

Códigos Visuales

Eduardo Fernández

Mapeo de Datos a Variables Visuales

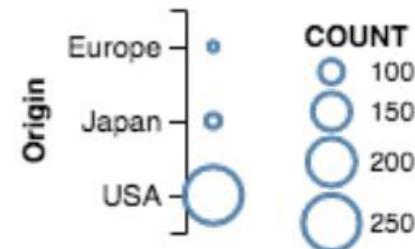
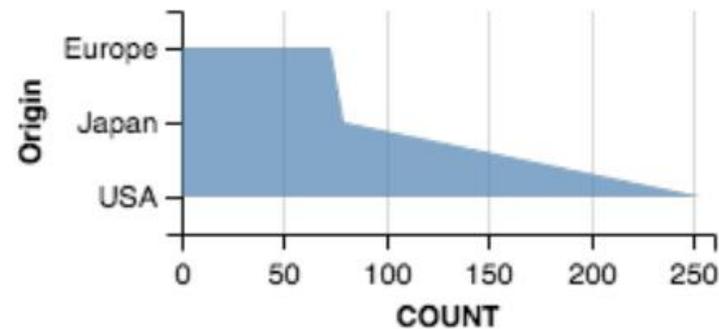
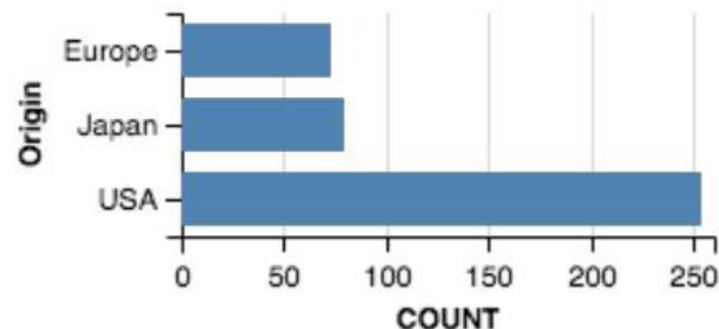
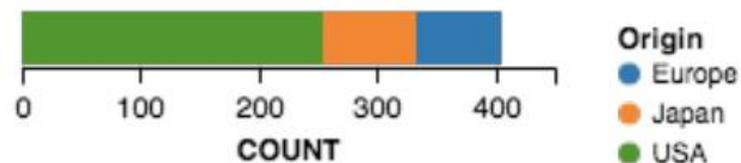
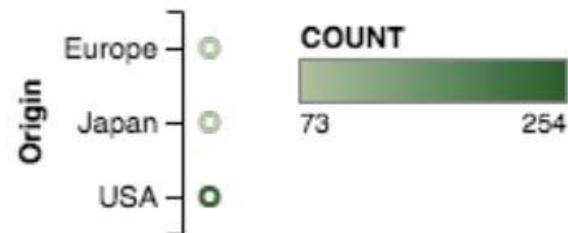
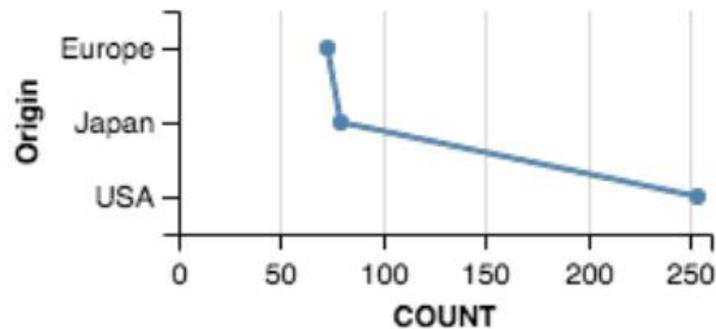
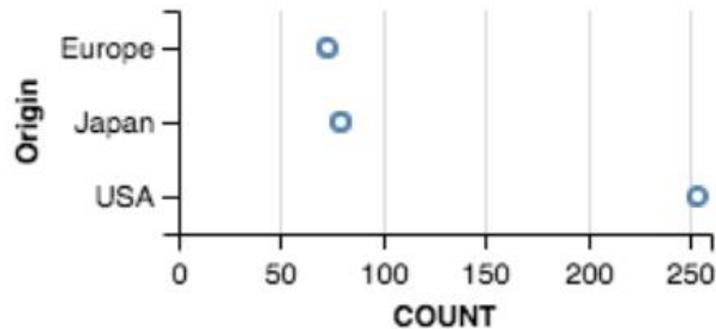
- Asignación de variables (Nominal, Ordinal, Cuantitativo) a canales visuales (x, y, color, forma, tamaño, ...) para un tipo de marca gráfica (punto, barra, línea, ...).
- Además: cómo realizar la codificación apropiada (escala lineal, logarítmica, ordenamiento, ...) y transformación de datos (discretizado -o binning- , agrupar, agregar, etc.)
- Estas opciones definen muchas combinaciones posibles, conteniendo visualizaciones útiles y cuestionables.

1D: Nominal

Raw

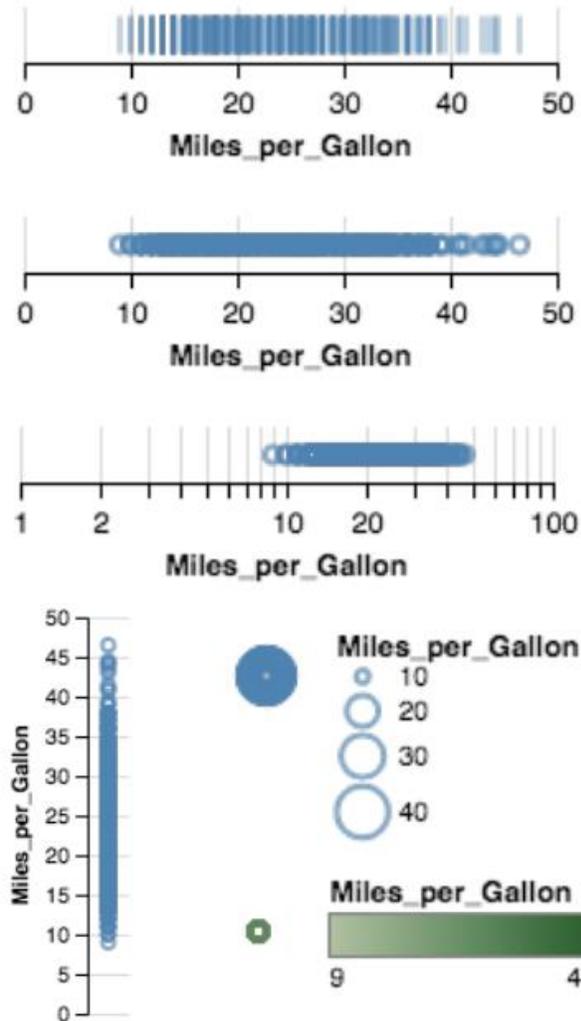


Aggregate (Count)

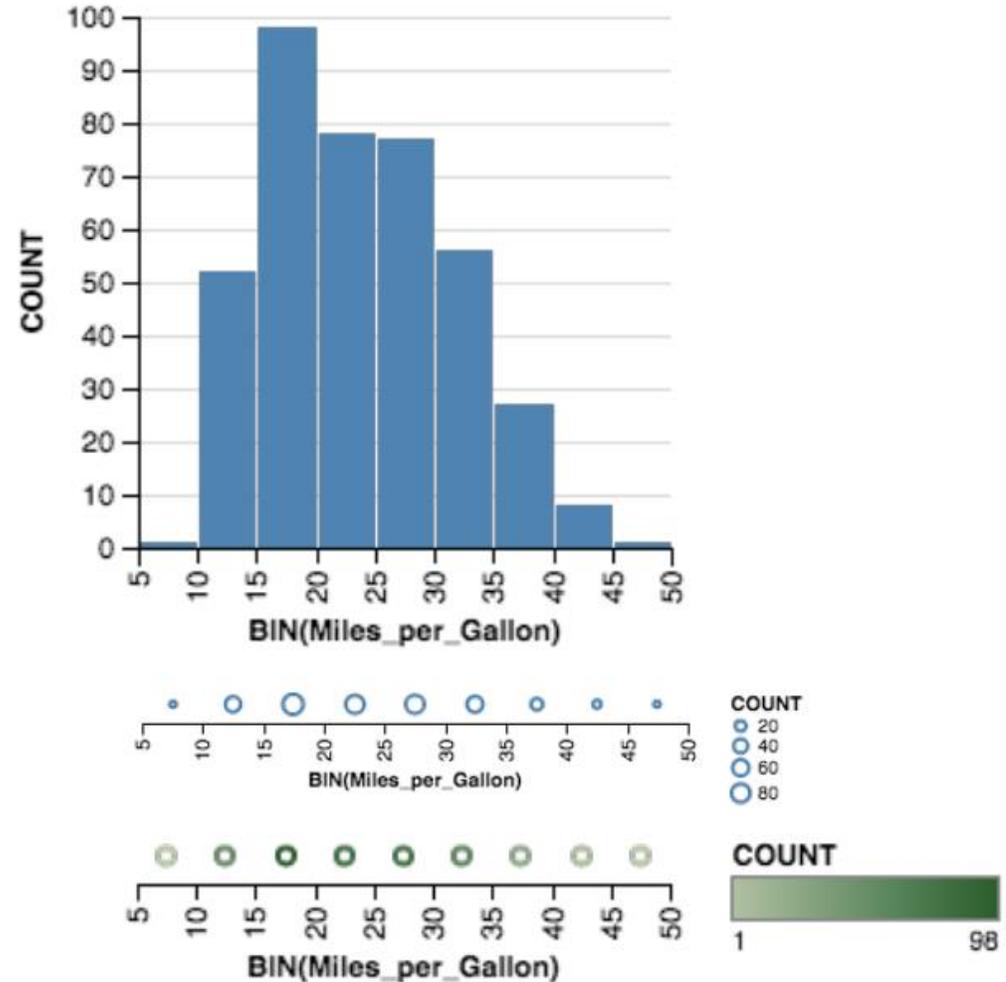


1D: Cuantitativo

Raw

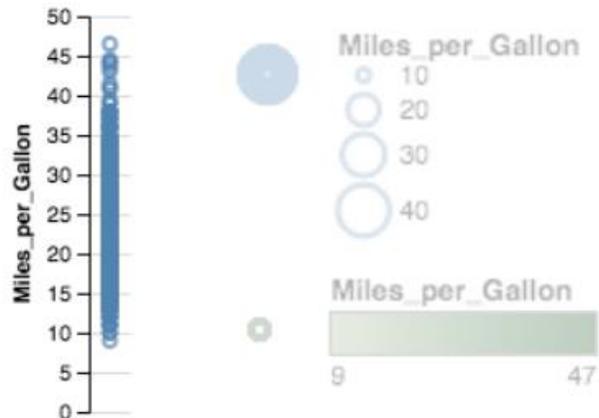
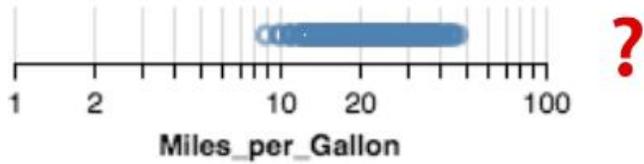
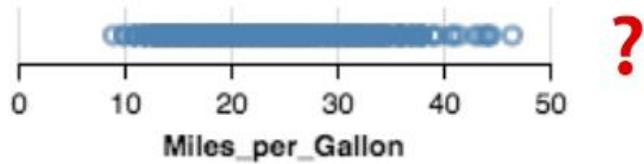
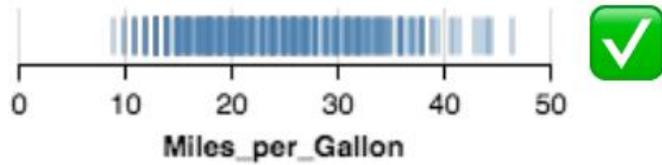


Aggregate (Count)

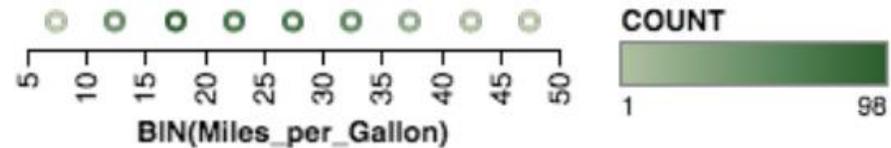
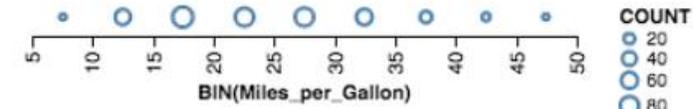
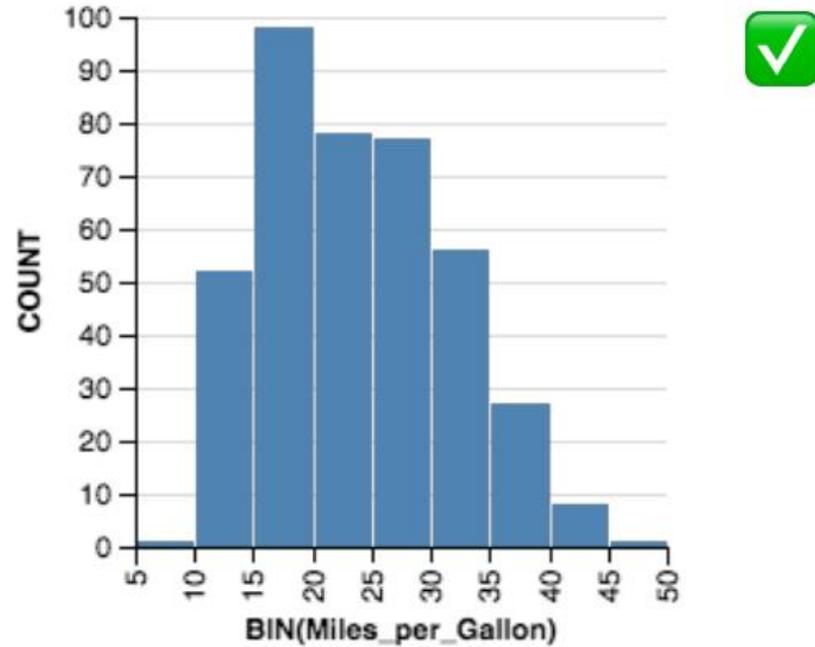


¿Efectividad?

Raw

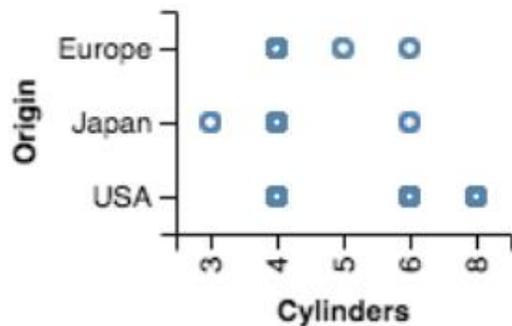


Aggregate (Count)

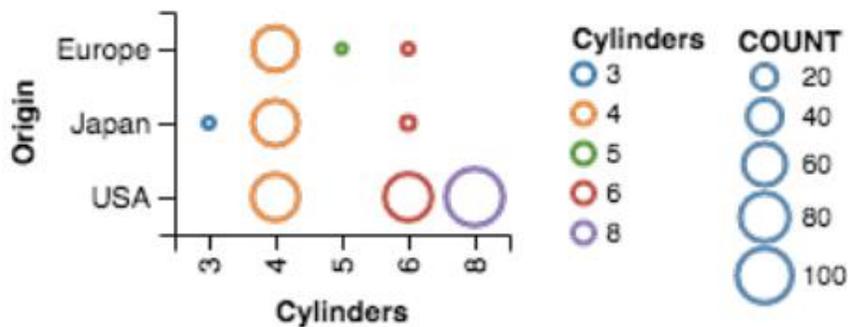
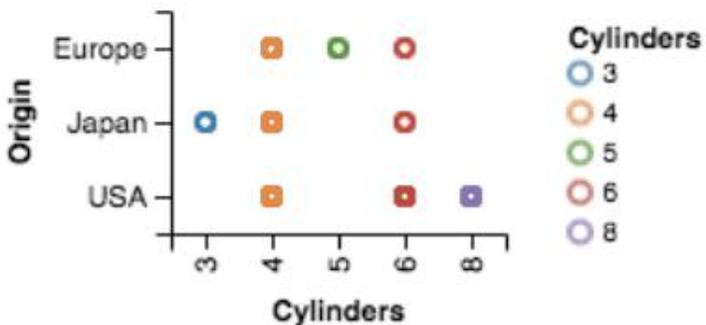
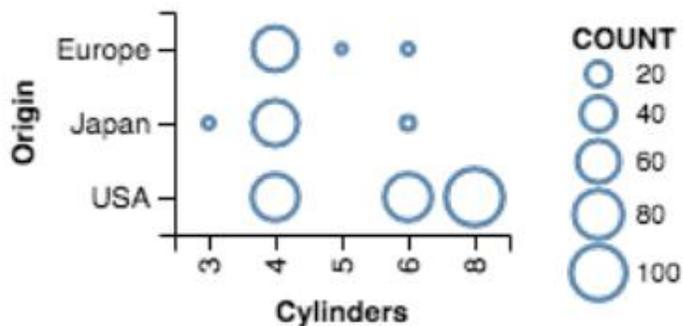
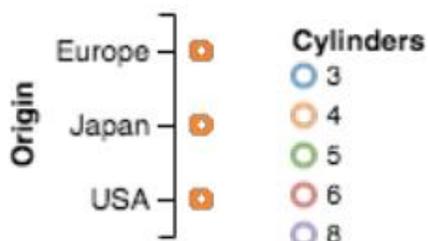
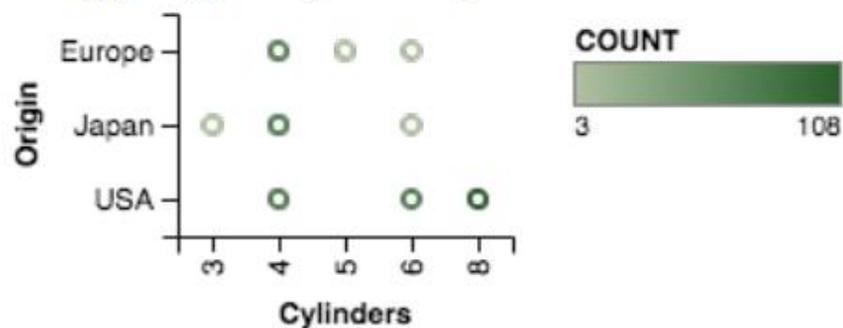


2D: Nominal x Nominal

Raw

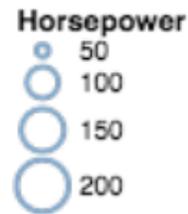
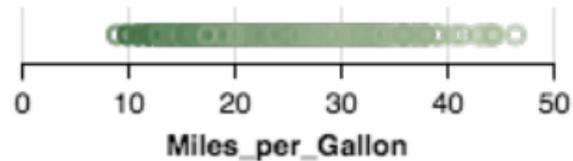
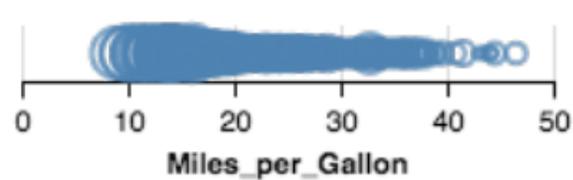
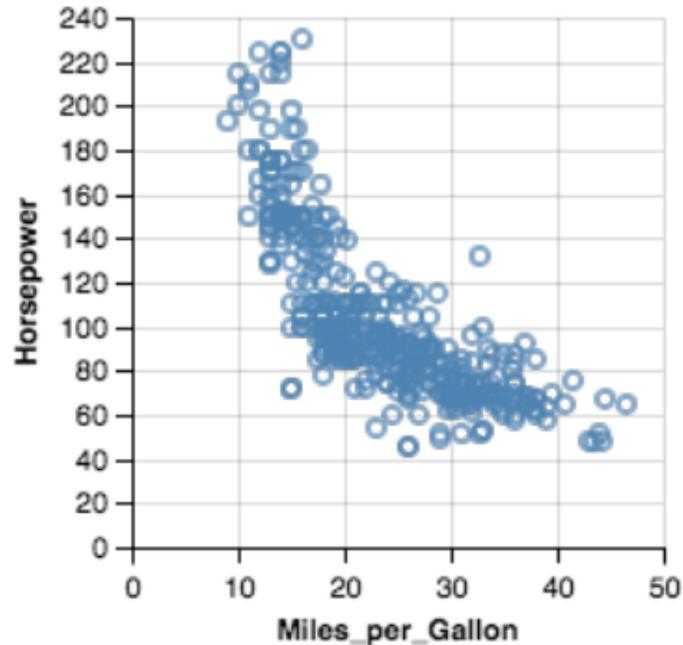


Aggregate (Count)

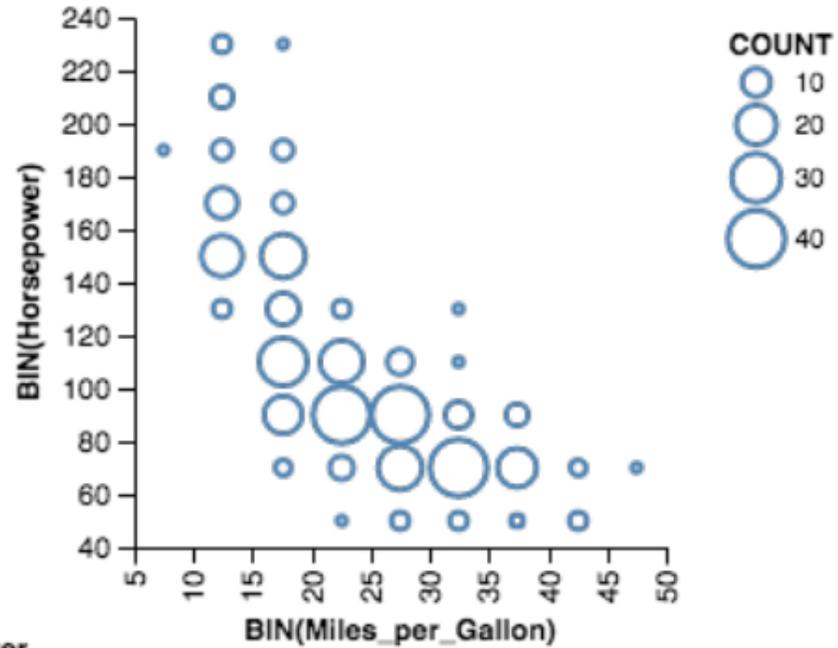


2D: Cuantitativo x Cuantitativo

Raw

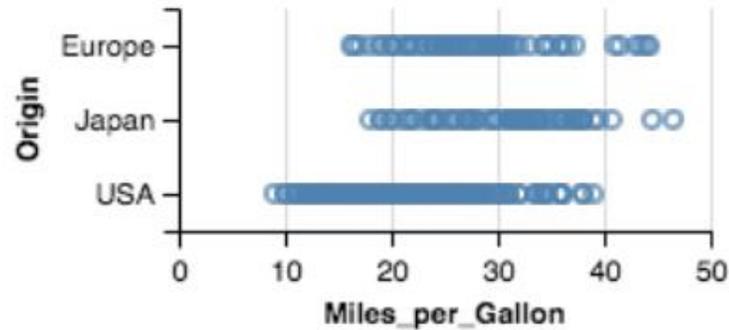


Aggregate (Count)

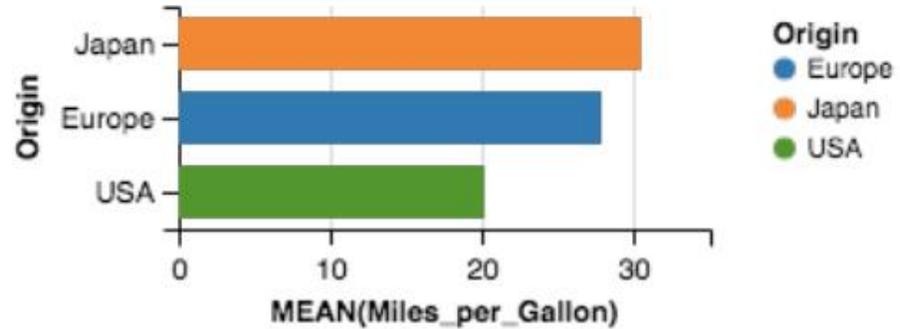
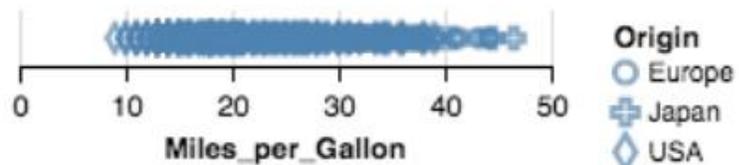
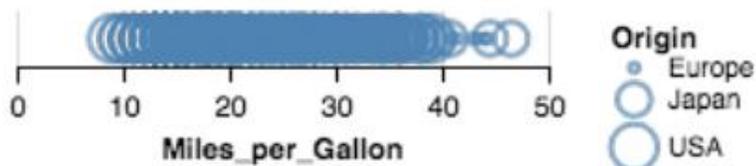
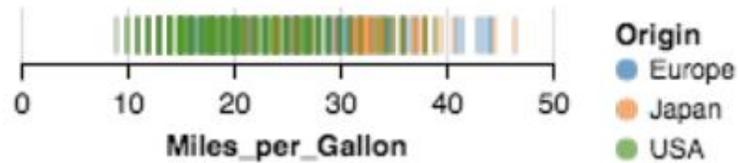
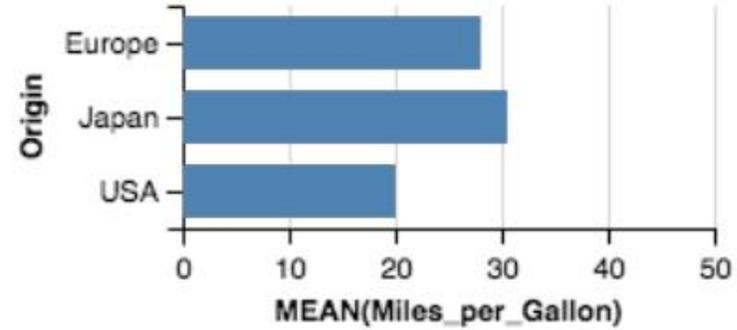


2D: Nominal x Cuantitativo

Raw

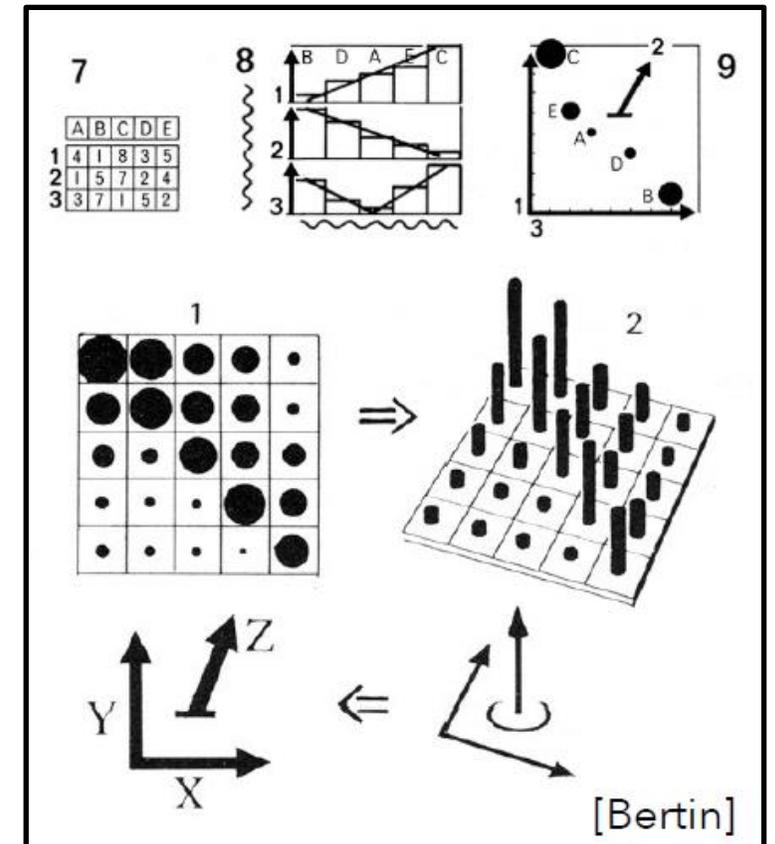
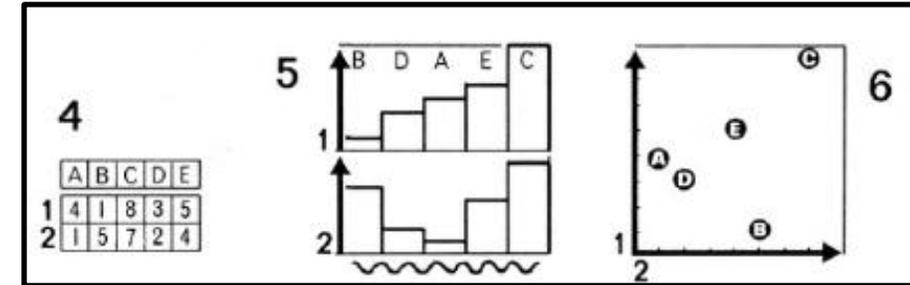
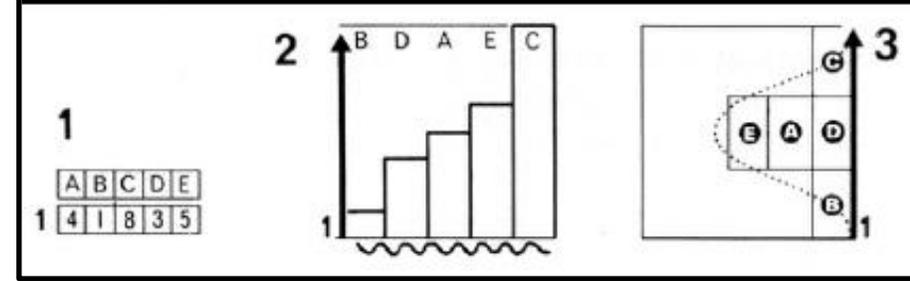


Aggregate (Mean)



3D y más

- 2 variables se mapean a gráficos 2D.
- La 3er variable se mapea a color, tamaño, opacidad, forma, o se particiona el espacio.
- ¿Rendering 3D?



¿Otras formas de visualización?

wind map

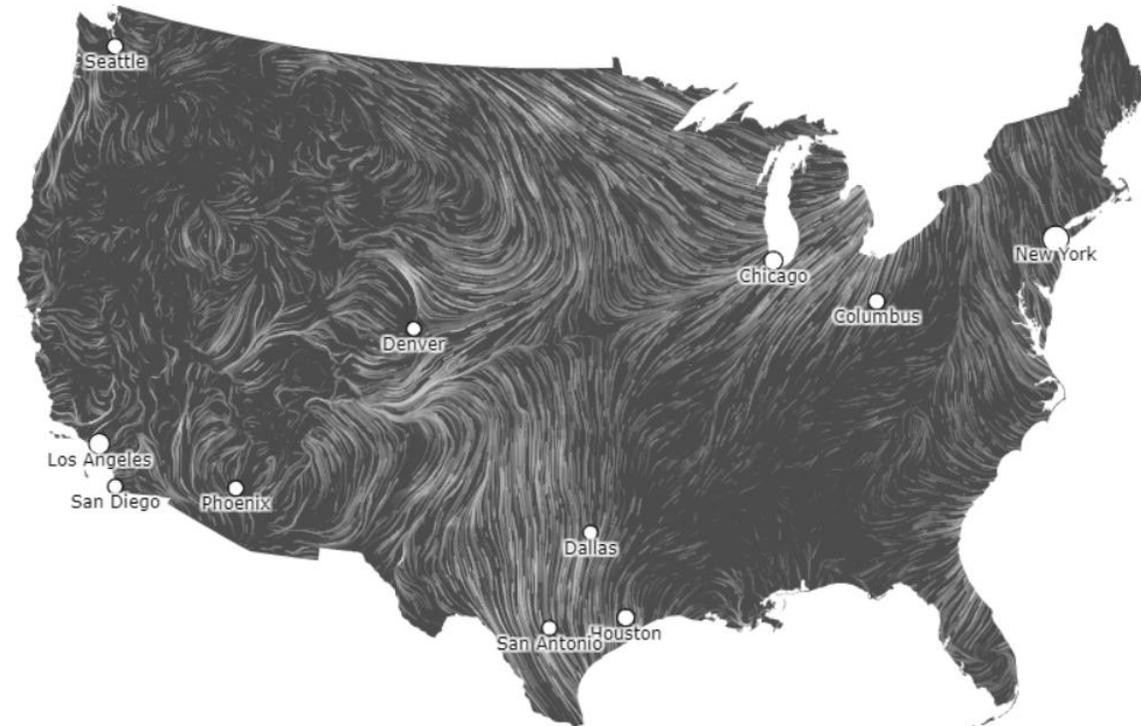
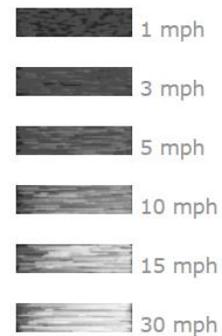
July 24, 2021

5:19 pm EST

(time of forecast download)

top speed: 29.1 mph

average: 7.2 mph



- <http://hint.fm/wind/> <http://hint.fm> <https://www.windy.com/>

Ranking de efectividad [Mackinlay 86]

QUANTITATIVE

Position

Length

Angle

Slope

Area (Size)

Volume

Density (Value)

Color Sat

Color Hue

Texture

Connection

Containment

Shape

ORDINAL

Position

Density (Value)

Color Sat

Color Hue

Texture

Connection

Containment

Length

Angle

Slope

Area (Size)

Volume

Shape

NOMINAL

Position

Color Hue

Texture

Connection

Containment

Density (Value)

Color Sat

Shape

Length

Angle

Slope

Area

Volume

Ranking de efectividad [Mackinlay 86]

QUANTITATIVE

Position
Length
Angle
Slope
Area (Size)
Volume
Density (Value)
Color Sat
Color Hue
Texture
Connection
Containment
Shape

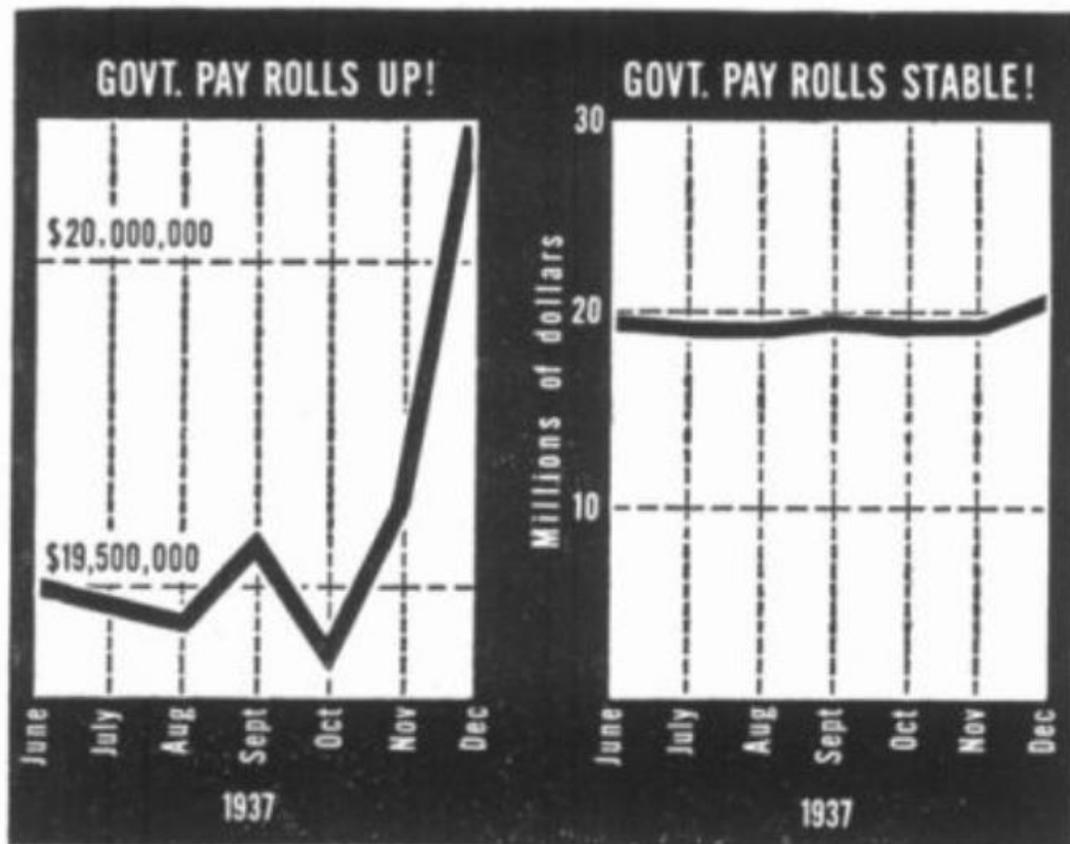
ORDINAL

Position
Density (Value)
Color Sat
Color Hue
Texture
Connection
Containment
Length
Angle
Slope
Area (Size)
Volume
Shape

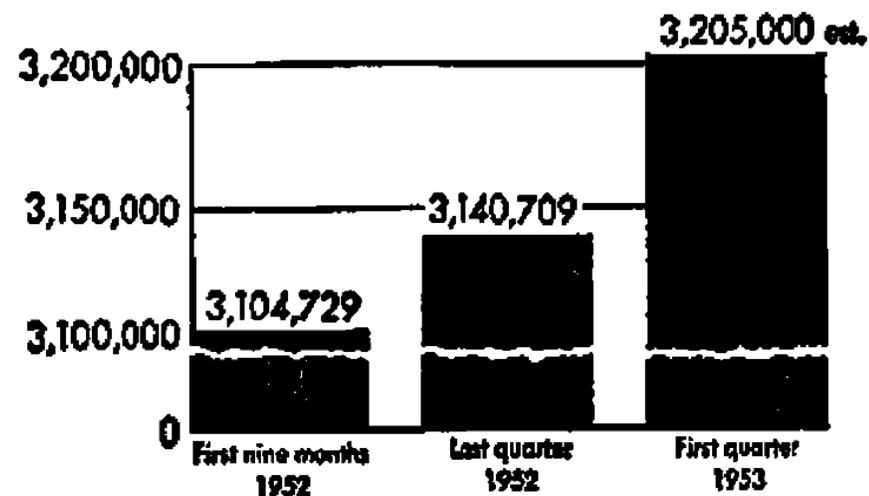
NOMINAL

Position
Color Hue
Texture
Connection
Containment
Density (Value)
Color Sat
Shape
Length
Angle
Slope
Area
Volume

Ejes y escalas: ¿Incluir el cero?



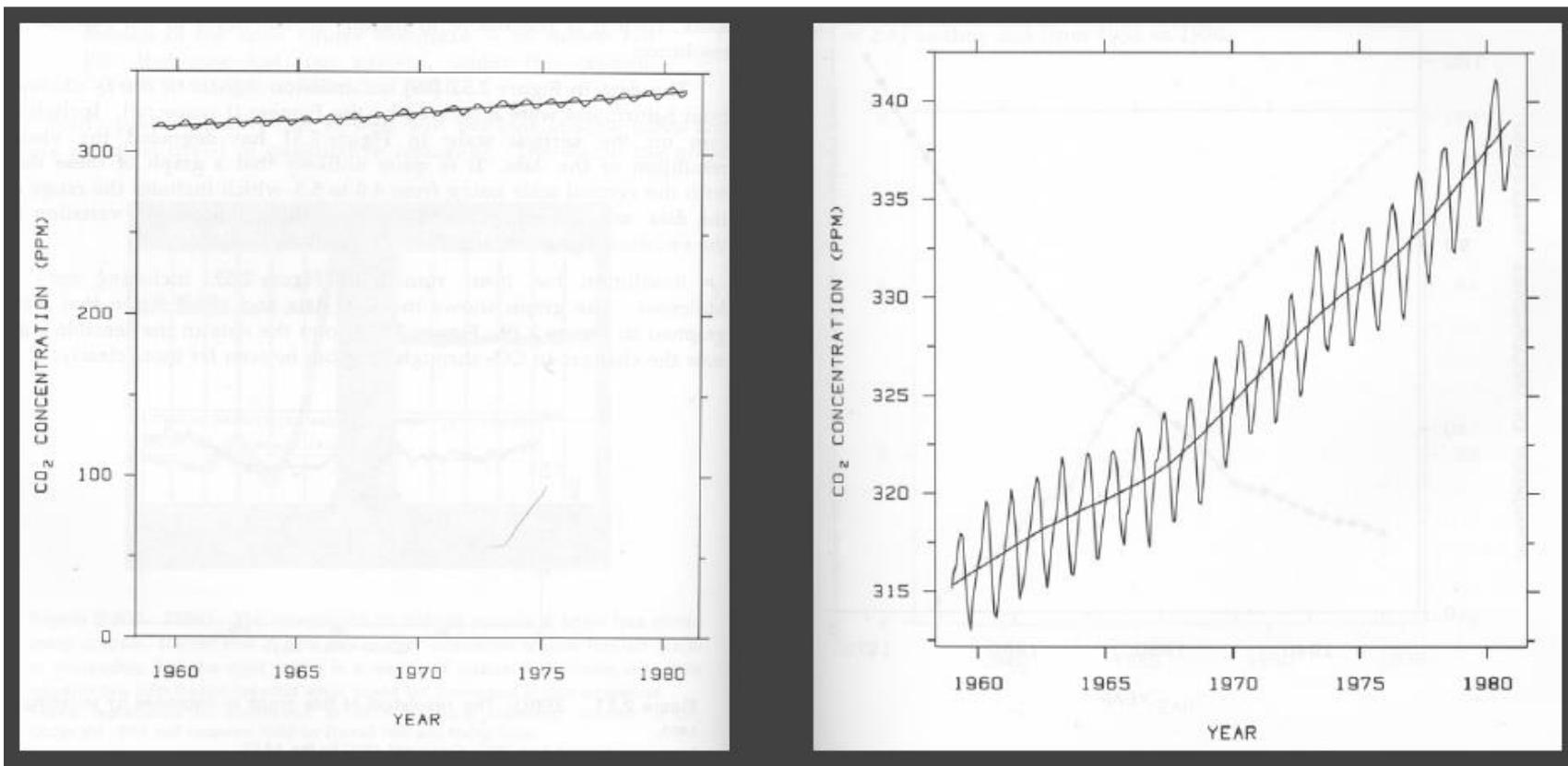
La misma información se muestra de dos formas distintas según el titular de la noticia.



From an April 24, 1953, newspaper advertisement for COLLIER'S

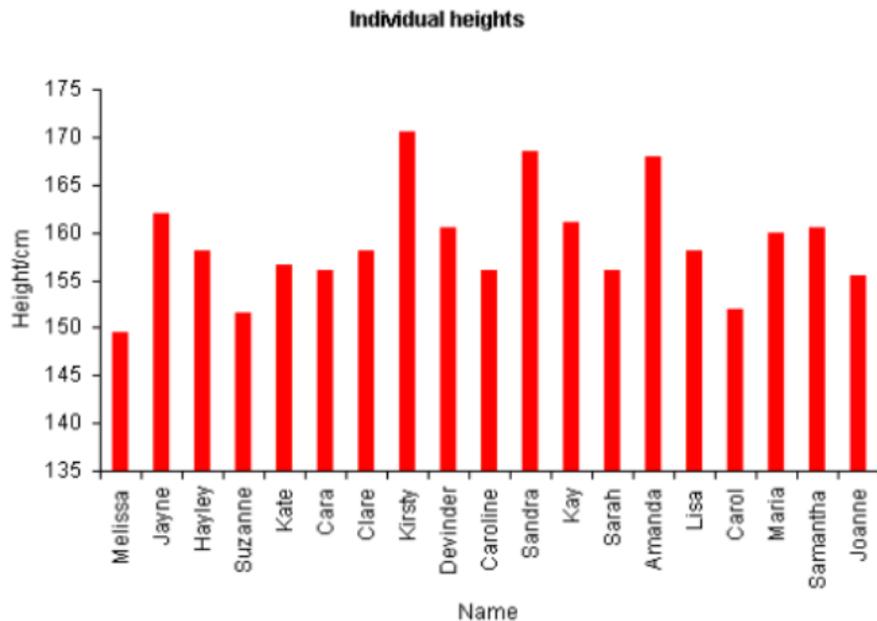
Si bien se incluye el 0, el incremento de la longitud de las barras está alterado.

Ejes y escalas: ¿Incluir el cero?



Concentración de CO₂ en la atmósfera. Por un lado se visualiza que el incremento real es “pequeño”, pero, si se quieren ver los detalles, parece imposible mantener el 0.

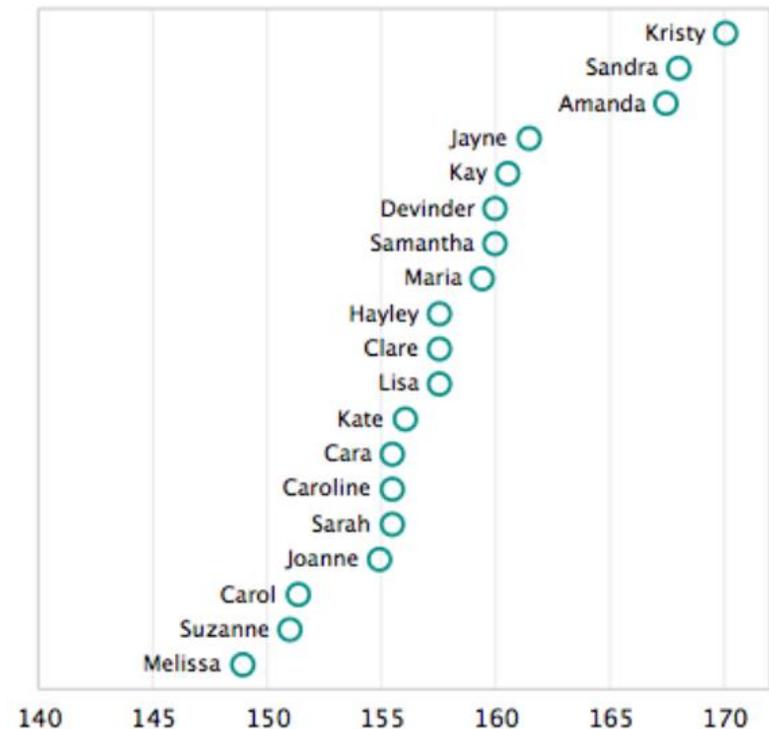
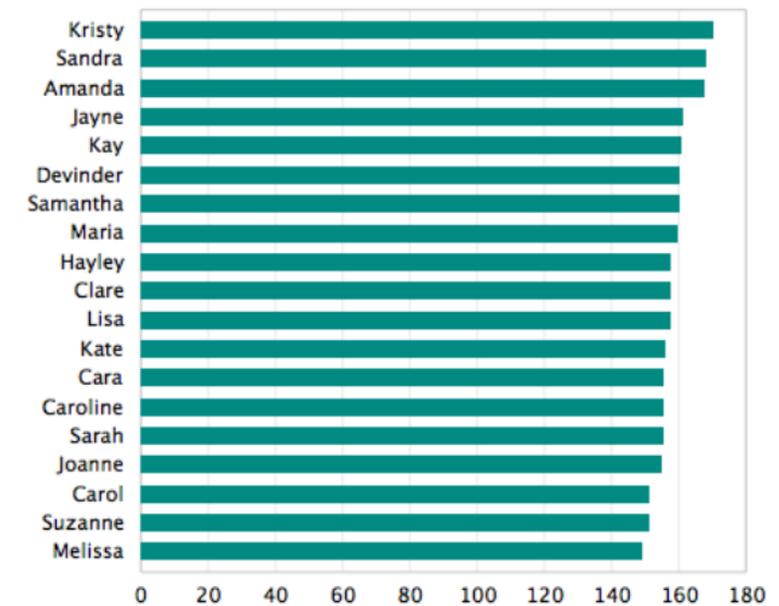
Ejes y escalas: ¿Incluir el cero?



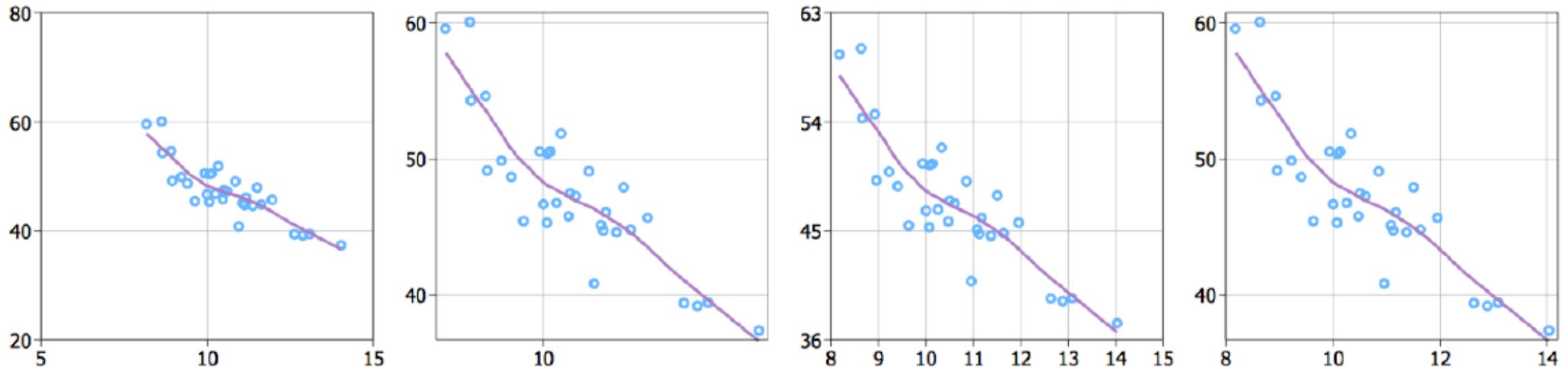
Se Viola el “**principio de expresividad**”: Se debe codificar la información y sólo la información.

Se pueden visualizar las proporciones (ratio o cociente).

Se comparan las posiciones relativas (intervalo o diferencia).

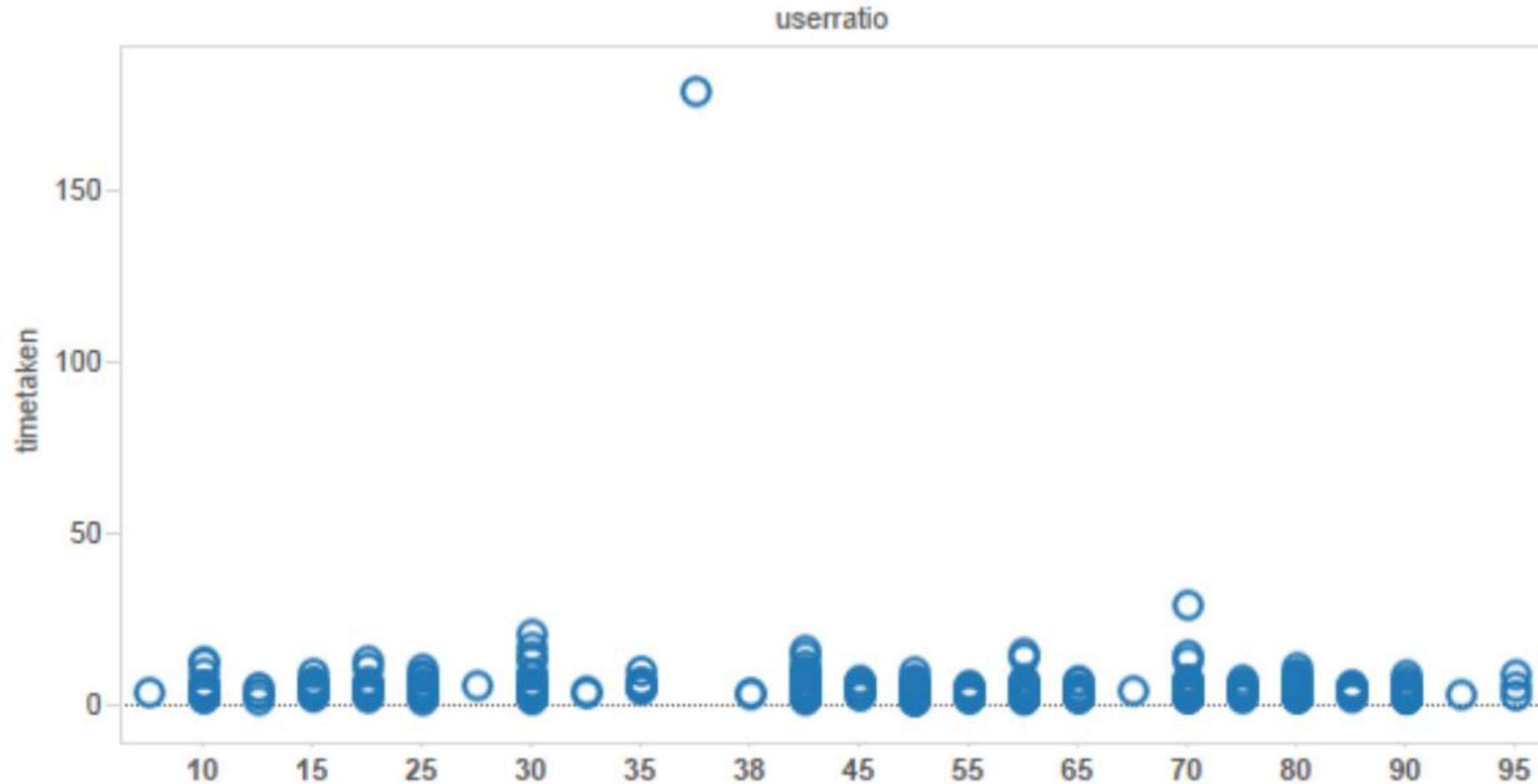


Selección de marcas en los ejes

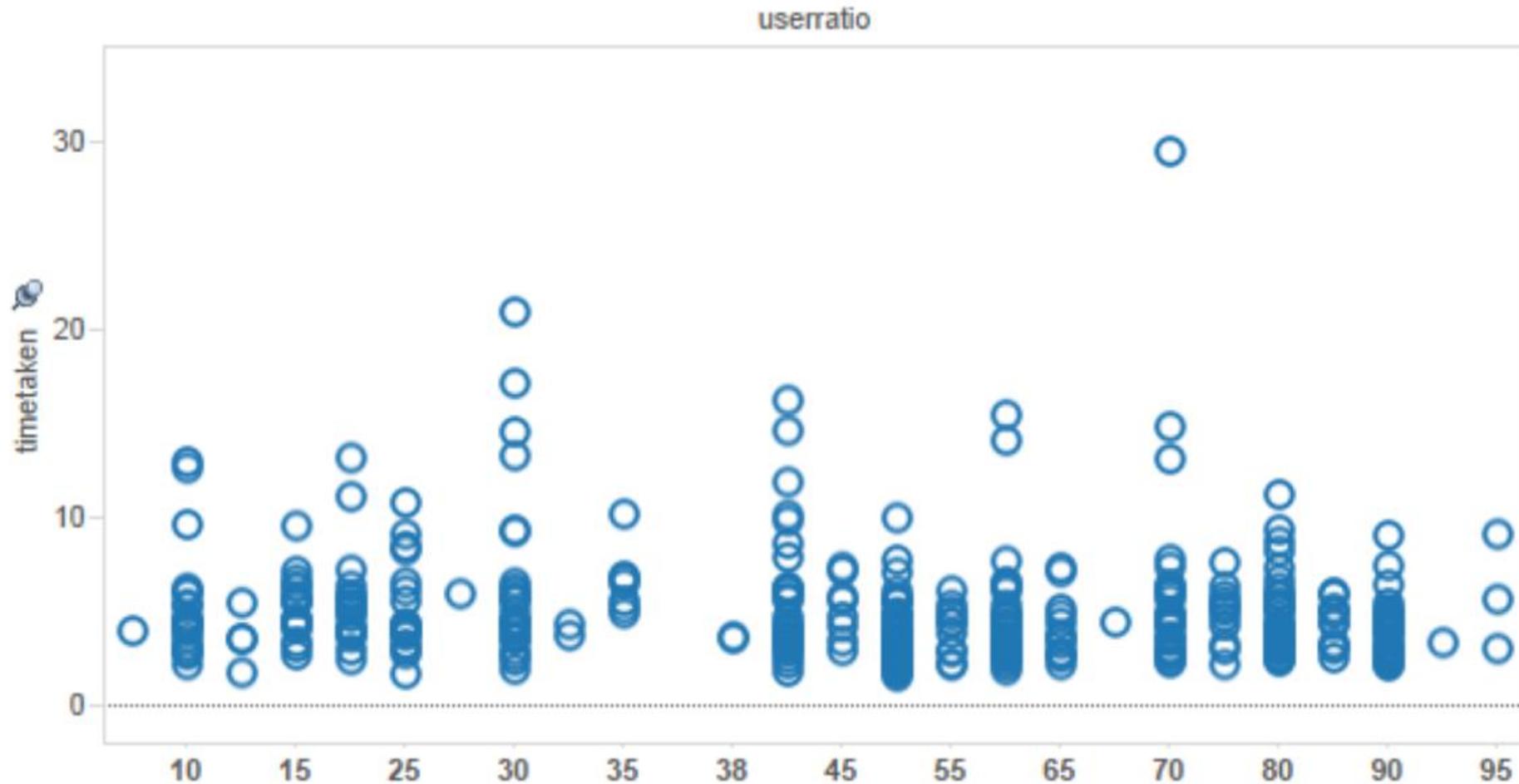


- **Simplicidad:** Números múltiplos de 10, 5 o 2.
- **Cobertura:** Marcas cerca de los extremos de los ejes.
- **Densidad:** Ni demasiadas ni muy pocas.
- **Legibilidad:** Espacios entre etiquetas, texto horizontal, tamaño adecuado.

¿Cuál es la escala correcta para los ejes?

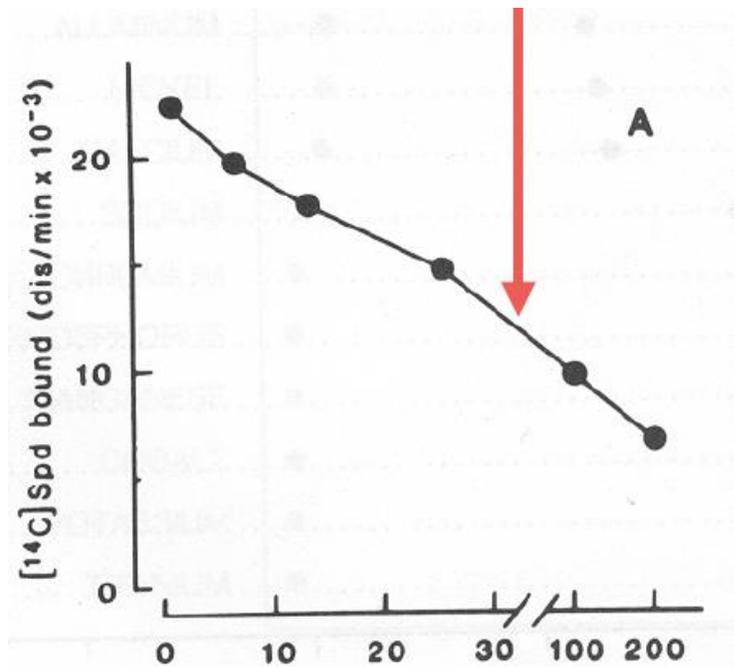


¿Cuál es la escala correcta para los ejes? Eliminar los outliers

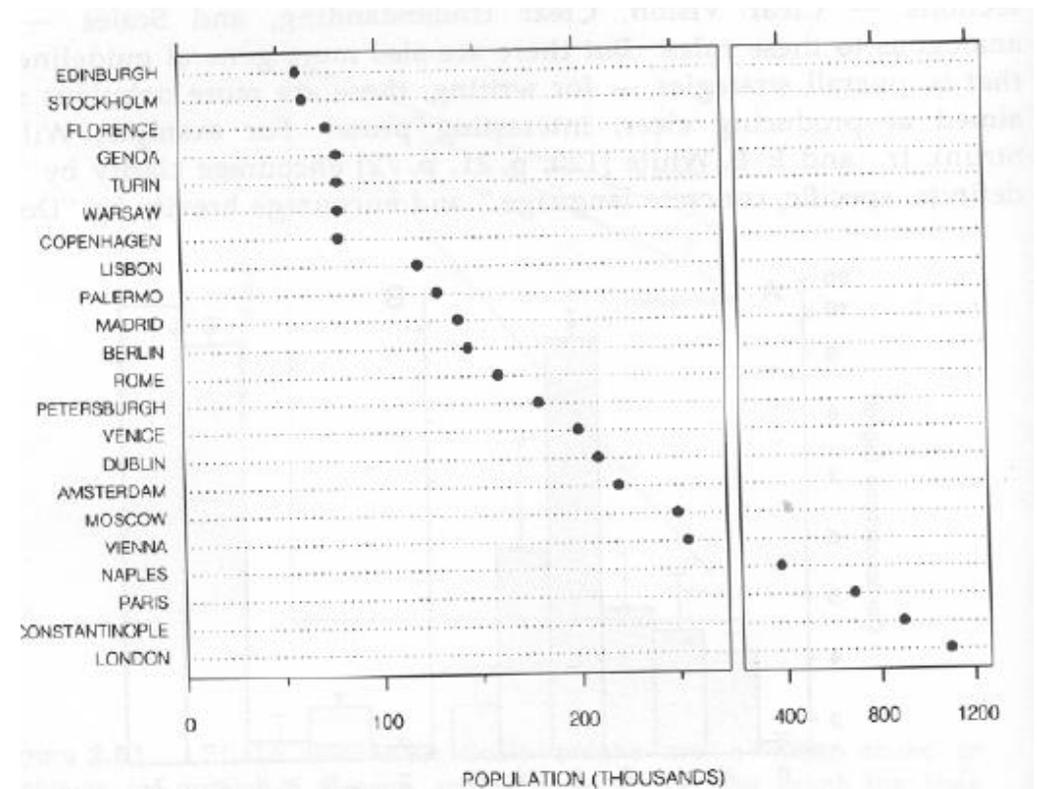


¿Cuál es la escala correcta para los ejes? Marcar claramente dónde se corta la escala

Se rompe el principio de expresividad, al unir información que no está unida, al menos de esa forma.



Mejor expresado el cambio de escala.

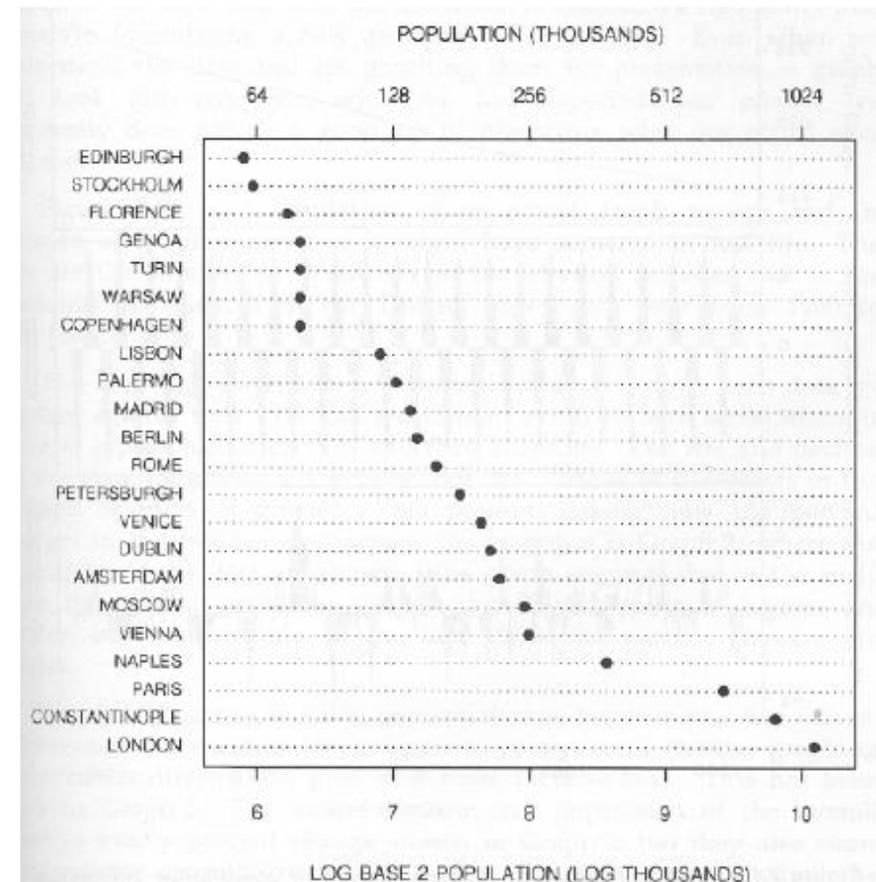
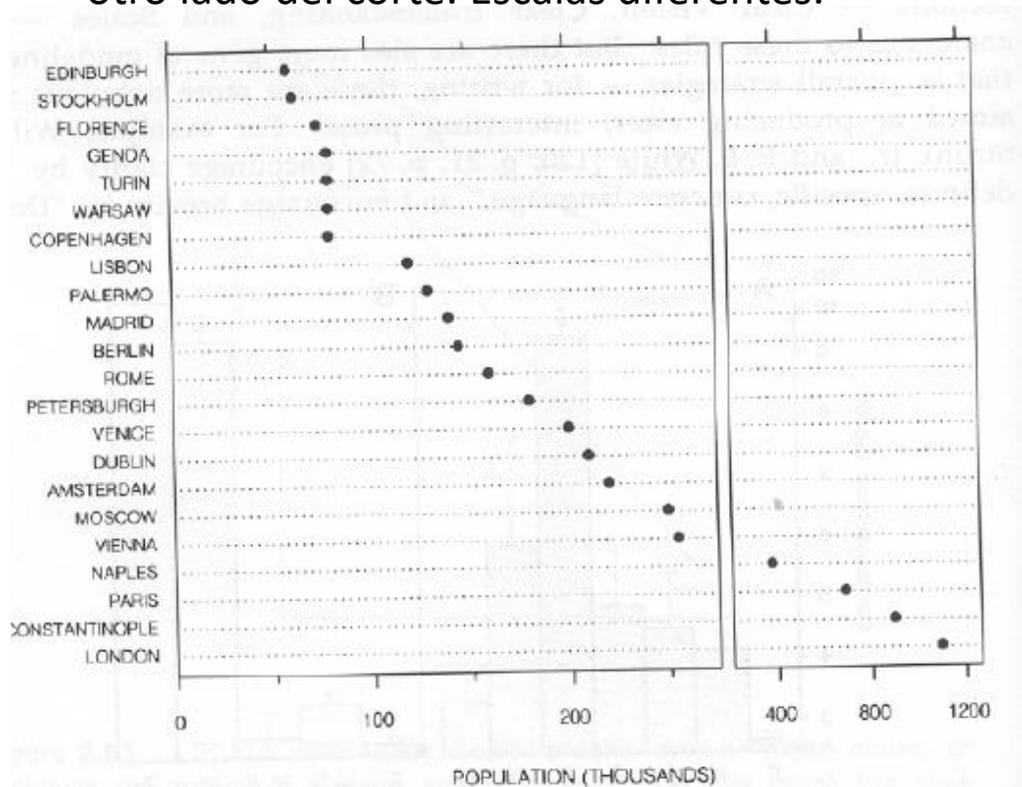


¿Cuál es la escala correcta para los ejes? ¿Escala logarítmica? ¿Cortes en la escala?

Ambas mejoran la resolución:

Difícil de comparar (ratio) los valores del otro lado del corte. Escalas diferentes.

Comparación directa, pero hay que estar habituado a la escala logarítmica



¿Cuál es la escala correcta para los ejes?

¿Escala lineal? ¿Escala logarítmica?

Escala lineal

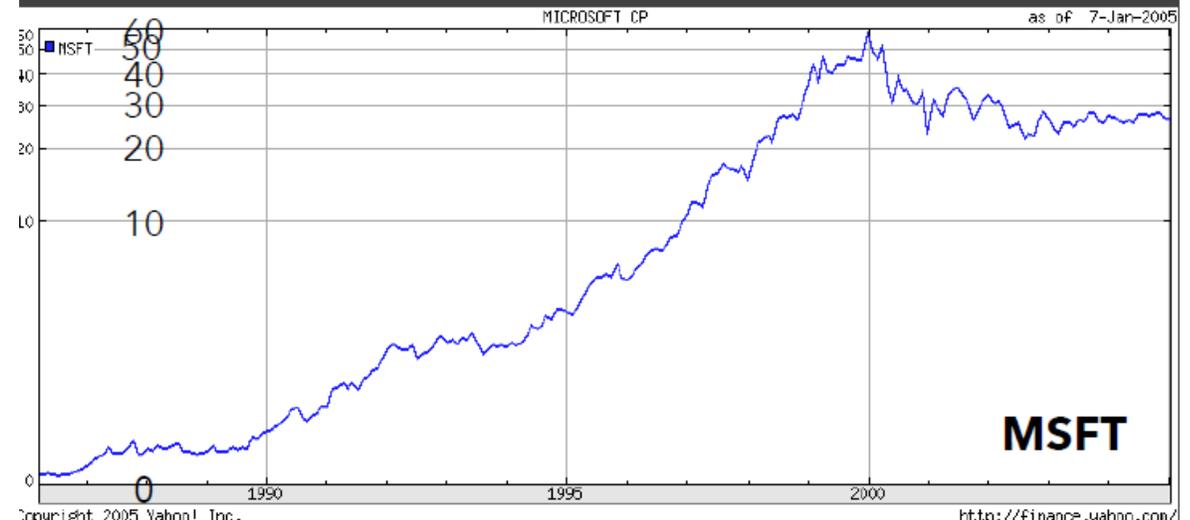
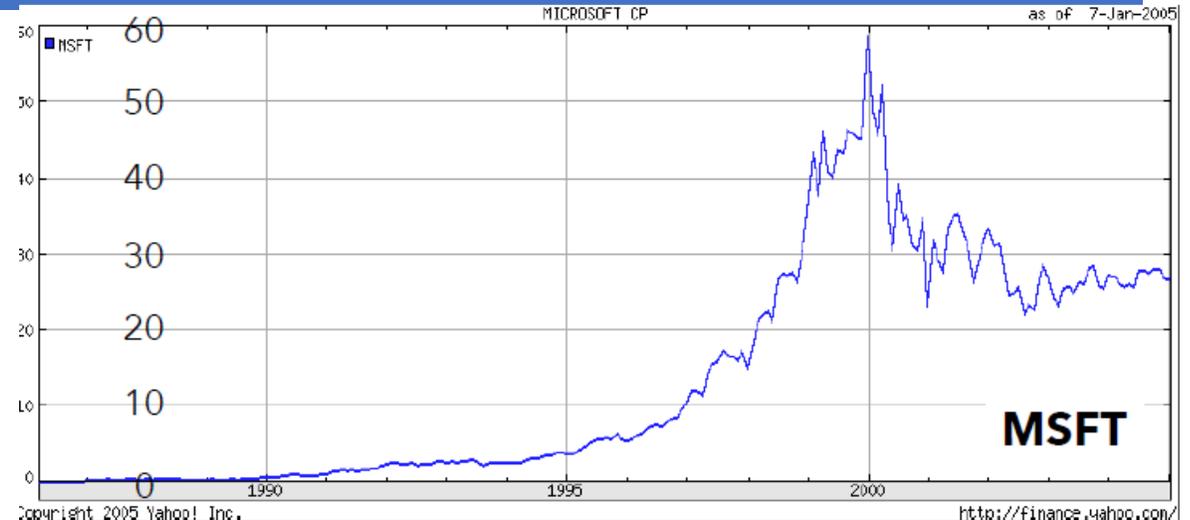
Cambios absolutos: Las variaciones entre 10 y 20 se observan de igual forma que las variaciones entre 40 y 50.

$$20-10 = 50 - 40$$

Escala logarítmica

Cambios en proporción: Las variaciones entre 10 y 20 se ven mucho mayores que las variaciones entre 40 y 50.

$$20/10 = 2 \gg 5/4 = 50/40$$



¿Cuándo aplicar la escala logarítmica?

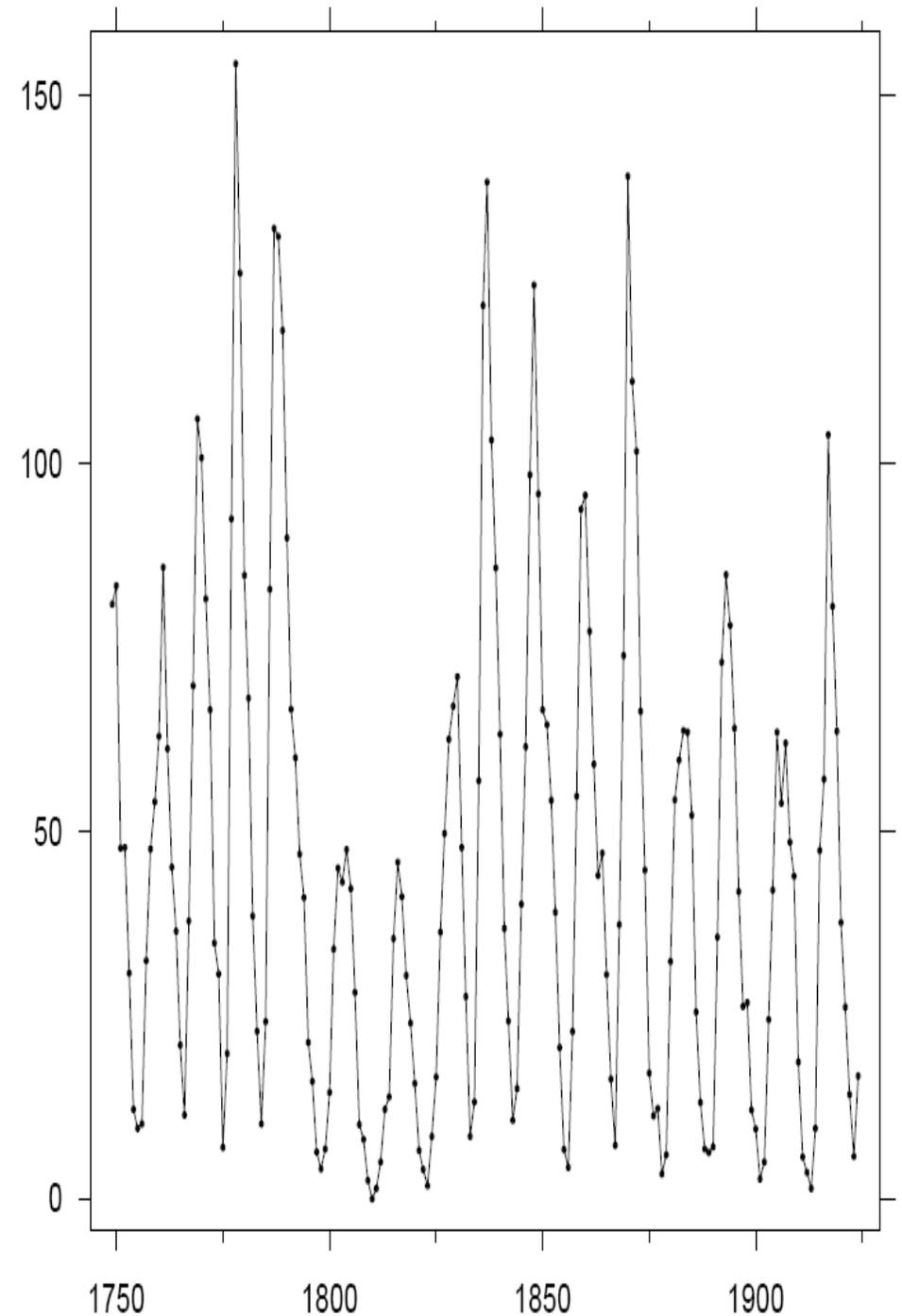
- Contemplar sesgo de datos (colas largas, valores atípicos). Permite la comparación dentro y entre varios órdenes de magnitud.
- El logaritmo transforma * en +. $\log(x * y) = \log(x) + \log(y)$
- Cambio en % o en ratio, no en la diferencia lineal.
- Restricciones:
 - Los valores deben ser positivos distintos de cero (log de negativo es complejo)
 - ¿familiaridad de la audiencia?

¿Cuándo aplicar la escala logarítmica?

- Contemplar sesgo de datos (colas largas, valores atípicos). Permite la comparación dentro y entre varios órdenes de magnitud.
- El logaritmo transforma * en +. $\log(x * y) = \log(x) + \log(y)$
- Cambio en % y en ratio, no en la diferencia lineal ni en la resta.
- Restricciones:
 - Los valores deben ser positivos distintos de cero (log de negativo es complejo)
 - Familiaridad del destinatario ¿Entenderá la escala?

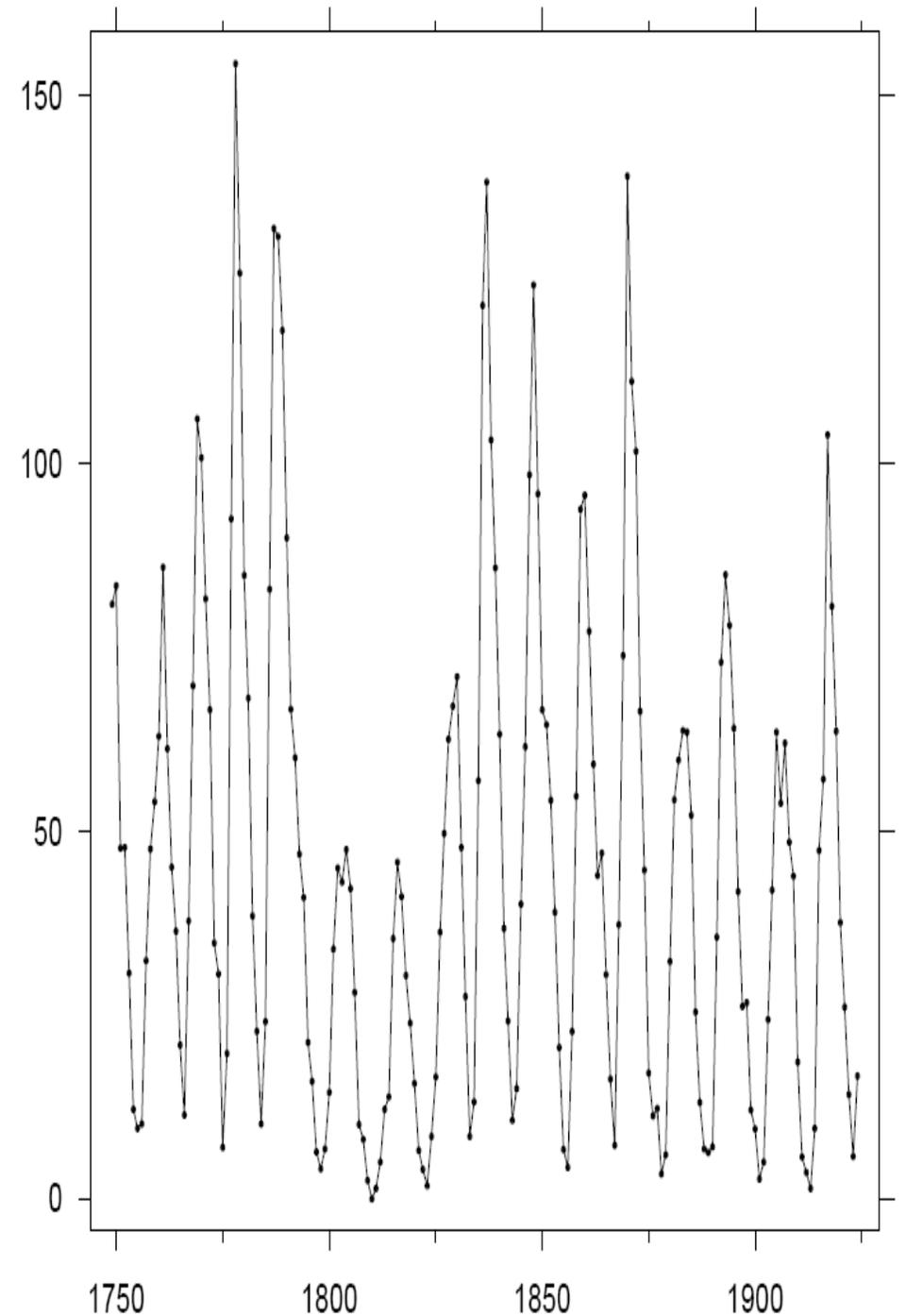
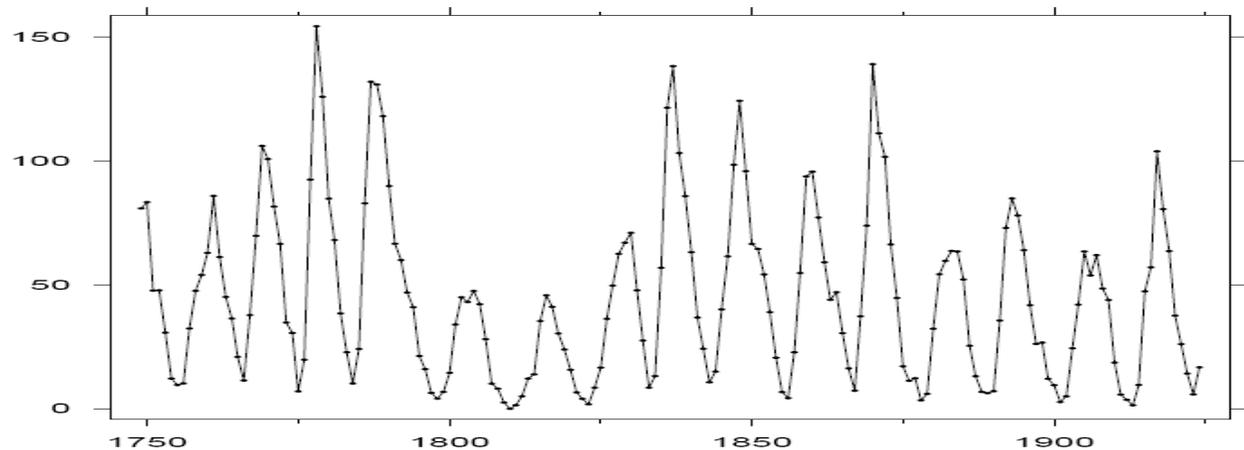
Relación ancho-alto (aspect ratio)

En esta gráfica se observa mucho mejor los valores individuales del eje Y.
Pero, ¿qué ocurre con la observación de las pendientes o de la concavidad?



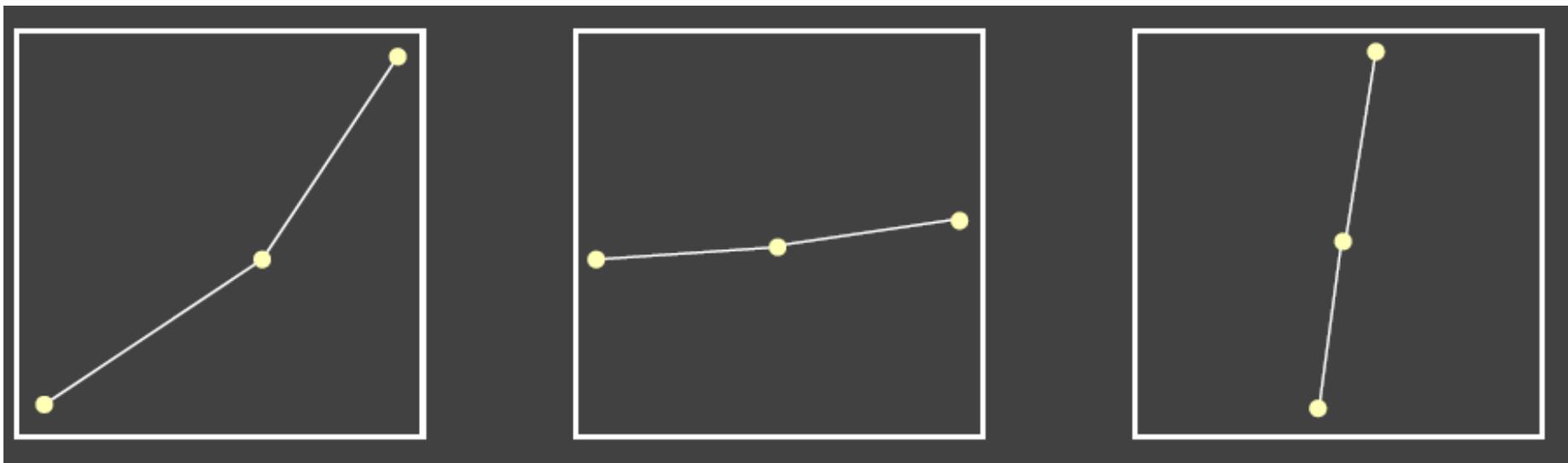
Relación ancho-alto (aspect ratio)

- En la gráfica se observa mucho mejor los valores individuales del eje Y.
- Es más difícil comparar visualmente las pendientes, o la concavidad.



Relación ancho-alto (aspect ratio)

- Para facilitar la percepción de tendencias, maximice la discriminabilidad de las orientaciones de los segmentos de línea.
- Dos segmentos de línea se pueden discriminar al máximo cuando su ángulo absoluto promedio es de 45° .
- **Método:** optimice la relación de aspecto de modo que el ángulo absoluto promedio de todos los segmentos sea de 45° .



Relación ancho-alto (aspect ratio)

- **Método:** optimice la relación de aspecto de modo que el ángulo absoluto promedio de todos los segmentos sea de 45°

[Cleveland et al. 1988]

Aquí en el panel A es donde se distingue mejor que la concavidad es positiva.

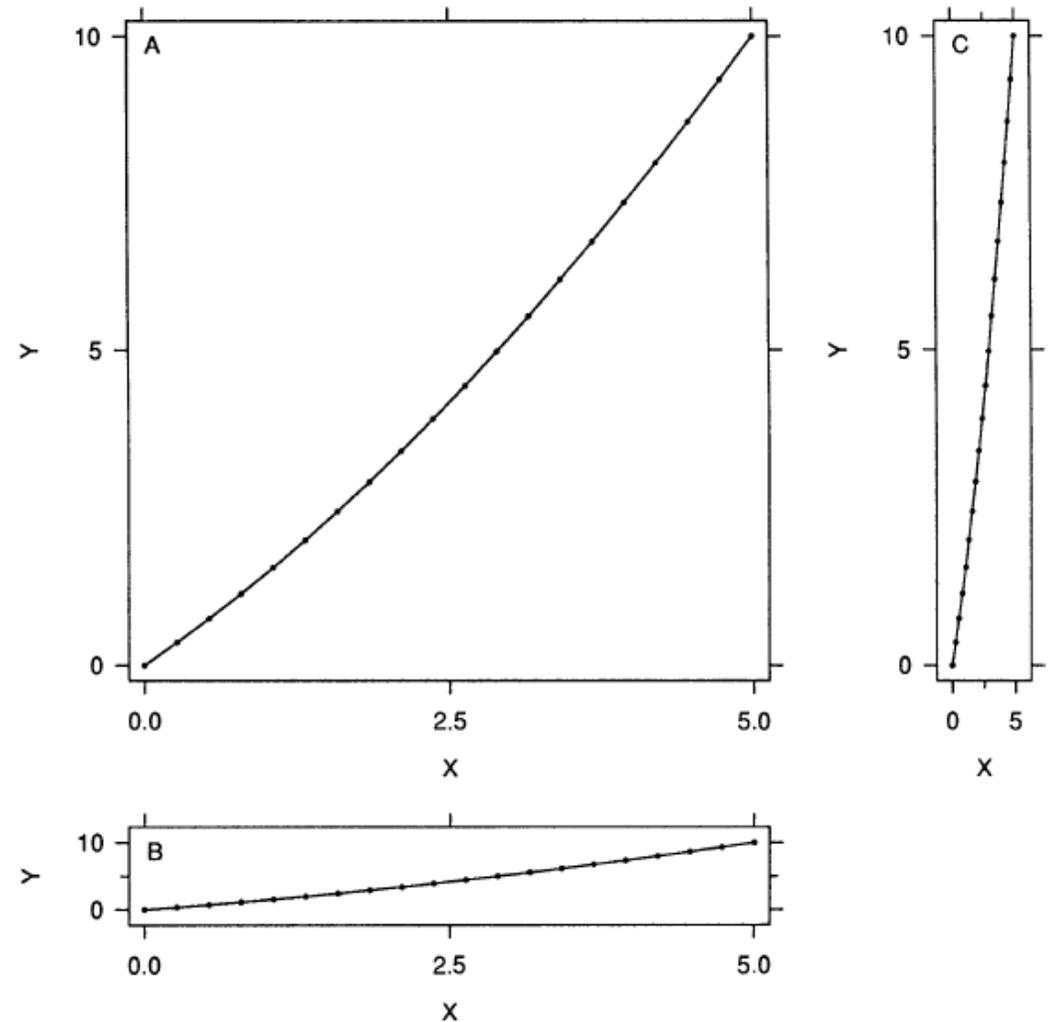


Figure 5. Orientation Resolution. The same data are graphed in all three panels. The orientation resolution is 18.4° in Panel A, 4.0° in Panel B, and 4.0° in Panel C. In Panels B and C it is difficult to perceive the curvature that is apparent in Panel A, because the resolutions are so much smaller. Orientation resolution is maximized when the midangle is 45° .

Relación ancho-alto (aspect ratio)

- **Método:** optimice la relación de aspecto de modo que el ángulo absoluto promedio de todos los segmentos sea de 45°

[Cleveland et al. 1988]

En la figura 2 se observan los valores, pero no se observa que la curva asciende más lentamente de lo que desciende.

Eso se observa mejor en la figura 3

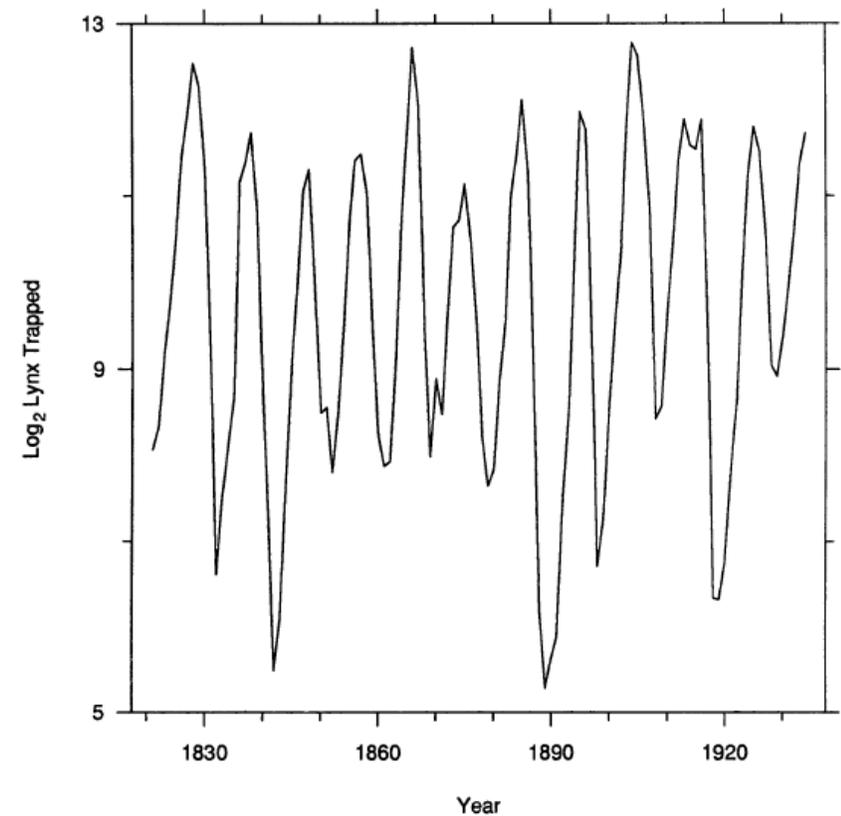


Figure 2. The Effect of Shape. The data are the Canadian lynx trapings from 1821 to 1934. The shape parameter is 1. The orientations of the line segments connecting successive data points are too close to 90° and -90° to allow us to see an important property of the data.

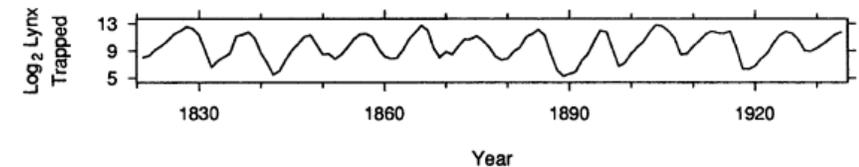
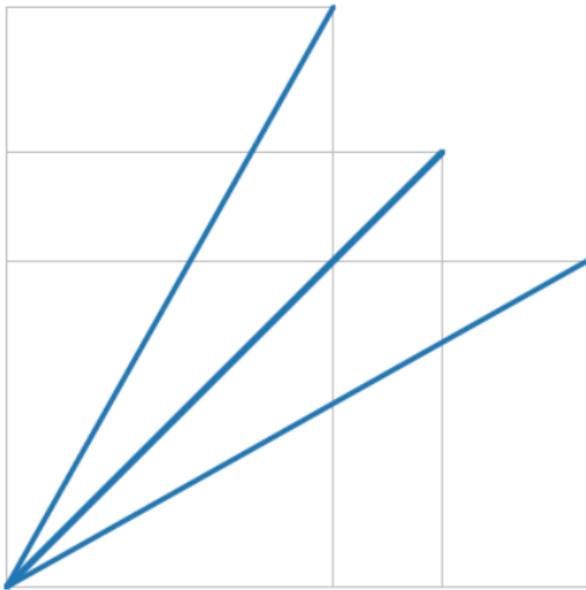


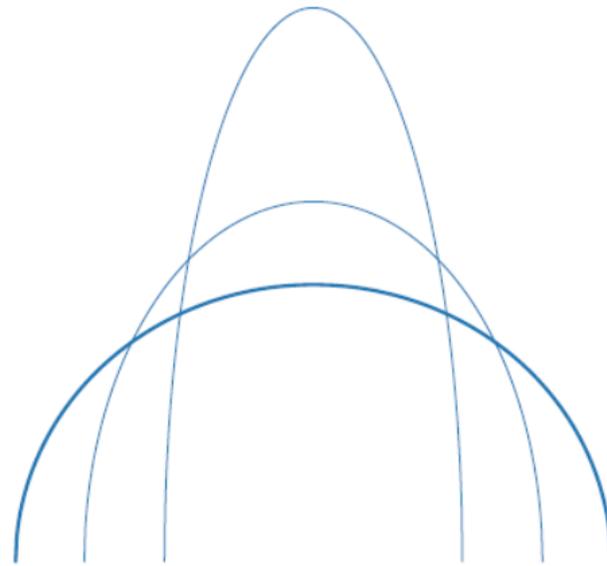
Figure 3. The Effect of Shape. The shape parameter of the graph is .074. The orientations of the line segments are in a range that allows better visual decoding of the slopes. We can now see what we could not see in Figure 2—the numbers tend to rise more slowly than they fall.

Alternativa [Talbot et al. 2011]:

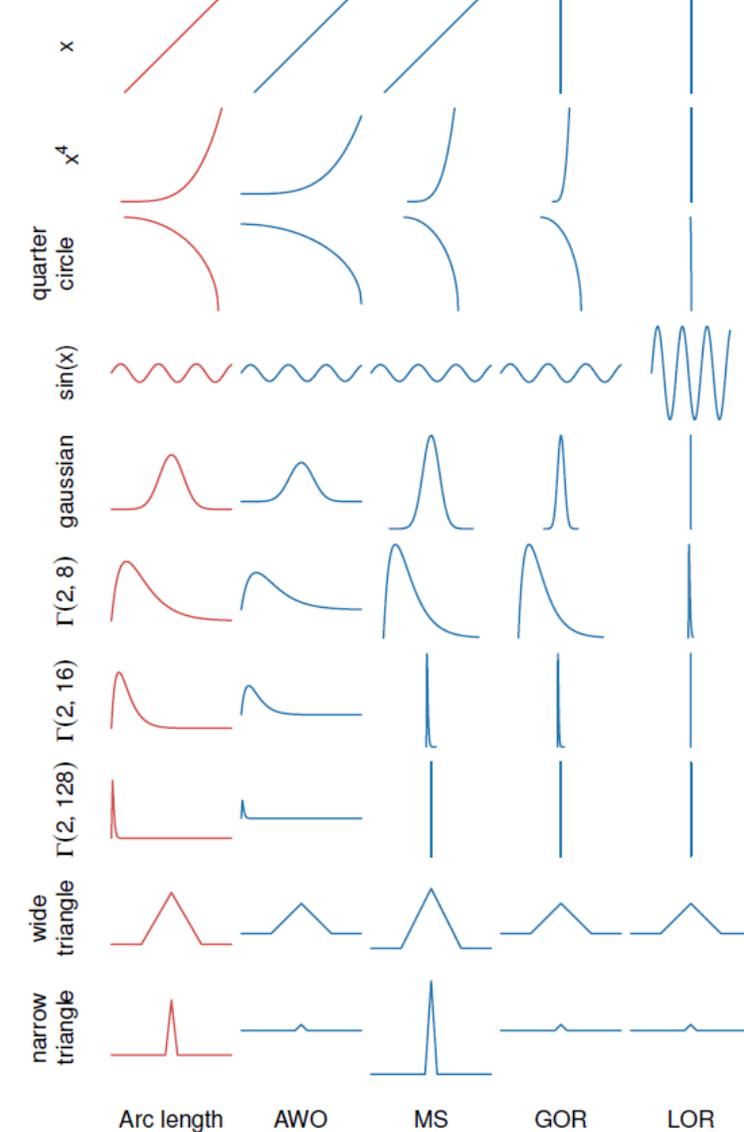
- Minimizar la longitud de arco, mientras se mantiene constante el área.



Straight line $\rightarrow 45^\circ$



Ellipse \rightarrow Circle



MS: Median Slope (45°)

AWO: Length Weighted Average Orientation

GOR: Global Orientation Resolution

LOR: Local Orientation Resolution

Variables visuales

- Posición (X)
- Posición (Y)
- Área
- Valor
- Textura
- Color
- Orientación
- Forma

~8 Variables?

LES VARIABLES DE L'IMAGE

	POINTS			LIGNES			ZONES	
XY 2 DIMENSIONS DU PLAN								
Z TAILLE								
VALEUR								

LES VARIABLES DE SÉPARATION DES IMAGES

GRAIN								
COULEUR								
ORIENTATION								
FORME								

Ejemplo: Ventas de Café

Cifras de venta de una cadena de café ficticia.

Ventas: Cuantitativo – Ratio

Ganancias: Cuantitativo – Ratio

Marketing: Cuantitativo – Ratio

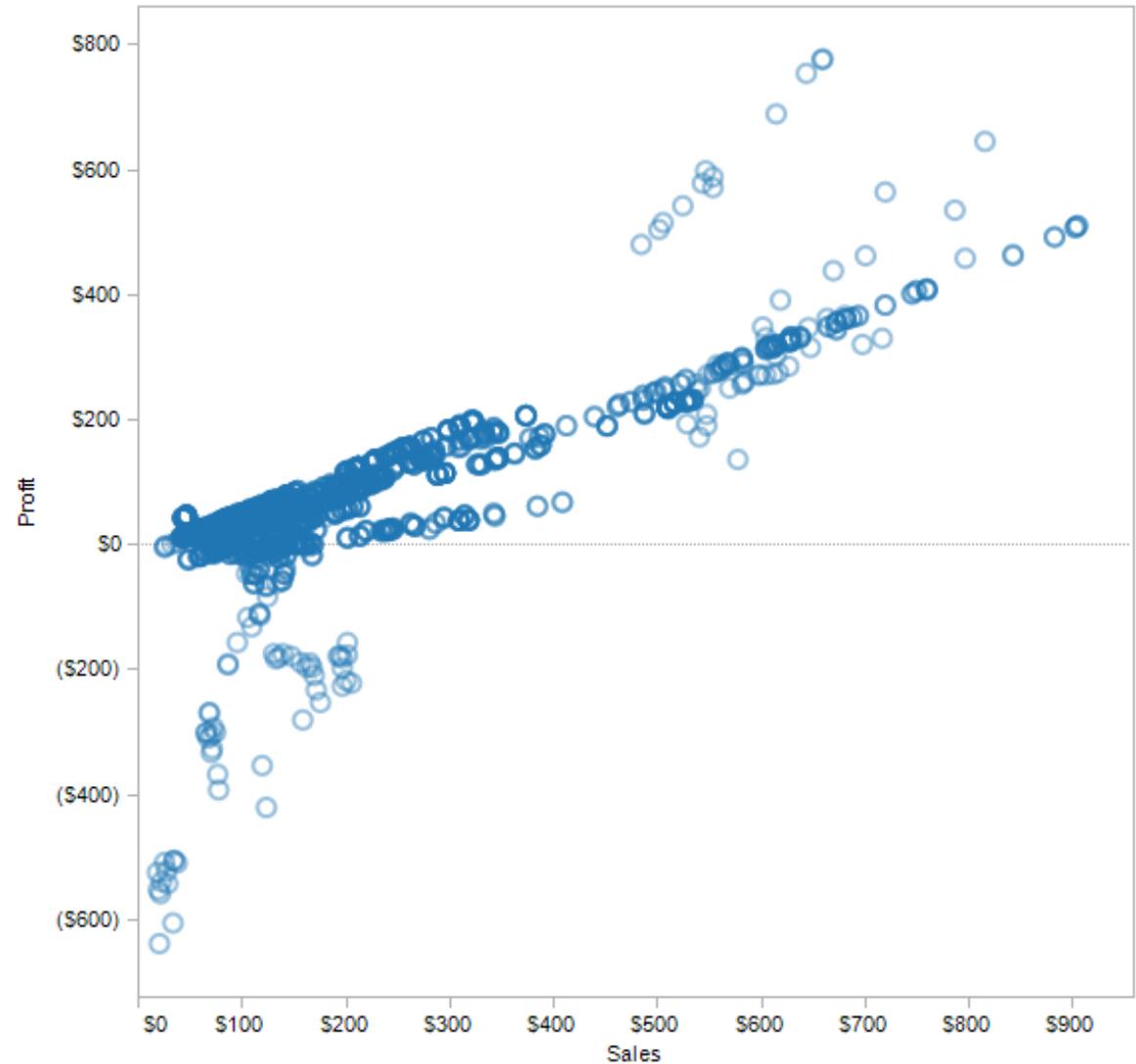
Tipo de producto: Nominal {Café, Espresso, Te de hierbas, Te}

Mercado: Nominal {Central, Este, Sur, Oeste}

Ejemplo: Ventas de Café

Cifras de venta de una cadena de café ficticia.

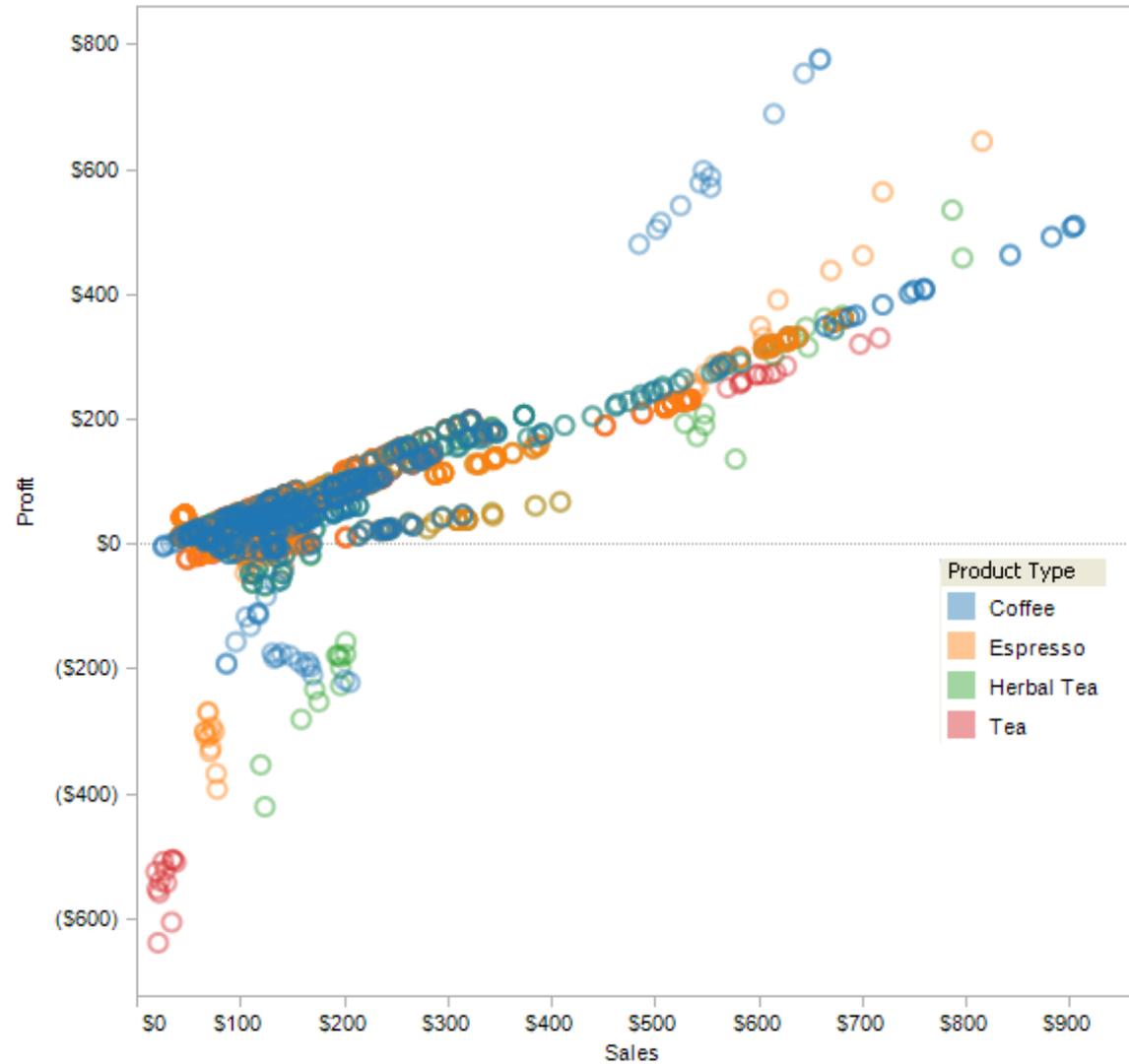
- Ventas y ganancia se codifican con posición.



Ejemplo: Ventas de Café

Cifras de venta de una cadena de café ficticia.

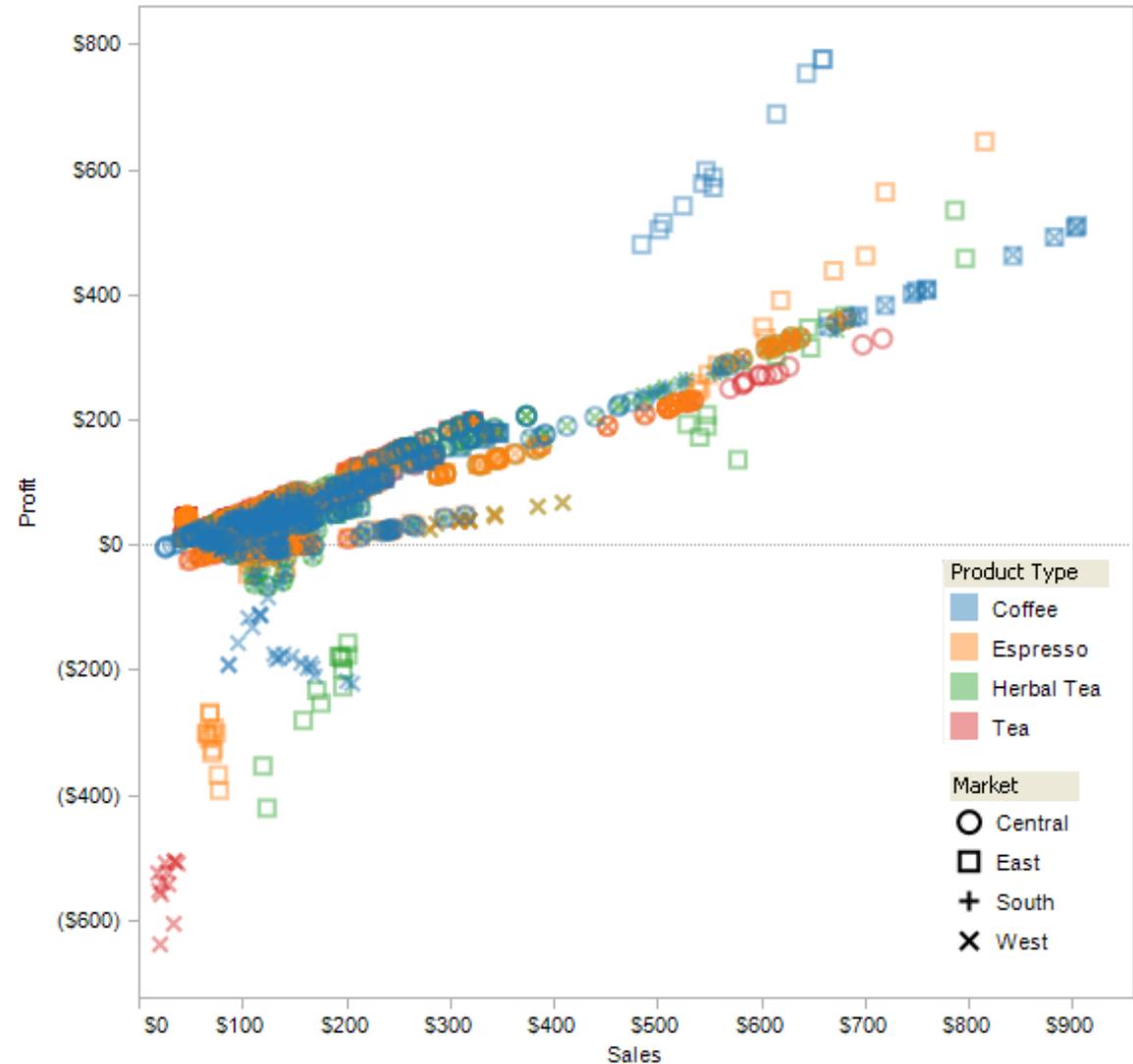
- Ventas y ganancia se codifican con posición.
- El tipo de producto se codifica con color.



Ejemplo: Ventas de Café

Cifras de venta de una cadena de café ficticia.

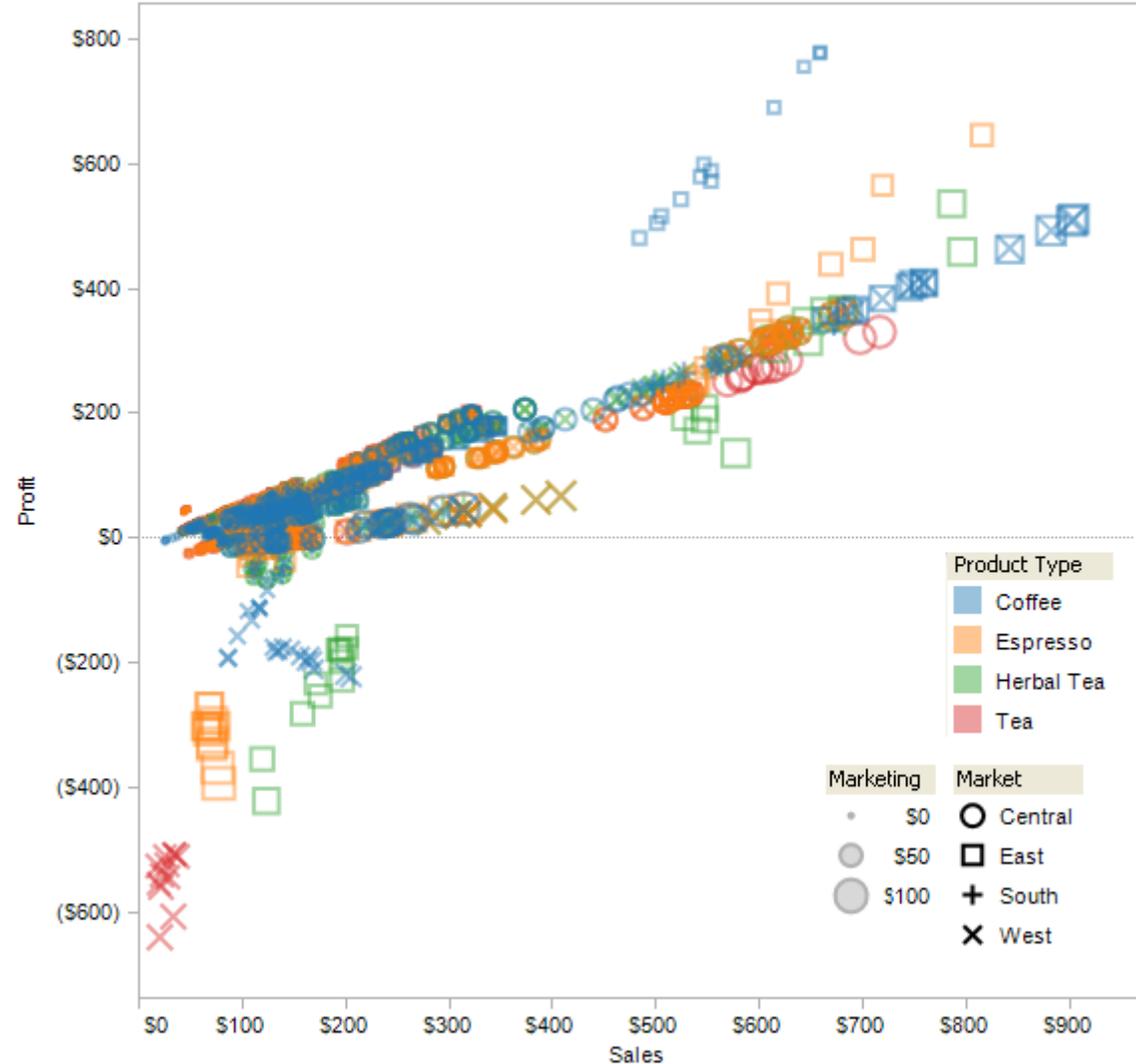
- Ventas y ganancia se codifican con posición.
- El tipo de producto se codifica con color.
- El mercado se codifica con la forma.



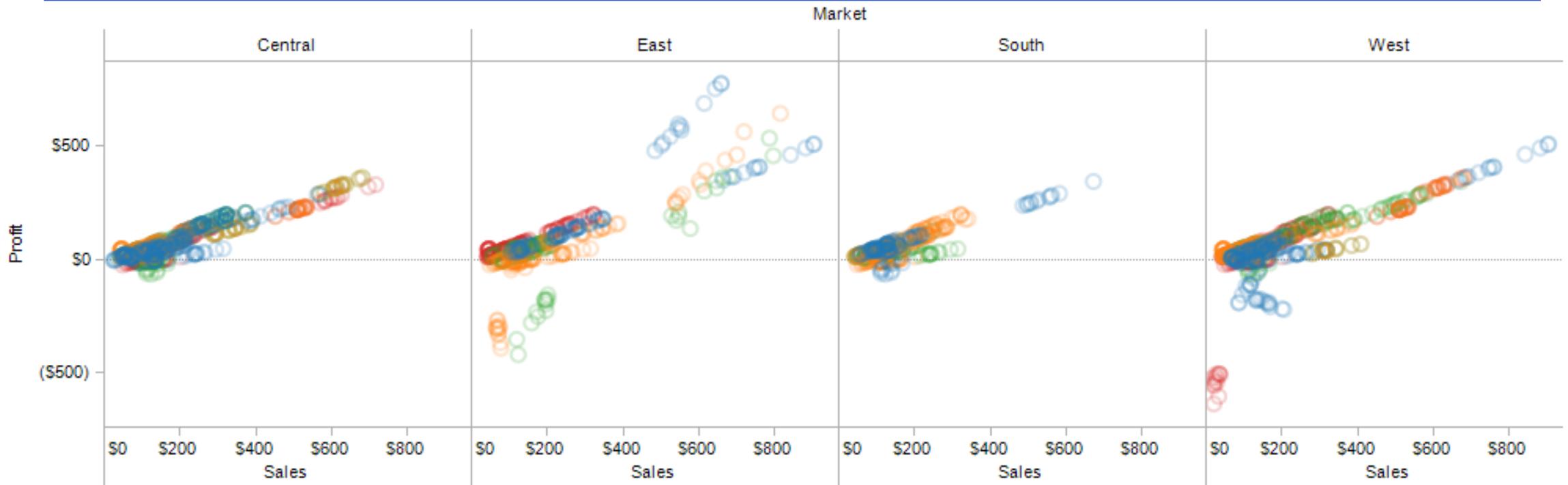
Ejemplo: Ventas de Café

Cifras de venta de una cadena de café ficticia.

- Ventas y ganancia se codifican con posición.
- El tipo de producto se codifica con color.
- El mercado se codifica con la forma.
- El Marketing se codifica con el área



Ejemplo: Ventas de Café

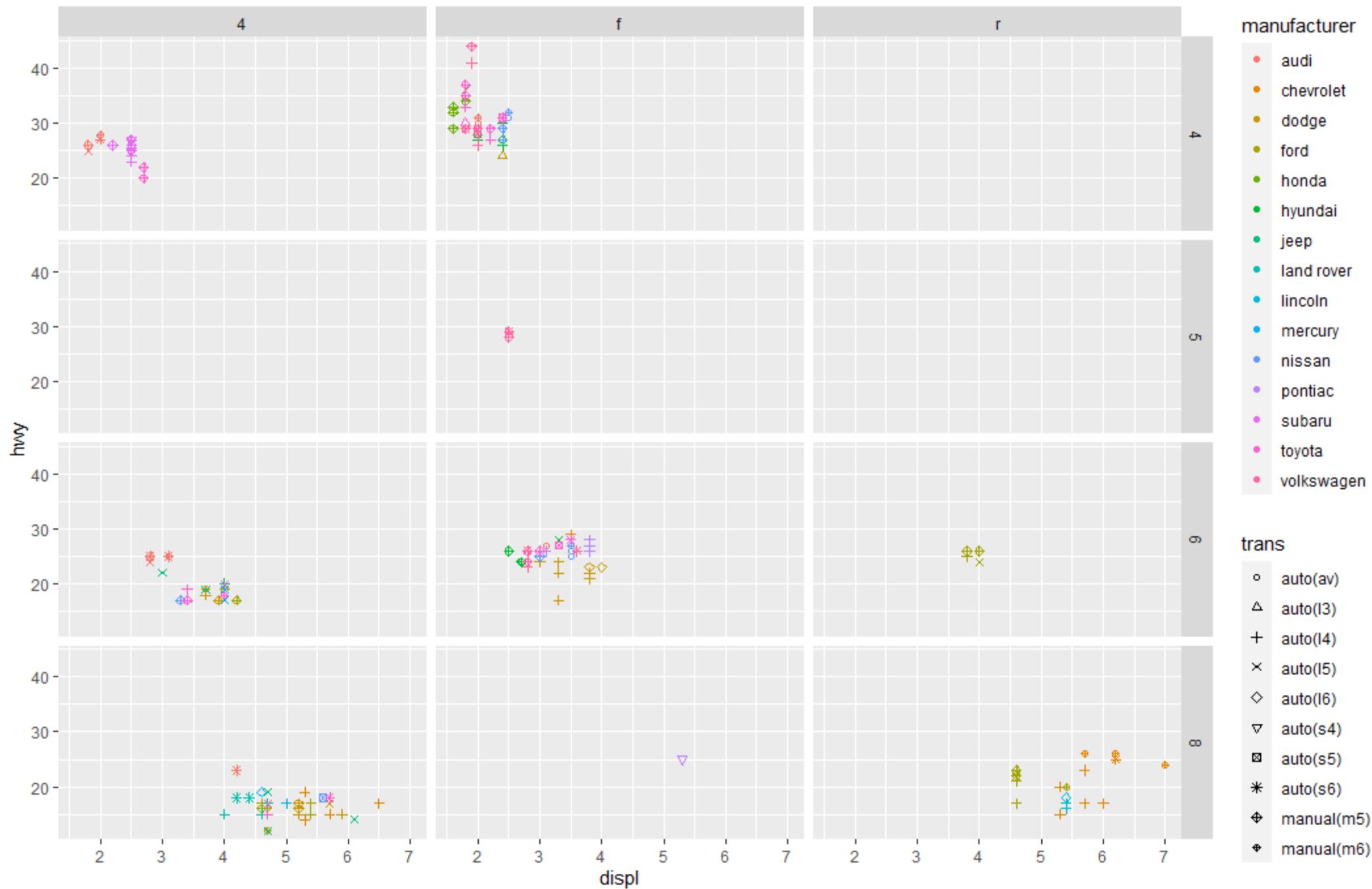


Se puede utilizar Facets (R) o gráfico Trellis o ... divide al gráfico en subgráficos basado en los valores de uno o más variables discretas. En este caso se utiliza Mercado.

Ejemplo en R: data.frame mpg

```
manufacturer model displ year cyl trans drv cty hwy fl class
<chr> <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
1 audi a4 1.8 1999 4 auto(l5) f 18 29 p compact
2 audi a4 1.8 1999 4 manual(m5) f 21 29 p compact
:
```

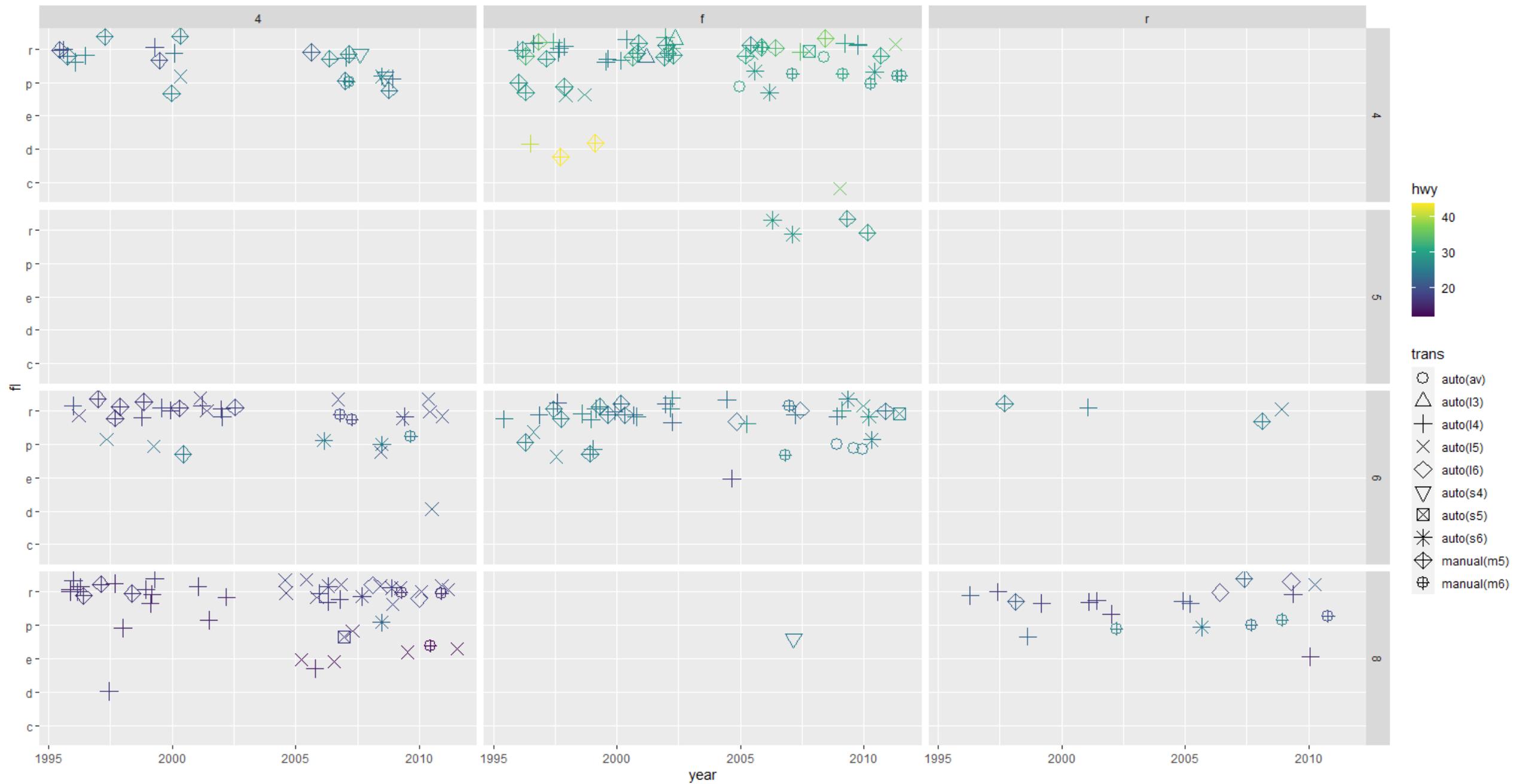
```
ggplot(data = mpg, aes(x = displ, y = hwy)) +
  geom_point(aes(color = manufacturer, shape=trans)) +
  scale_shape_manual(values=1:10) +
  facet_grid(cyl~ drv);
```



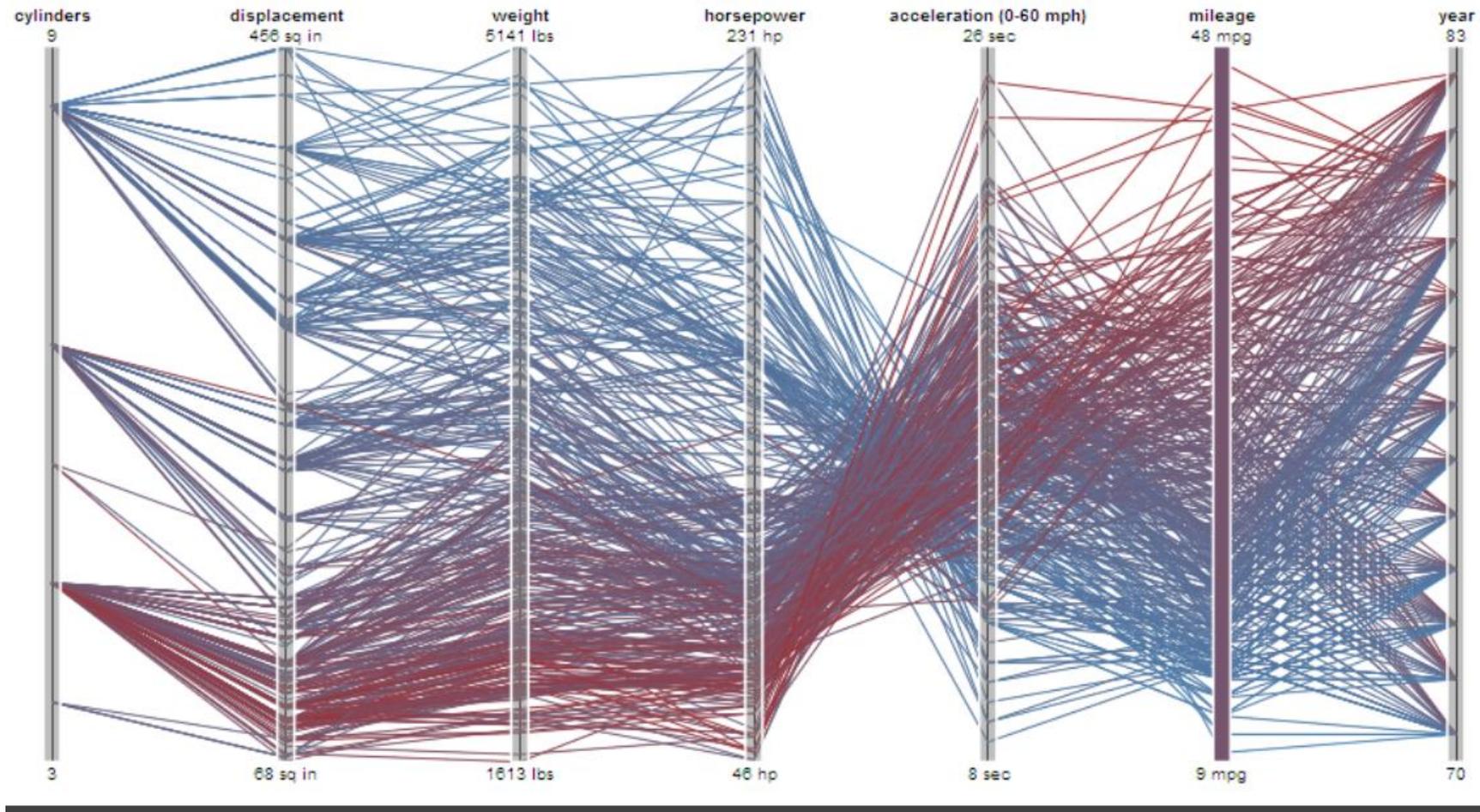
Ejemplo en R: data.frame mpg

```
manufacturer model displ  year   cyl trans      drv   cty   hwy fl   class
  <chr>          <chr> <dbl> <int> <int> <chr>    <chr> <int> <int> <chr> <chr>
1 audi          a4      1.8  1999     4 auto(l5)  f     18    29 p    compact
2 audi          a4      1.8  1999     4 manual(m5) f     21    29 p    compact
:
```

```
ggplot(mpg, aes(y=fl, x=year, color = hwy, shape=trans)) +
  scale_shape_manual(values=1:10) +
  scale_color_continuous(type = "viridis") +
  geom_jitter(size=4) +
  facet_grid(cyl~ drv);
```



Coordenadas Paralelas



<https://plotly.com/r/parallel-coordinates-plot/>

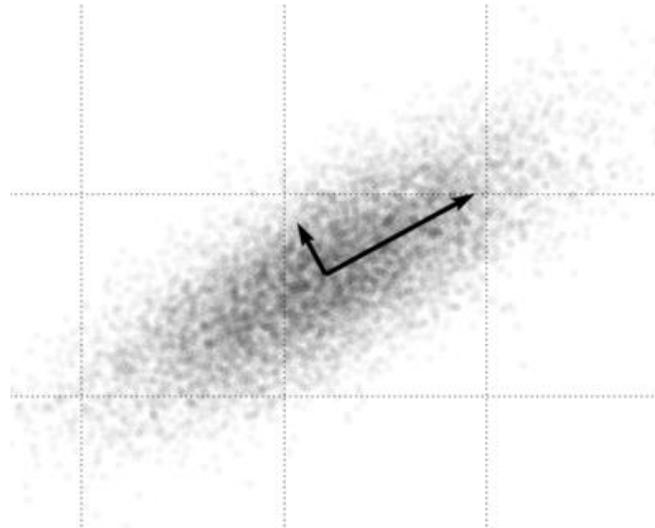
<https://plotly.com/ggplot2/parallel-coordinates-plot/>

Coordenadas Paralelas

- Permite visualizar hasta ~ dos docenas de dimensiones a la vez
- 1. Dibuja ejes paralelos para cada variable.
- 2. Para cada tupla, conecta puntos en cada eje.
- **Entre ejes adyacentes:** los cruces de líneas implican una correlación negativa, las pendientes compartidas implican una correlación positiva.
- La trama completa puede estar desordenada.
- La selección interactiva se puede utilizar para evaluar relaciones multivariadas.
- Muy sensible a la escala y al ordenamiento de los ejes.
- **Se requiere experiencia para usar de manera efectiva**

Reducción de la dimensionalidad (RD)

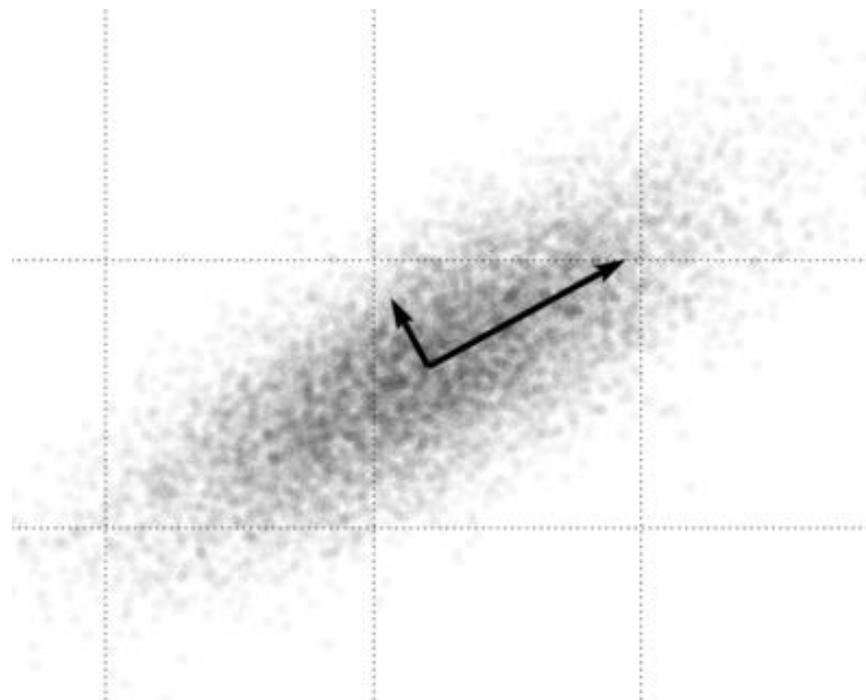
- **Se proyectan datos n dimensionales en 2D o 3D para su visualización.**
- Los diferentes métodos de RD hacen diferentes balances:
 - Se puede preservar la estructura global (p. Ej., PCA)



- Se puede enfatizar la estructura local (p. Ej., Enfoques de vecino más cercano, incluidos t-SNE y UMAP) <https://distill.pub/2016/misread-tsne/#citation>

Análisis de Componentes Principales

1. Desplazar los datos para que la media quede en el origen.
2. Encuentre ejes ortogonales que maximicen la varianza de los datos.
 - Cálculo de valores y vectores propios.
3. Proyecte los datos sobre los dos o tres vectores con mayor varianza.
 - Utilice los 2 o 3 vectores propios con mayores valores propios.
4. Grafique usando esos valores como nuevas dimensiones



problema: las nuevas dimensiones suelen no tener un significado intuitivo

Análisis de Componentes Principales

- En PCA se realizan transformaciones lineales:
- Se rota, se escala y se proyectan los datos originales.
- Las líneas y vectores se proyectan como líneas rectas y vectores rectos.
- Se preservan las distancias globales.

