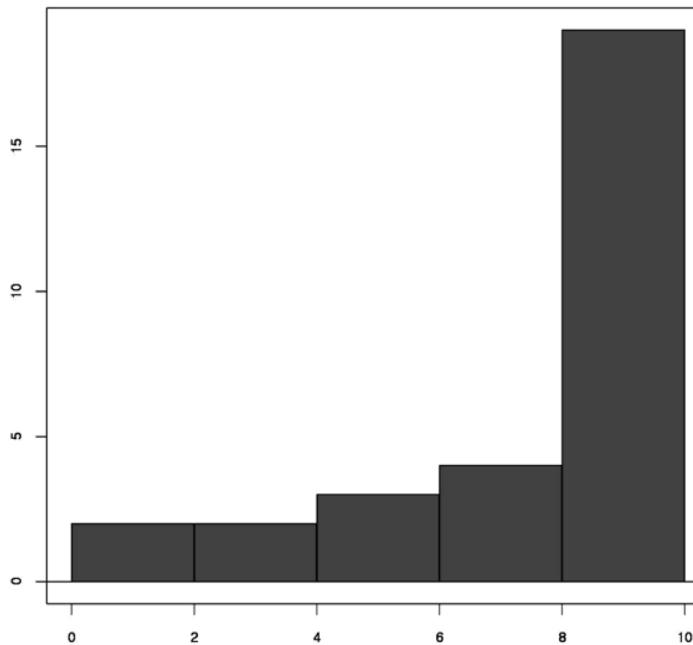
Diagrama de barras ejemplo

Ejemplo: afluencia de personas en un lugar (cafetería de la UFV).



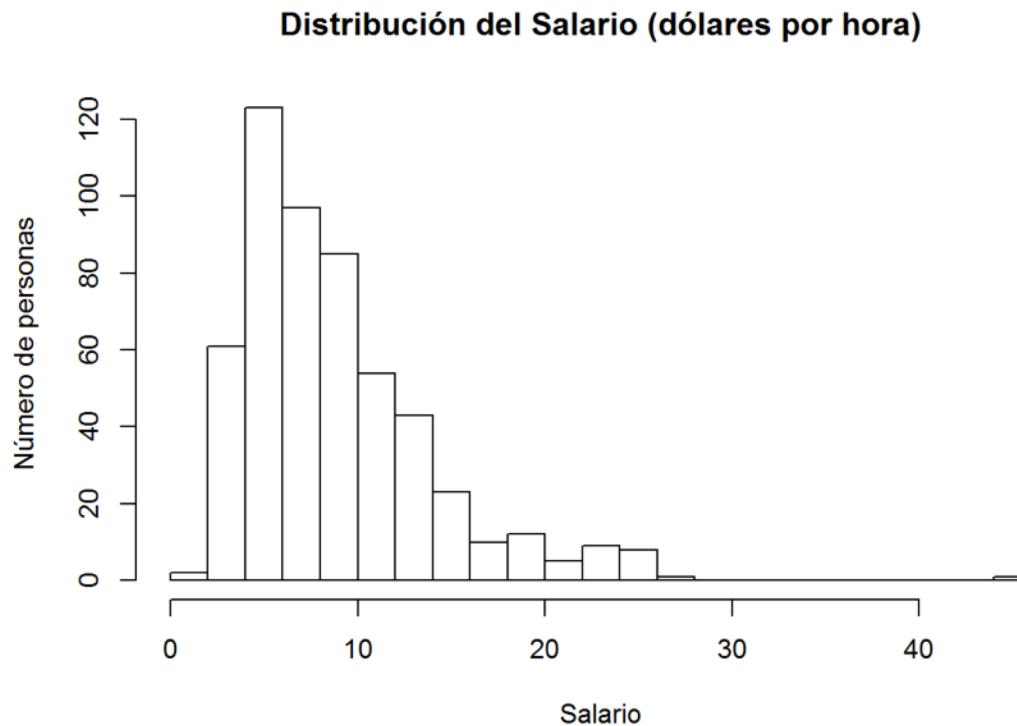
Histograma

- Permite la representación de la frecuencia de una variable cuantitativa (edad o altura) mediante intervalos. Se puede hacer en 2 dimensiones y medir la relación entre ellas. Las barras están juntas.



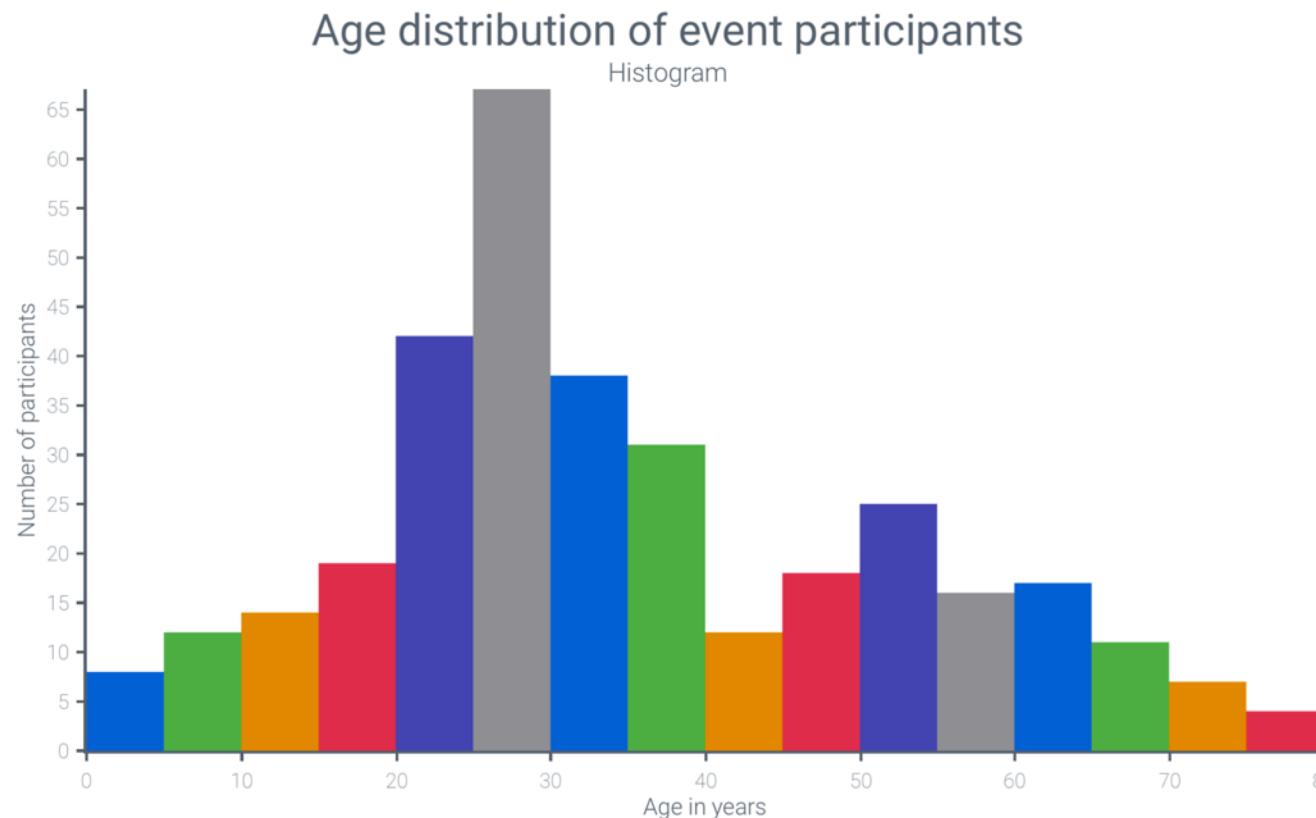
Histograma

Ejemplo: Distribución de los salarios.



Histograma

Ejemplo: Edad de los participantes en un evento.



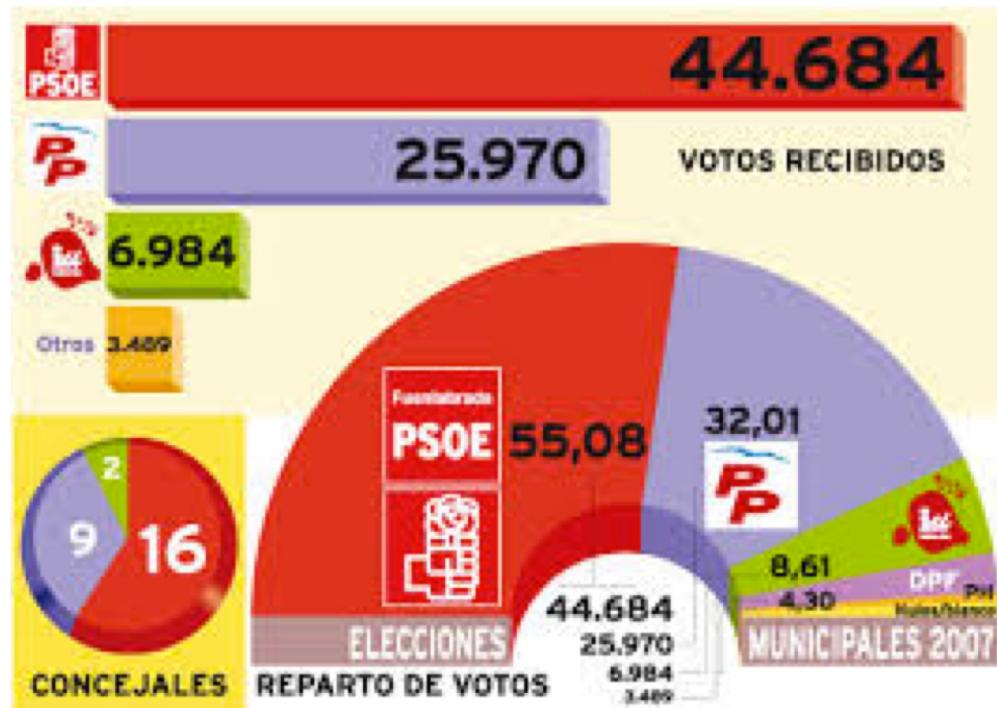
Pie chart

- Se dividen en sectores; cada uno muestra el tamaño de un fragmento de información relacionado. Los gráficos circulares suelen utilizarse para mostrar tamaños relativos de partes de un todo (porcentajes...).



Pie chart

Ejemplo: Votos en las elecciones.



Pie chart

Ejemplo: Uso de exploradores Web.

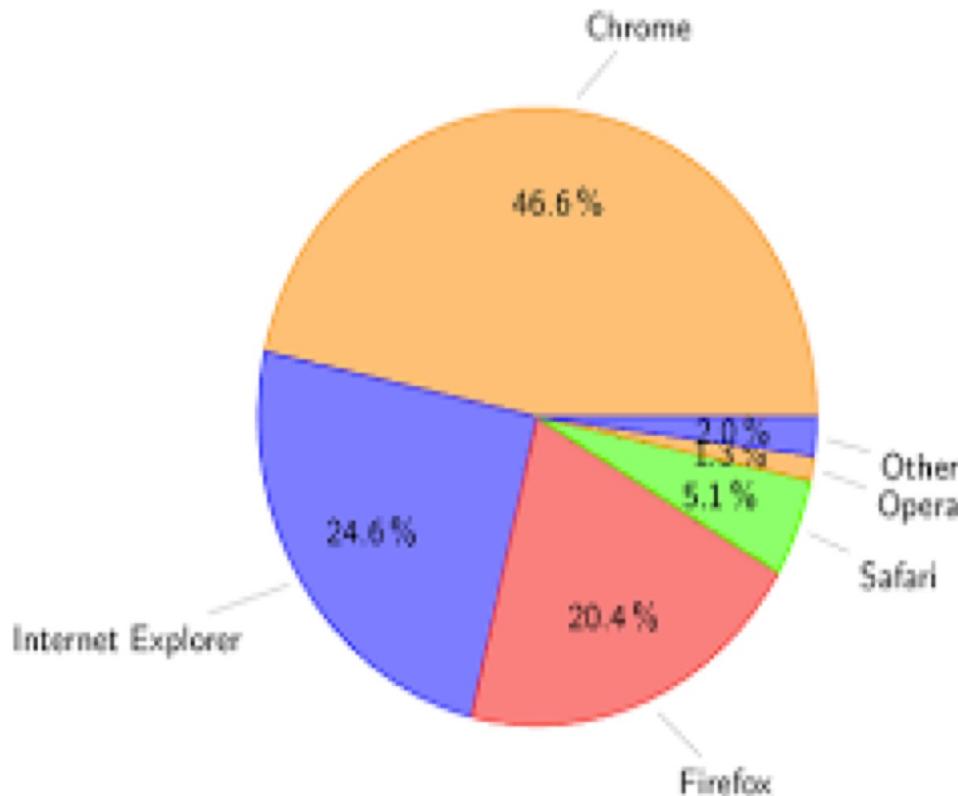


Diagrama de dispersión

- Los valores de cada atributo determinan la posición.
- Los más comunes son los de 2 dimensiones, pero también se usan los de 3.
- A veces se muestran atributos adicionales haciendo diferentes los puntos: modificando su tamaño, forma o el color.
- Son muy útiles para mostrar de forma intuitiva las relaciones entre parejas de atributos

Diagrama de dispersión

Ejemplo: Ventas de helados vs Temperatura.

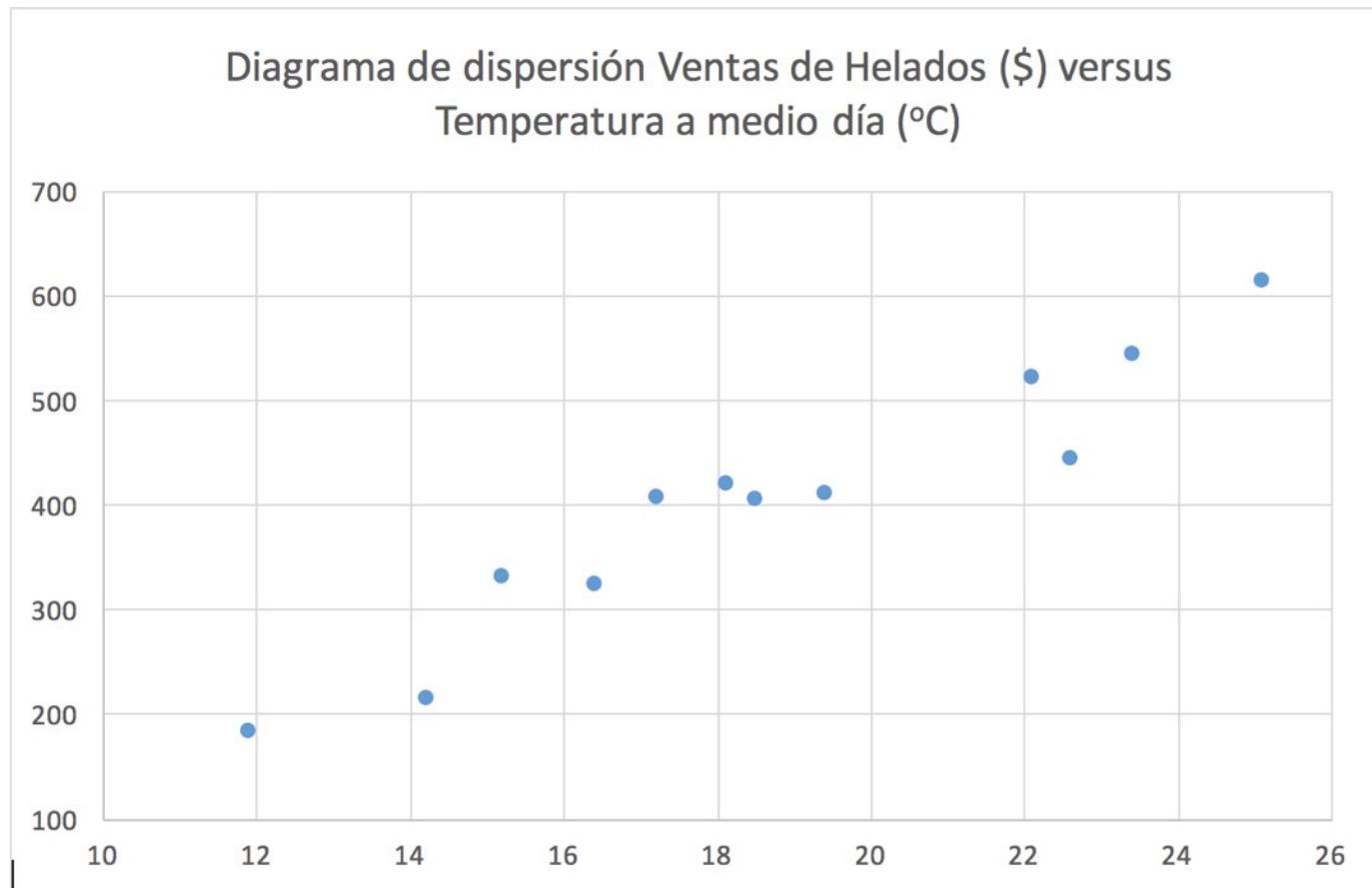


Diagrama de dispersión

Ejemplo: Tiempo estudiando vs Tiempo viendo TV.

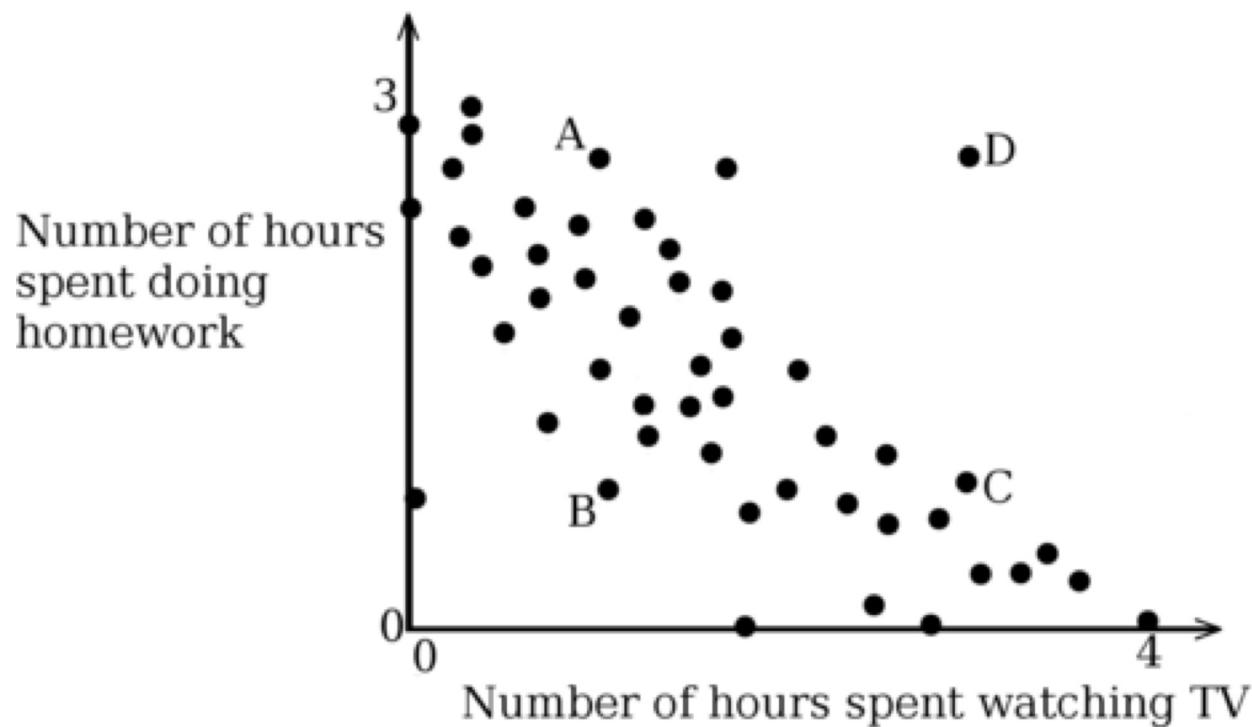


Diagrama de burbujas

- Es un tipo especial de diagrama de dispersión al que se le introduce una tercera variable que indica el tamaño.
- Si hay muchas instancias puede ser difícil de leer.

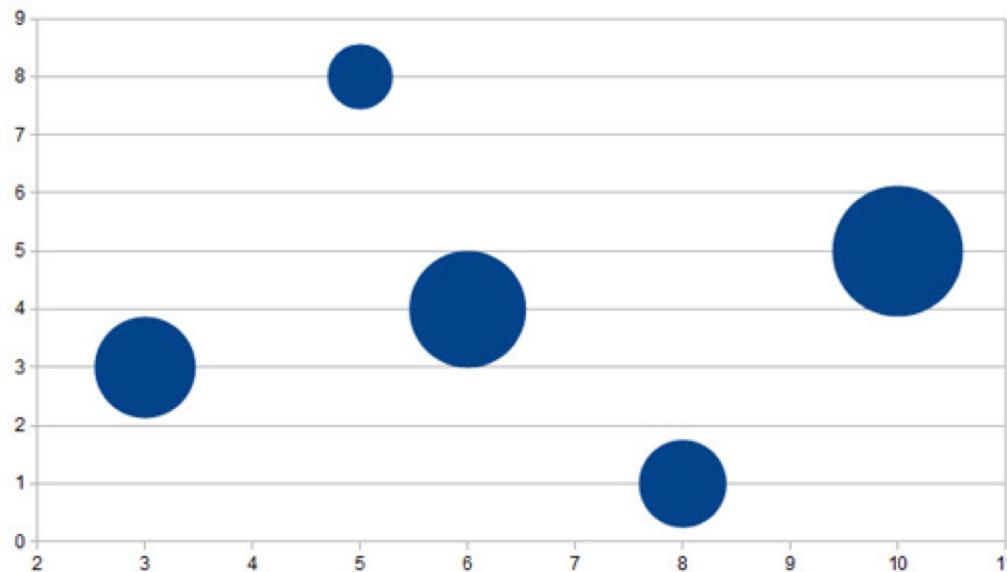


Diagrama de burbujas

Ejemplo: Millas de carreteras públicas (Eje X), AADT (Eje Y), Número de muertes (Tamaño burbuja). ¿Se podría meter otra dimensión?

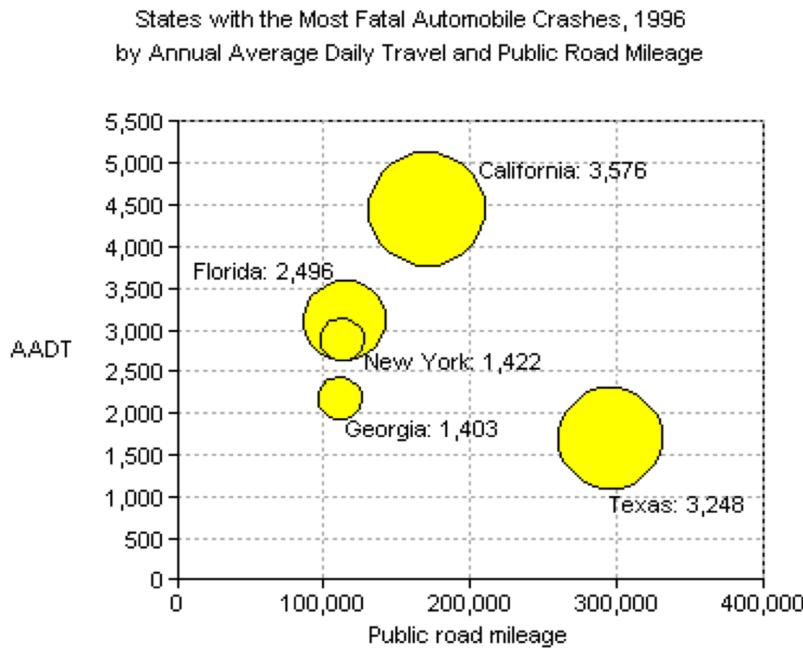
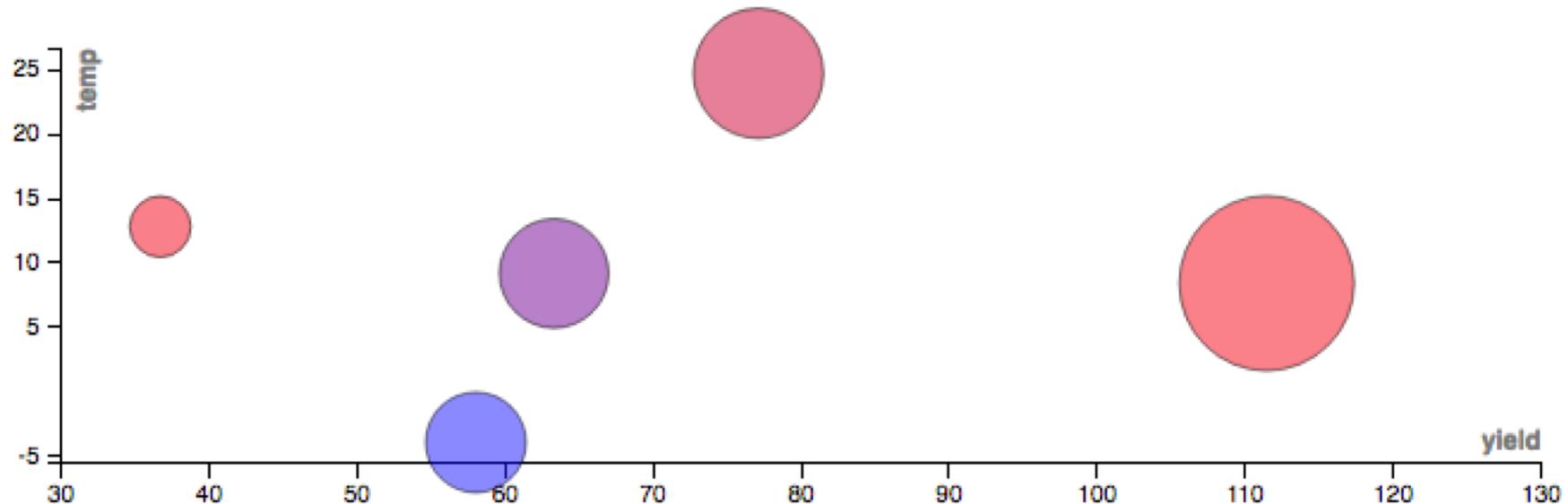


Diagrama de burbujas

Ejemplo: Rendimiento agrícola (Eje X), Temperatura media (Eje Y), PIB (Tamaño burbuja). Si añadimos tiempo como los años (Las burbujas se mueven).

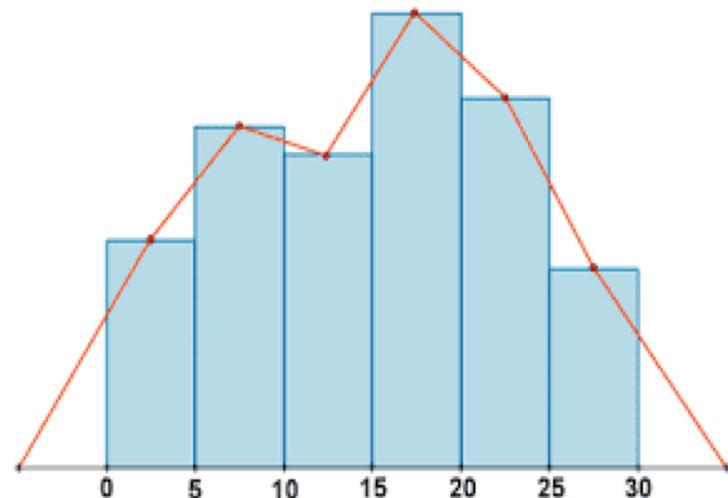


Polígono de frecuencias → Diagrama líneas

- Esta representación se basa en el histograma
- Sólo es útil para variables cuantitativas.
- Los puntos que permiten la unión de las líneas representa el centro de clase y el valor de un punto en el caso de crear un diagrama de líneas.

Polígono de frecuencias → Diagrama líneas

- Polígono de frecuencias



- Diagrama de líneas

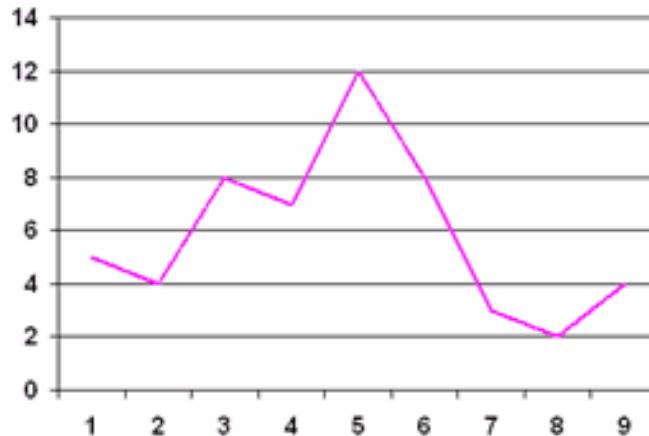


Diagrama de líneas

Ejemplo: La temperatura del día por horas.



28 ^{°C | °F}

Precipitaciones: 0%
Humedad: 36%
Viento: 6 km/h

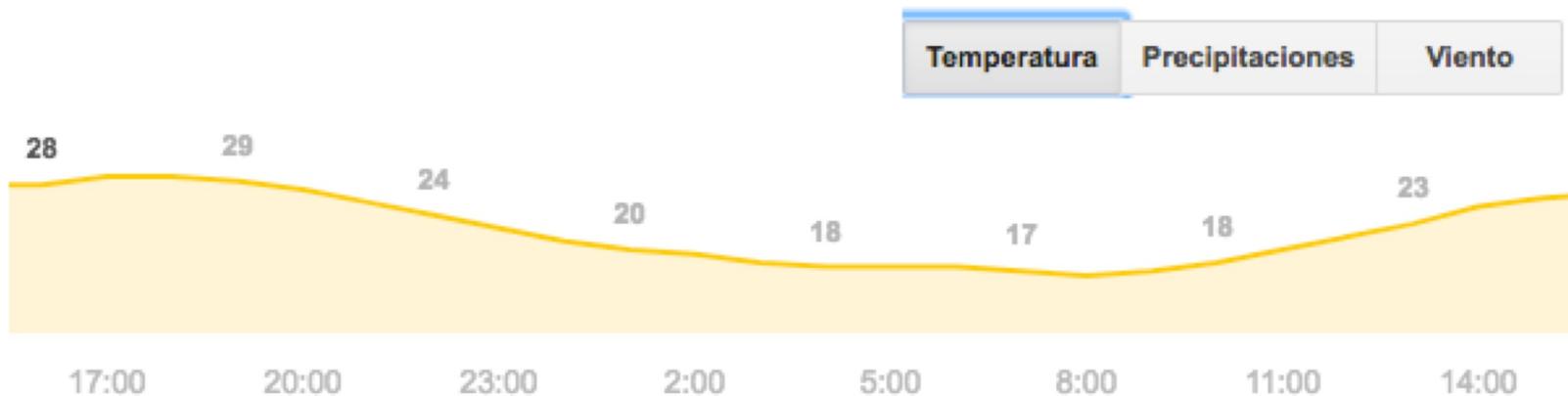


Diagrama de líneas

Ejemplo: Comparar dos empresas en la bolsa.

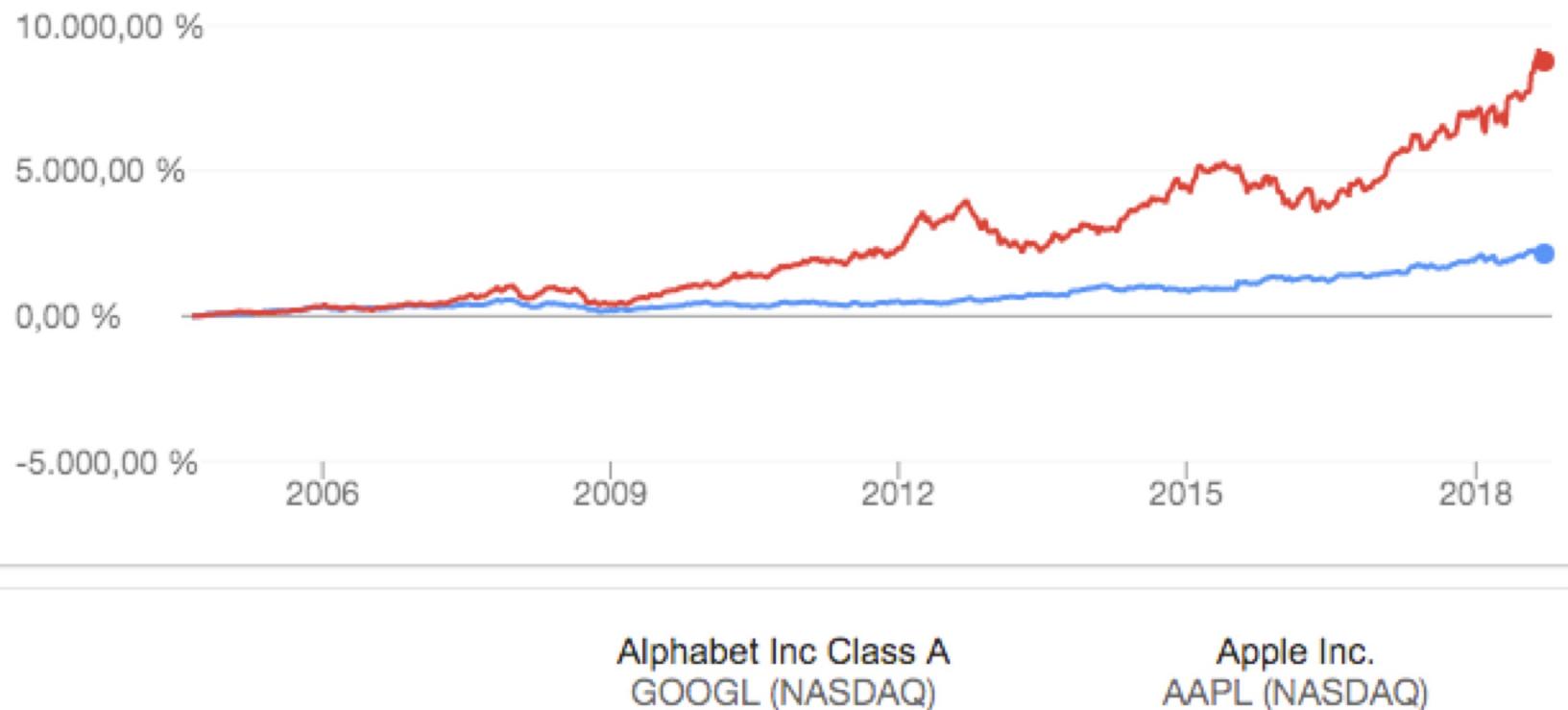


Diagrama de tallos y hojas

- Permite obtener simultáneamente una distribución y las frecuencias de la variable junto a su representación gráfica.

4	3	5	6	6	9
5	0	0	1	2	2
6	0	0	0	1	1
7	0	0	0	2	2
8	1	2	2	5	5
9	2				

Diagrama de tallos y hojas

- Como se crea:
 1. Ordenar los valores de menor a mayor o viceversa.
 2. Elegir el tallo y separarlo de las hojas.
 3. Dibujar los tallos de arriba hacia abajo y sus hojas asociadas de izquierda a derecha.

Diagrama de tallos y hojas

Ejemplo: Los horarios de un tren entre dos ciudades.

05	03
06	02 18 37 48 55
07	02 07 20 25 32 37 50
08	02 05 20 24 32 37 51
09	02 07 24 32 37
10	02 07 32 37
11	02 07 32 37
12	02 07 32 37
13	02 07 20 32 37 50
14	02 07 20 32 37 50
15	02 07 20 32 37 50
16	02 07 20 32 37 50
17	02 07 20 32 37 50
18	02 07 20 32 37 50
19	02 07 20 32 37 50
20	02 07 20 32 37 50
21	02 07 20 32 37
22	38

Diagrama de tallos y hojas

Ejemplo: Las longitudes de una muestra.

Tallo	Hoja
14	5 7 9
15	2 3 4 4 6 7 8
16	1 2 2 3 3 4 4 5 7 7 8 9 9
17	0 1 1 2 3 4 4 5 6 6 8 9
18	0 1 3 5 6

Diagrama de cajas

- Un diagrama de cajas y bigotes es una manera conveniente de mostrar visualmente grupos de datos numéricos a través de sus cuartiles.

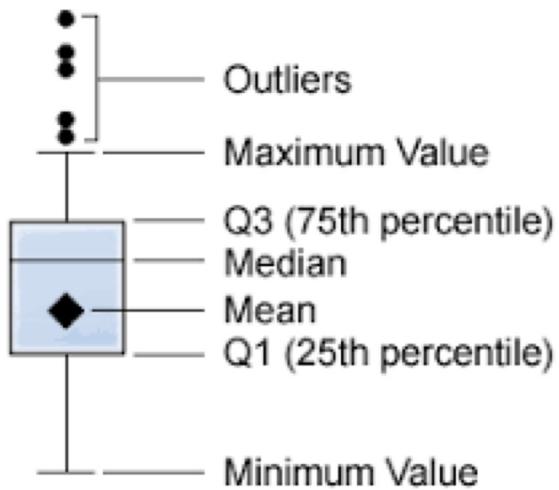


Diagrama de cajas

- Como se crea:

1. Dada un dataset con la edad de 20 personas.

36 25 37 24 39 20 36 45 31 31

39 24 29 23 41 40 33 24 34 40

2. Ordenar la distribución

20 23 24 24 24 25 29 31 31 33

34 36 36 37 39 39 40 40 41 45

Diagrama de cajas

3. Q_1 , es el valor mayor que el 25% de los valores de la distribución. N° de individuos=20

$$20/4 = 5$$

Q_1 es la media del quinto valor y el siguiente:

$$Q_1 = (24+25) / 2 = 24,5$$

4. Q_2 , es la mediana de la distribución, es el valor de la variable que ocupa el lugar central (50%).

$$20/2 = 10 \text{ entonces } Q_2 = (33+34)/2 = 33,5$$

Diagrama de cajas

5. Q_3 , es el valor que mayor que el 75% de los valores de la distribución.

$$3*20 / 4 = 15$$

$$Q_3=(39+39) / 2 = 39$$

6. Valores atípicos son aquellos que $<Q_1-1,5*IQR$ o $>Q_3+1,5*IQR$ teniendo en cuenta que $IQR=Q_3-Q_1$
Por lo tanto los lim. inferior y superior son el menor y el mayor de aquellos que no son atípicos.

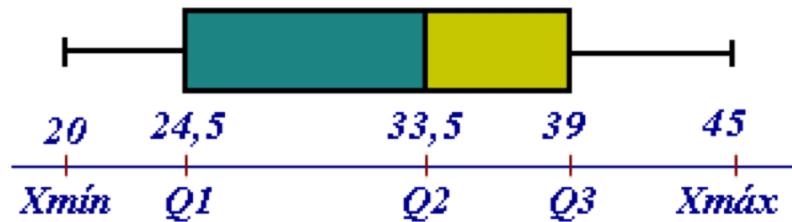
Diagrama de cajas

$$IQR = 39 - 24,5 = 14,5$$

Valores atípicos $<2,75$ y $>60,75 \rightarrow \text{NO HAY!!!}$

$$LI=20 \ LS=45$$

7. Se dibuja la caja y los bigotes con los datos.



¿Qué se puede interpretar del gráfico?

Diagrama de cajas

- Que se puede interpretar:

1. La parte izquierda de la caja es mayor que la de la derecha; ello quiere decir que las edades comprendidas entre el 25% y el 50% de la población está más dispersa que entre el 50% y el 75%.

Diagrama de cajas

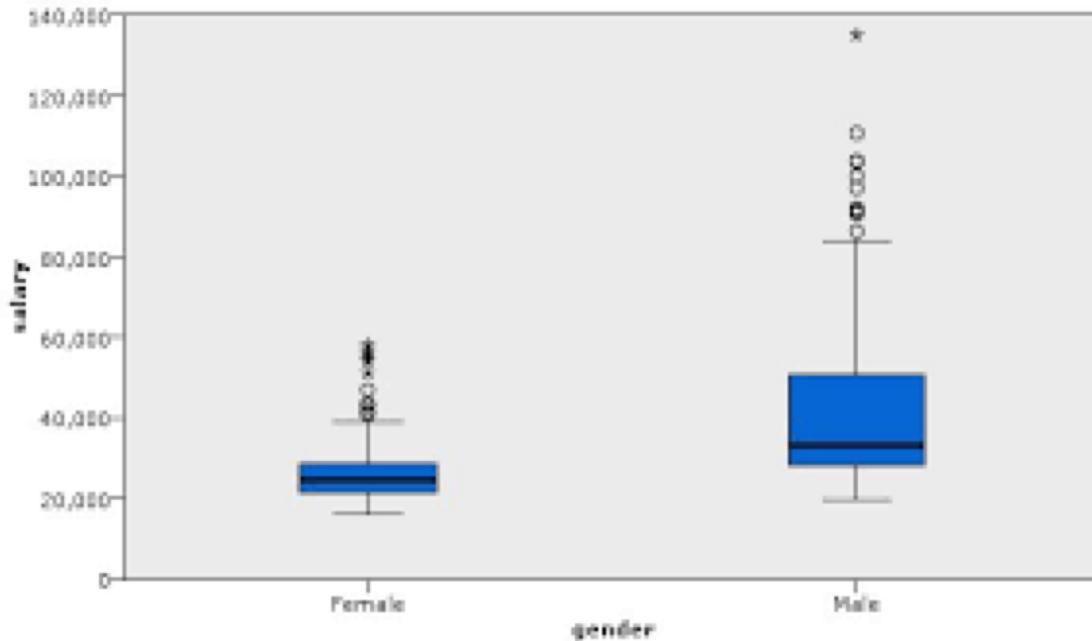
2. El bigote de la izquierda ($X_{mín}$, Q_1) es más corto que el de la derecha; por ello el 25% de los más jóvenes están más concentrados que el 25% de los mayores.
3. El *rango intercuartílico* = $Q_3 - Q_1 = 14,5$; es decir, el 50% de la población está comprendido en 14,5 años.

Diagrama de cajas

4. Si hay valores atípicos, cuáles y sus valores.
5. Los datos son asimétricos.

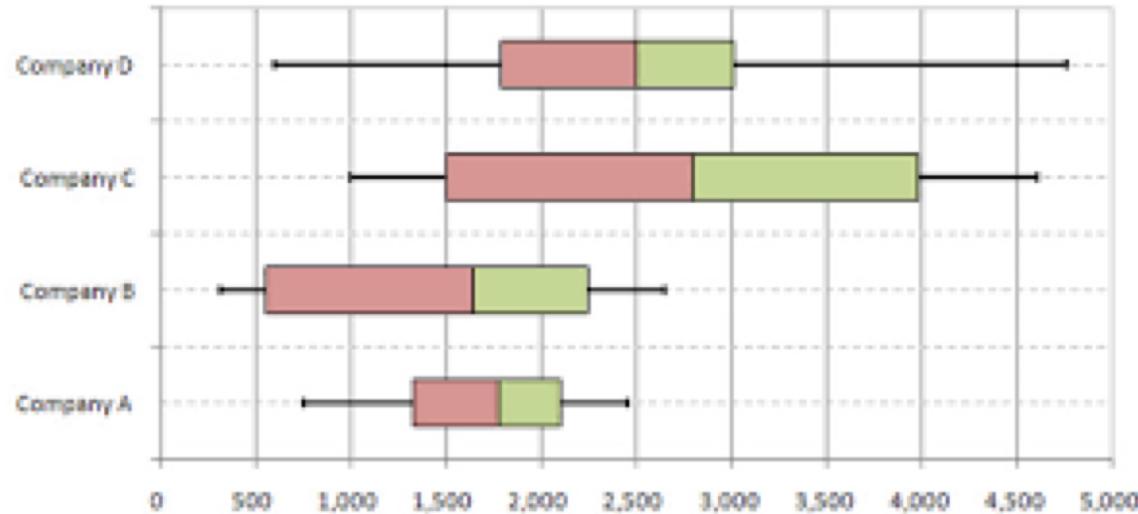
Diagrama de cajas

Ejemplo: Comparar salarios entre genero.



Polígono de frecuencias

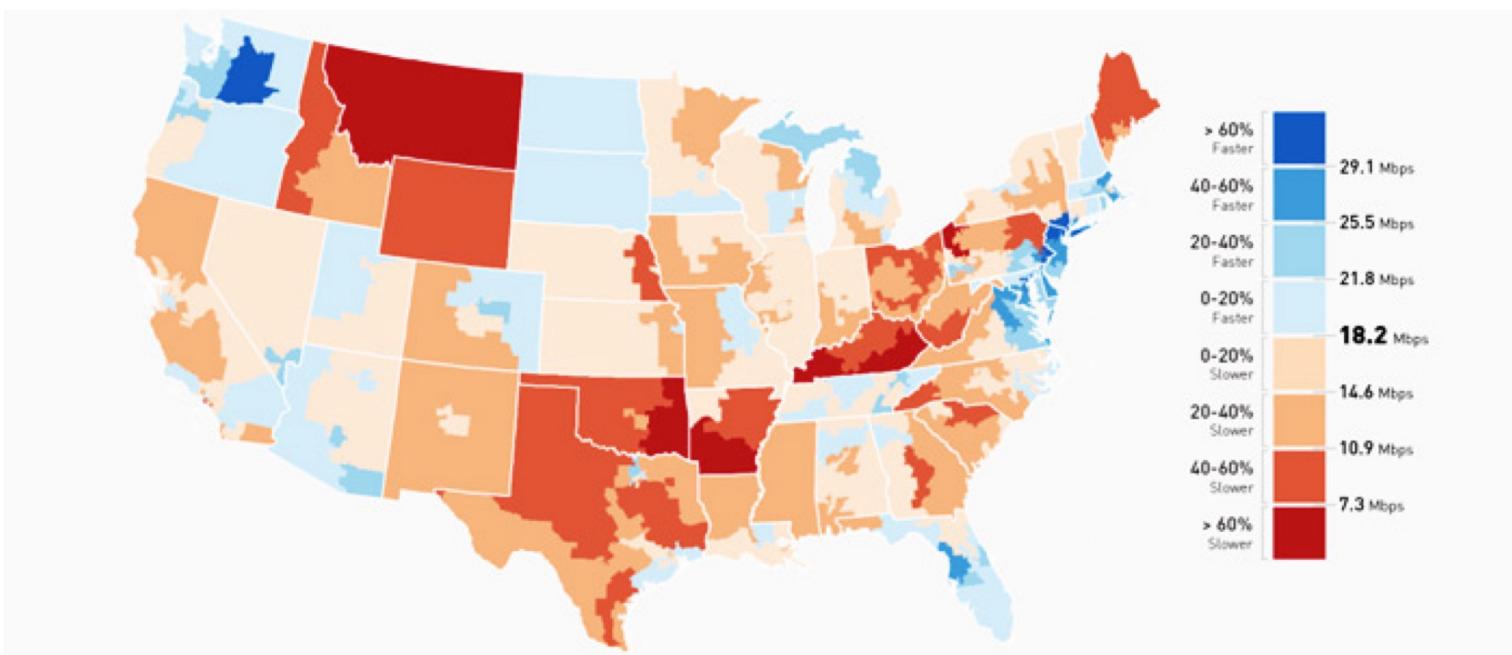
Ejemplo: Salarios entre distintas empresas.



Otros: mapas coropléticos

- Muestran zonas geográficas o regiones divididas en colores, con sombras o dibujos en relación con una variable de datos.

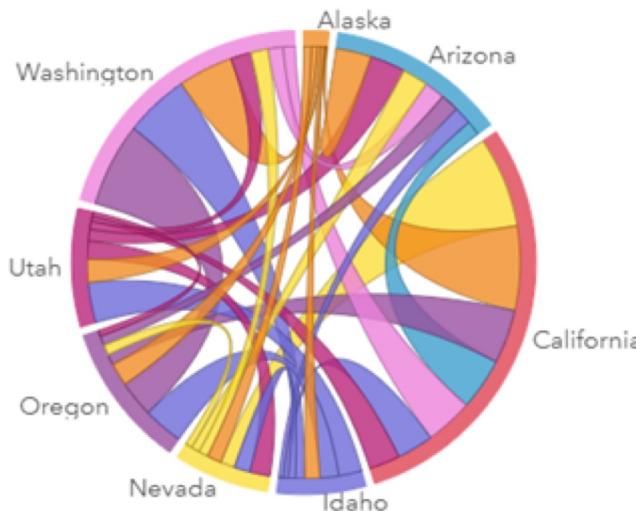
Velocidad de Internet en USA



Otros: diagrama de cuerdas

- Visualizar las interrelaciones entre entidades. Las conexiones entre las entidades se utilizan para mostrar el hecho de que comparten algo en común.

Flujo de migración



Otros: diagrama de diamante

- Permite mostrar visualmente qué atributos se asocian con un conjunto de individuos

Características de jugadores de fútbol



Otros: mapas de calor

- Visualizar datos a través de las variaciones de color.

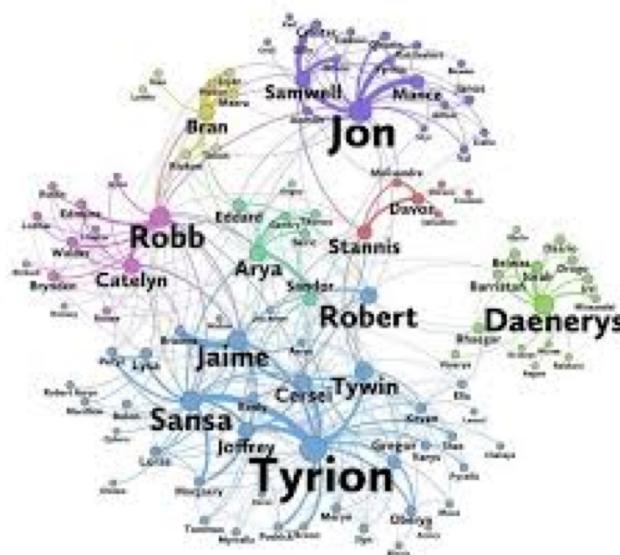
Ventas por meses de los empleados

	William Brown	Candy Farrell	Dylan Dodson	Gloria Berrington	Shanna Barrington	Doris Derrick	Luke Gilson	Lewis Gardner
Jan	\$65105	\$51822	\$56606	\$79225	\$89419	\$43012	\$51335	\$25133
Feb	\$19045	\$110424	\$78979	\$92798	\$86699	\$82851	\$47829	\$30739
Mar	\$62027	\$17851	\$60183	\$111425	\$50378	\$80876	\$36999	\$99908
Apr	\$23523	\$32915	\$78491	\$76731	\$91094	\$64318	\$58721	\$102126
May	\$80205	\$29441	\$57456	\$50683	\$100547	\$35440	\$110832	\$23017
Jun	\$70405	\$89681	\$15134	\$90672	\$46091	\$46407	\$65346	\$84899

Otros: grafos

- Entender como se conectan los diferentes individuos

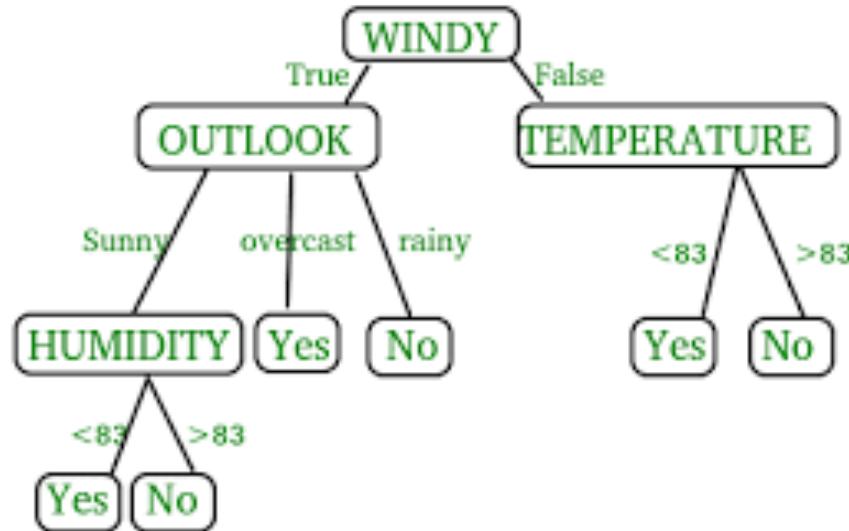
Game of Thrones



Otros: dendrograma

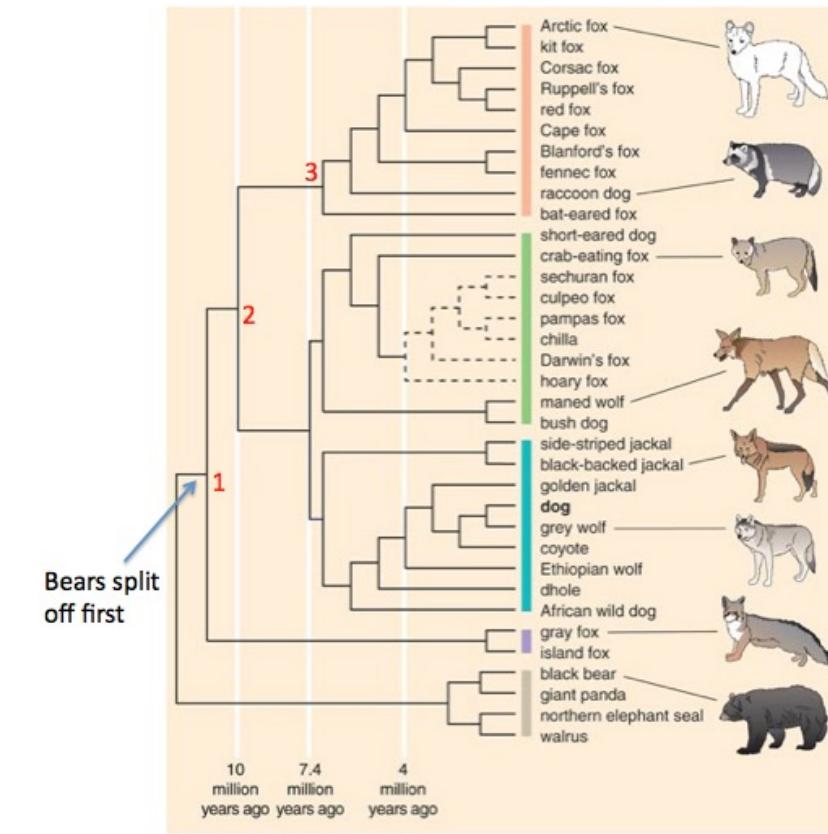
- Entender las jerarquías mediante una representación en forma de árbol. Lo usaremos en los árboles de decisión y en los clusters jerárquicos.

Árboles de decisión



Otros: dendrograma

Clustering jerárquico



Ejercicio

Plantear una hipótesis y aceptarla o refutarla mediante el uso de gráficas. Características de diferentes países.

País	Población	Costa (km/extensión)	Migración neta	Renta per cápita
1	1.313.973.713	0,15	-0,4	5.000
2	1.095.351.995	0,21	-0,07	2.900
3	298.444.215	0,21	3,41	37.800
4	65.773	194,34	2,49	3.600
5	11.382.820	3,37	-1,58	2.900
6	47.246	79,84	1,41	2.200
7	108.004	8.070,66	-20,90	2.000
8	57.794	58,29	-20,70	8.000
9	31.056.997	0,00	23,06	700
10	45.436	61,07	18,75	3.500