

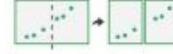
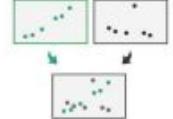


Visualización de Información y Analítica Visual

Hernán Valdivieso López (hfvaldivieso@uc.cl)

Resumen clase 3

How

How?			
Encode	Manipulate	Facet	Reduce
<ul style="list-style-type: none"> ⊕ Arrange <ul style="list-style-type: none"> → Express  → Separate  → Order  → Align  → Use  	<ul style="list-style-type: none"> ⊕ Change  ⊕ Select  ⊕ Navigate  	<ul style="list-style-type: none"> ⊕ Juxtapose  ⊕ Partition  ⊕ Superimpose  	<ul style="list-style-type: none"> ⊕ Filter  ⊕ Aggregate  ⊕ Embed 
<ul style="list-style-type: none"> ⊕ Map from categorical and ordered attributes → Color <ul style="list-style-type: none"> → Hue  → Saturation  → Luminance  → Size, Angle, Curvature, ...  → Shape  → Motion Direction, Rate, Frequency, ...  			

How aplicado a altair

- *encode* → para indicar canales (X, Y, color)
- *sort_values* → para ordenar
- *add_params* → para permitir selección de uno o más datos
- *interactive* → para *panning* y zoom
- | (columnas) y & (filas) → para yuxtaposición de visualización
- + → para superponer visualizaciones
- *transform_filter* → para filtrar según una selección
- *binding_select* → para hacer un selector (*dropdown*)
- *average, count, min, max* → para agregación de datos
- *tooltip* → para embeber información adicional

How?

¿Cómo elegimos entre las opciones que tenemos disponibles?

Debemos considerar al menos:

- Percepción
- Memoria
- Tarea a resolver
- Canales disponibles
- Marcas disponibles
- Interacciones entre marcas y canales
- Eficiencia de canales
- Algunas reglas basadas en la experiencia
- El usuario objetivo

How?

¿Cómo elegimos entre las opciones que tenemos disponibles?

Debemos considerar al menos:

- Percepción
- Memoria
- Tarea a resolver
- Canales disponibles
- Marcas disponibles
- Interacciones entre marcas y canales
- Eficiencia de canales
- **Algunas reglas basadas en la experiencia**
- El usuario objetivo

From data to Viz

Numeric

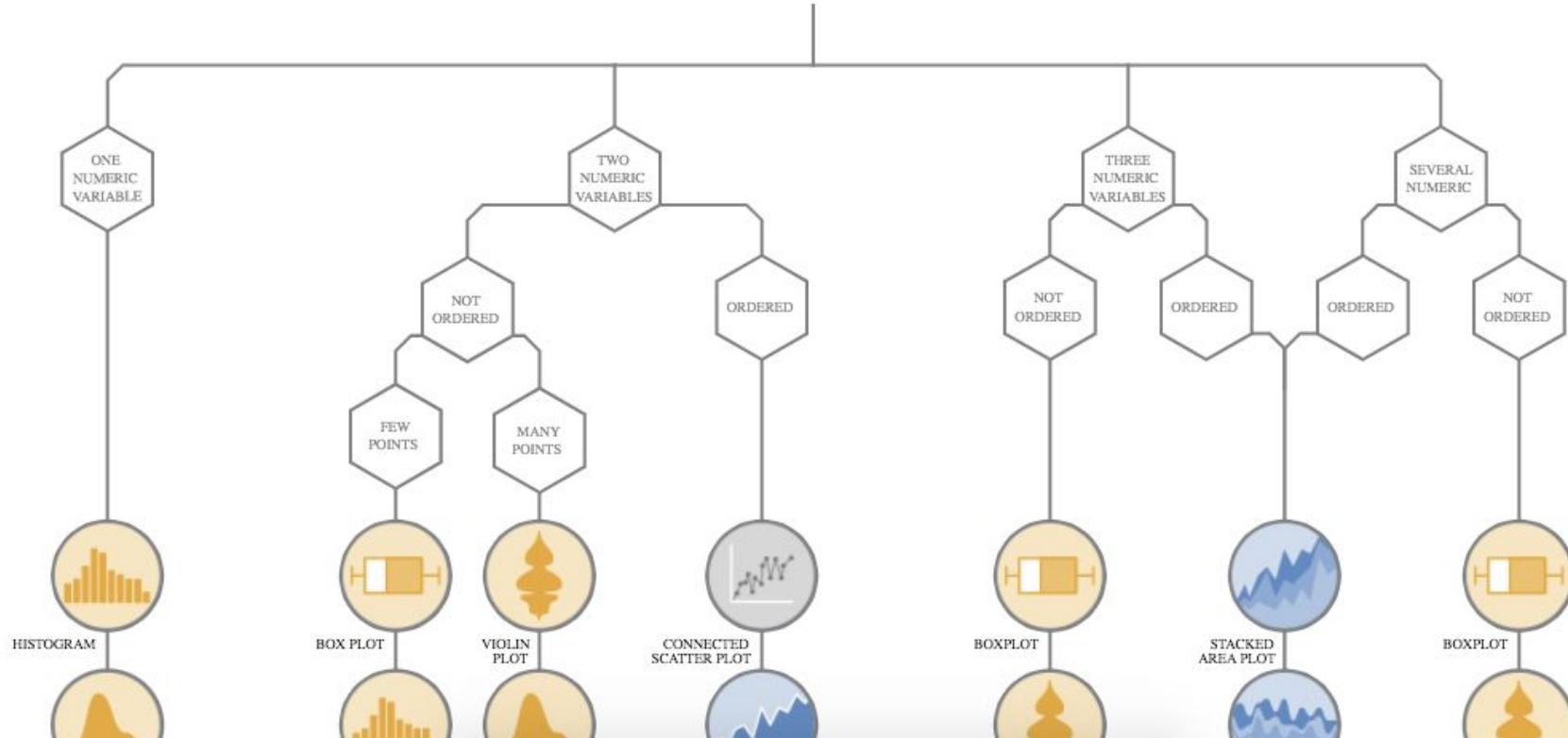
Categoric

Num & Cat

Maps

Network

Time series

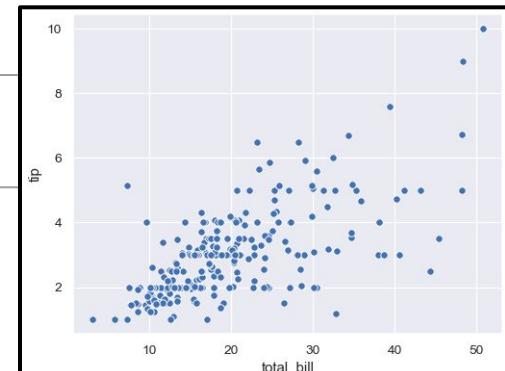


¿Cómo Tamara Munzner utiliza su *framework* para el análisis?

Ejemplo de cómo ocupar Framework de Munzner

Gráfico de dispersión o *scatterplot*

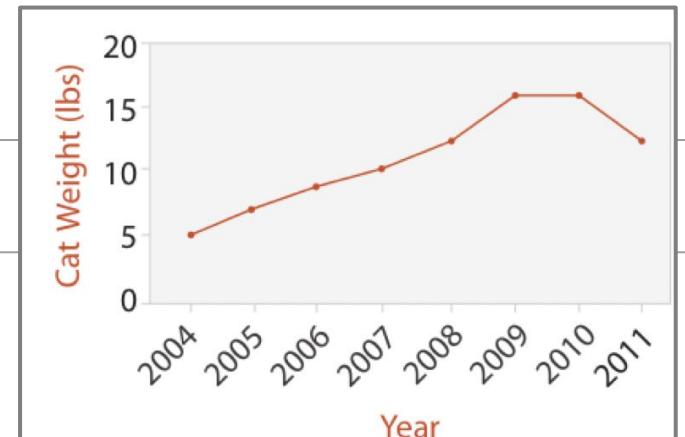
¿Qué?	Dos atributos cuantitativos.
¿Cómo?	Expresar valores con posición espacial horizontal y vertical mediante marcas de punto.
¿Por qué?	Encontrar tendencias, <i>outliers</i> , distribuciones, correlación; localizar aglomeraciones.
Escala	Cientos de ítems.



Ejemplo de cómo ocupar Framework de Munzner

Gráfico de línea o *line chart*

¿Qué?	Un atributo cuantitativo y un atributo ordenado.
¿Cómo?	Gráfico de punto con marcas de conexión entre puntos.
¿Por qué?	Presentar tendencia.
Escala	Cientos de valores en el eje X.



Clase 4: Principios de diseño & visualizaciones con Matplotlib/Seaborn

Contenidos

- ¿Cómo diseñar una visualización? - Principios de diseño
 - Relacionados con el uso correcto de canales.
 - Relacionados con HCI.
 - Relacionados con el diseño de la visualización.
- Visualizaciones tabulares con Matplotlib/Seaborn

¿Cómo diseñar una visualización?

Principios de diseño

¿Cómo diseñar una visualización?

- El espacio de posibles soluciones es **enorme**. Ejemplo: 1 dataset. 100 visualizations.
- Muchas decisiones de diseños tienen a dificultar el mensaje... son **inefectivos**.
- Existen **pocas verdades** en esta disciplina y **validar** un diseño de visualización es un proceso **sumamente difícil**.
- No hay un claro método para optimizar, pero si existen **guidelines** para apoyarte.

¿Cómo diseñar una visualización?

- El espacio de posibles soluciones es **enorme**. Ejemplo: 1 dataset. 100 visualizations.
- Muchas decisiones de diseños tienen a dificultar el mensaje... son **inefectivos**.
- Existen **pocas verdades** en esta disciplina y **validar** un diseño de visualización es un proceso **sumamente difícil**.
- No hay un claro método para optimizar, pero si existen **guidelines** para apoyarte.

Malos gráficos hasta el día de hoy...



Malos gráficos hasta el día de hoy...



Malos gráficos hasta el día de hoy...

EXPLORATION OF INTERPRETABILITY TECHNIQUES FOR DEEP COVID-19 CLASSIFICATION USING CHEST X-RAY IMAGES

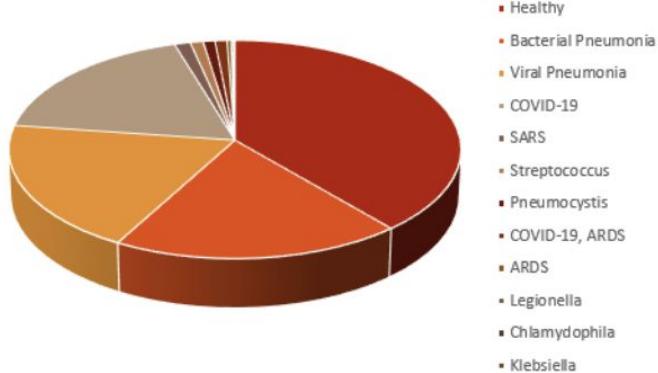
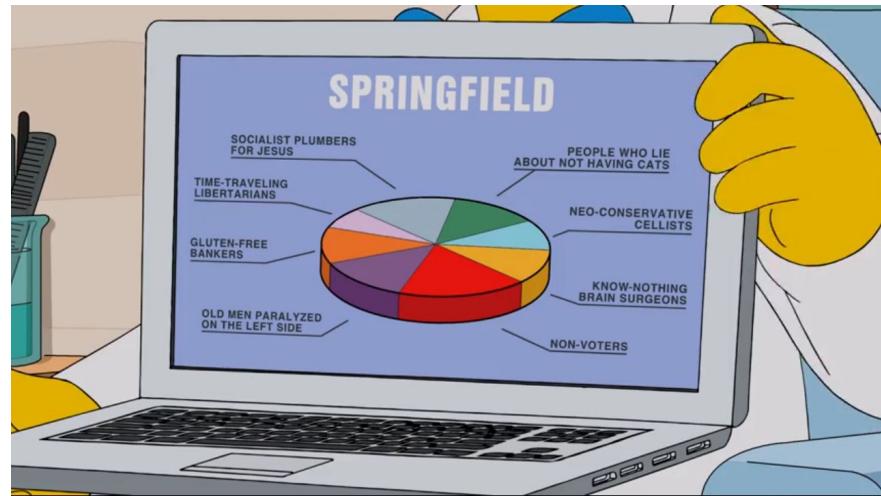


Fig. 1. CXR images distribution for each infection type in the dataset



Principios de diseño ...

... relacionados con el uso correcto de canales

- Factor de la mentira (*Lie factor*)
- Ejes engañosos
- No al 3D injustificado
- Lograrlo en blanco y negro (*Get it right in black and white*)

... relacionados con HCI

- *Overview first, details on demand*
- Los ojos le ganan a la memoria (*Eyes beat memory*)
- Tiene que ser receptivo (*Responsive is required*)

... relacionados con el diseño de la visualización

- Tasa de tinta de datos (*Data ink ratio*)
- Consistencia interna y externa
- Autocontención

... relacionados con el uso correcto de canales

... relacionados con el uso correcto de canales

- Principios de diseño relacionados con una correcta elección de canales o la forma que se están mostrando en la visualización.
- Vamos a analizar 4 principios:
 - Factor de la mentira (*Lie factor*) propuesto por Edward Tufte
 - Ejes engañosos
 - No al 3D injustificado
 - Lograrlo en blanco y negro (*Get it right in black and white*).

... relacionados con el uso correcto de canales

Factor de la mentira (*lie factor*)

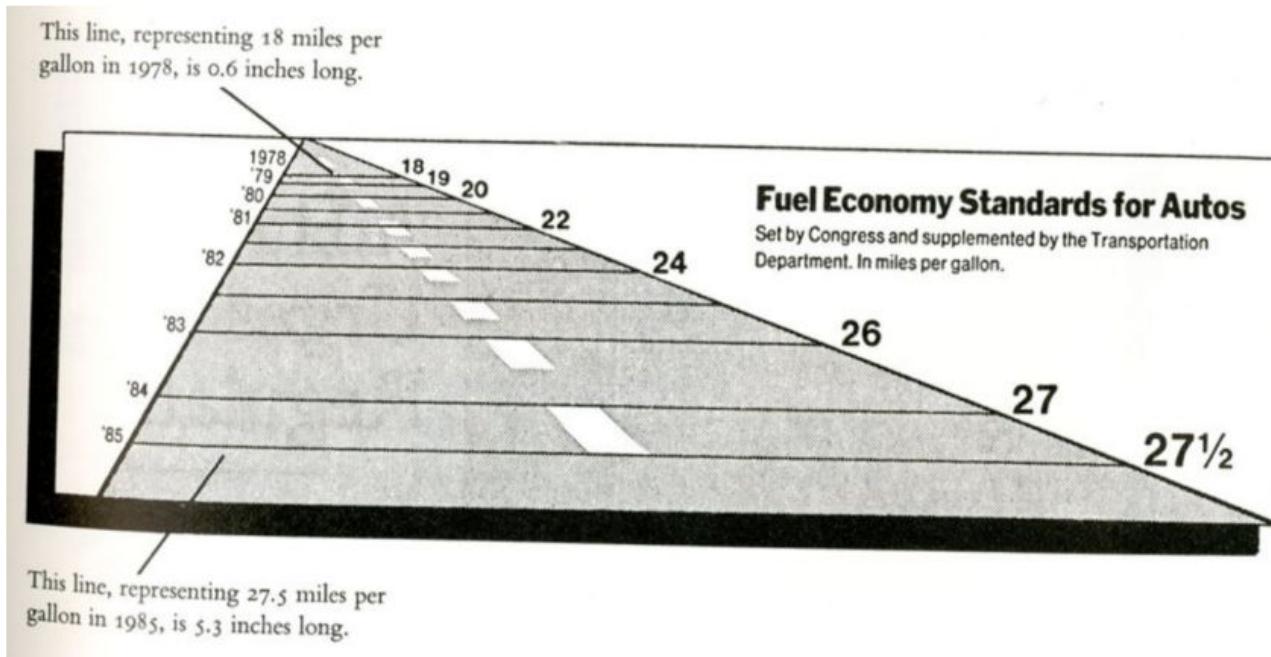
$$\text{Lie factor} = \frac{\text{size effect in graphic}}{\text{size effect in data}}$$

- El espíritu de este principio es que **visualmente entendamos lo mismo que declaran los datos.**
- La tasa de cambio entre los datos debe ser fielmente reflejada por el efecto que se muestra gráficamente.
- En nuestras visualizaciones, buscamos llegar que esta proporción sea 1.
- Cuando no ocurre esto, no esperar que el usuario realice comparaciones precisas de los datos.

... relacionados con el uso correcto de canales

Factor de la mentira (*lie factor*)

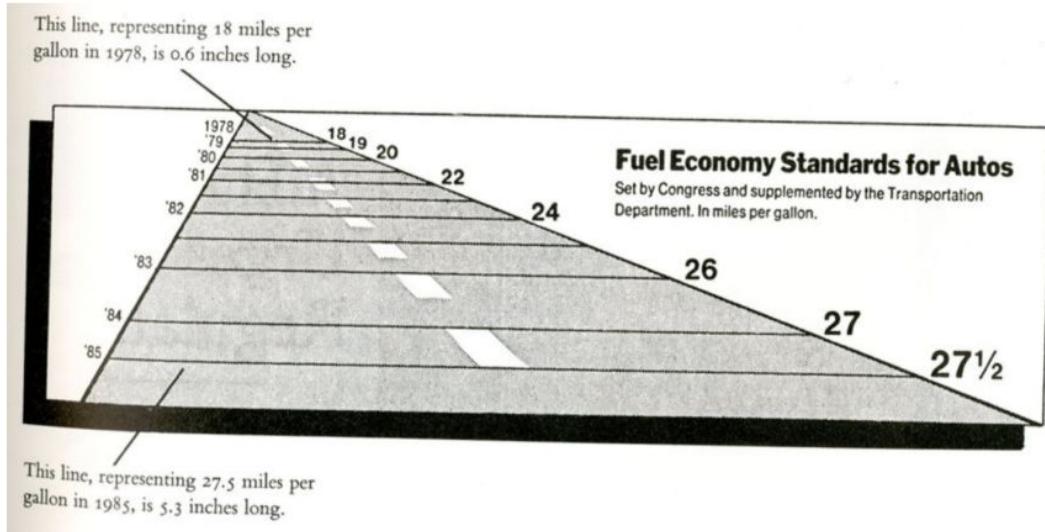
Cuántas millas por galón se logran transitar a medida que pasan los años



... relacionados con el uso correcto de canales

Factor de la mentira (*lie factor*)

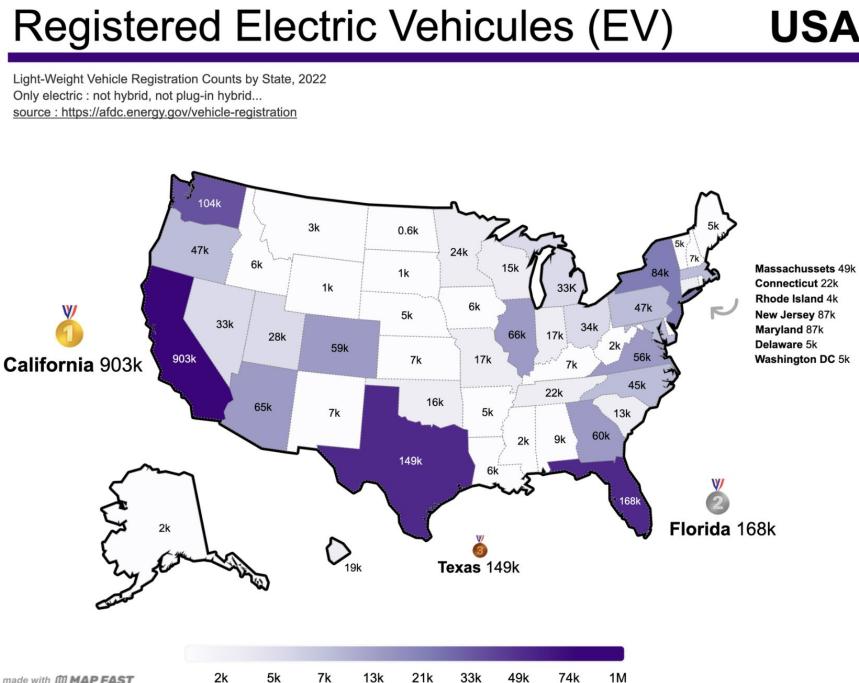
Cuántas millas por galón se logran transitar a medida que pasan los años



$$\text{Lie Factor} = \frac{\frac{5.3 - 0.6}{0.6}}{\frac{27.5 - 18}{18}} = 14.8$$

... relacionados con el uso correcto de canales

Factor de la mentira (*lie factor*)



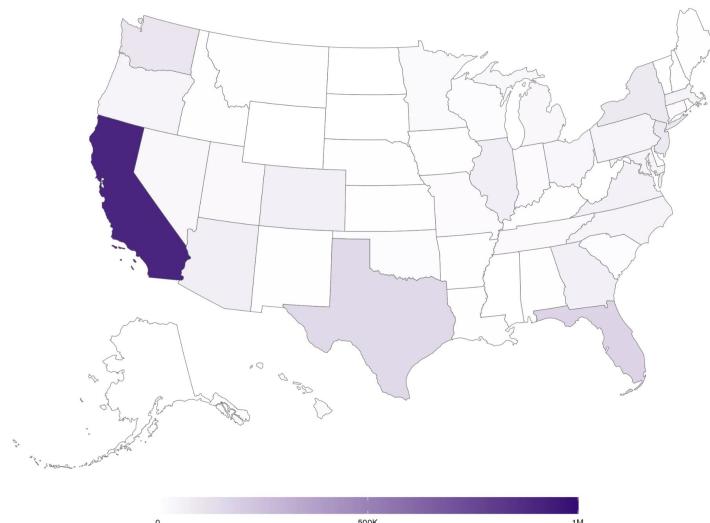
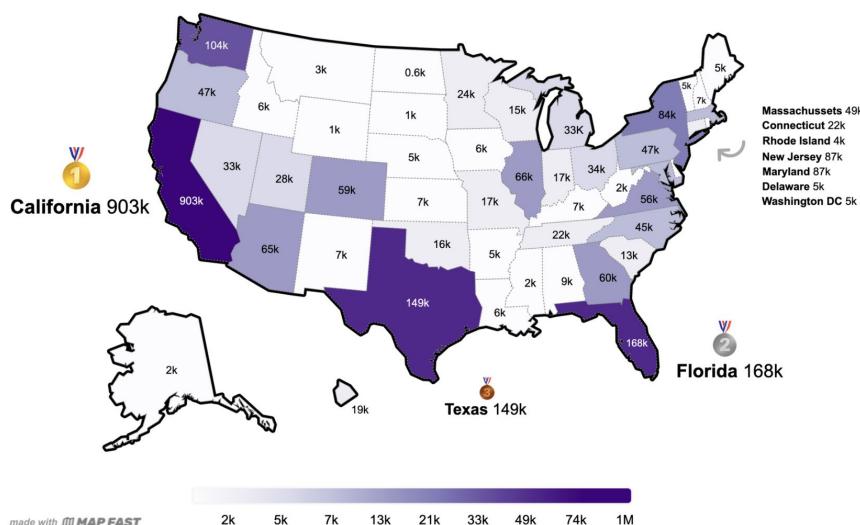
Fuente: The scale is horrendous : r/dataisugly

... relacionados con el uso correcto de canales

Factor de la mentira (*lie factor*)

Registered Electric Vehicles (EV) USA

Light-Weight Vehicle Registration Counts by State, 2022
Only electric : not hybrid, not plug-in hybrid...
source : <https://afdc.energy.gov/vehicle-registration>



... relacionados con el uso correcto de canales

Factor de la mentira (*lie factor*)

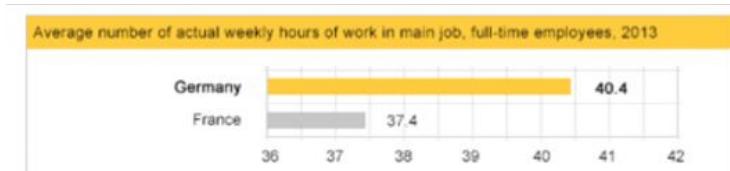
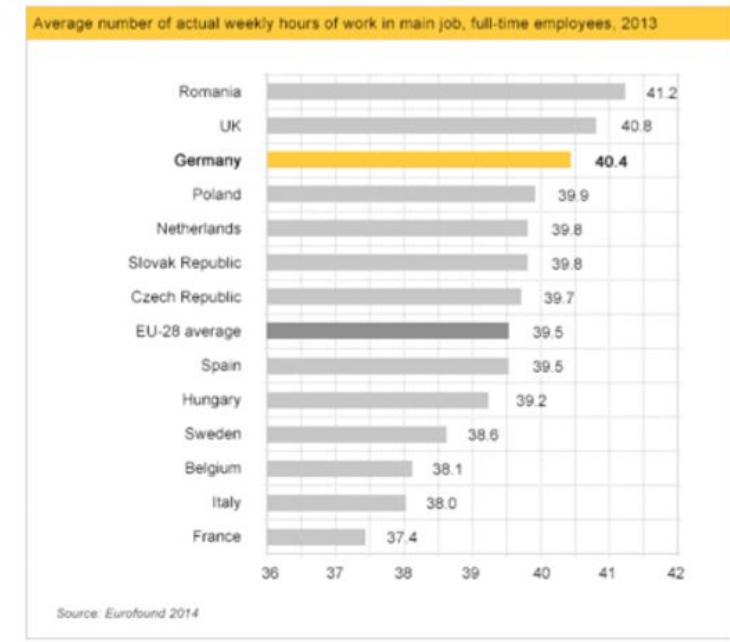


Fuente: [size of bubbles in a bubble chart | Statistical Modeling, Causal Inference, and Social Science](#)

... relacionados con el uso correcto de canales

Lie factor → Ejes engañosos

- Fenómeno ocurrido dentro del *Lie Factor* cuando el origen de una diferencia en percepción es producto a un mal uso de ejes.
- Si este tipo de visualización se ve muy rápido, las personas no alcanzan a determinar ese cambio de eje.



... relacionados con el uso correcto de canales

Lie factor → Ejes engañosos



... relacionados con el uso correcto de canales

Lie factor → Ejes engañosos



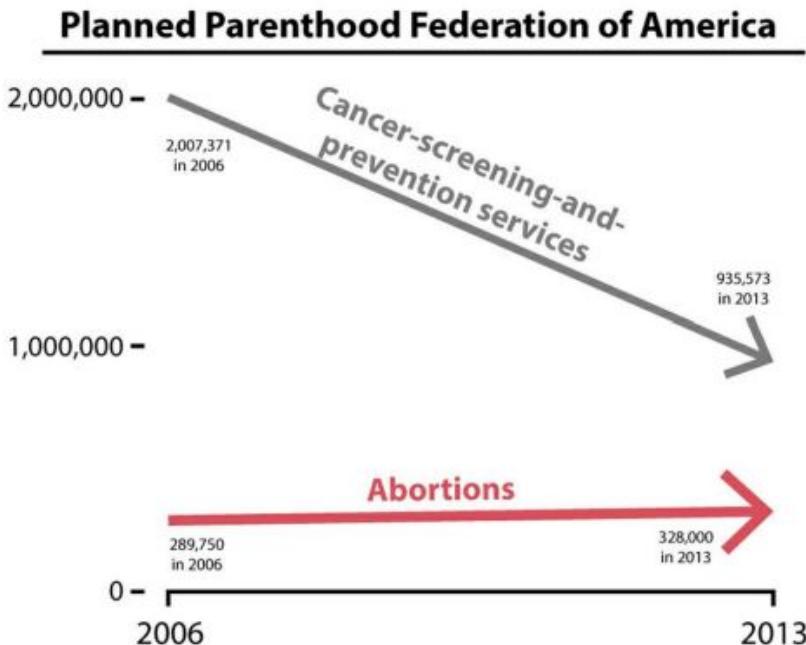
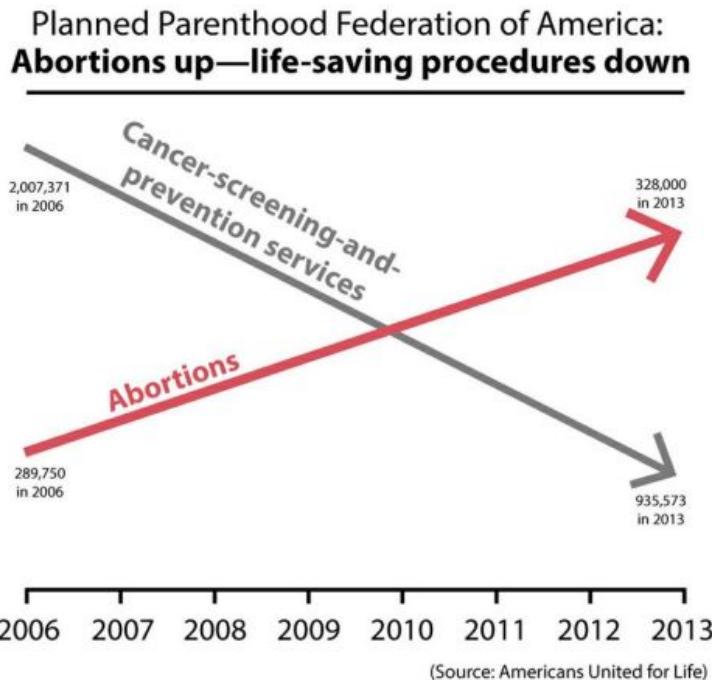
... relacionados con el uso correcto de canales

Lie factor → Ejes engañosos



... relacionados con el uso correcto de canales

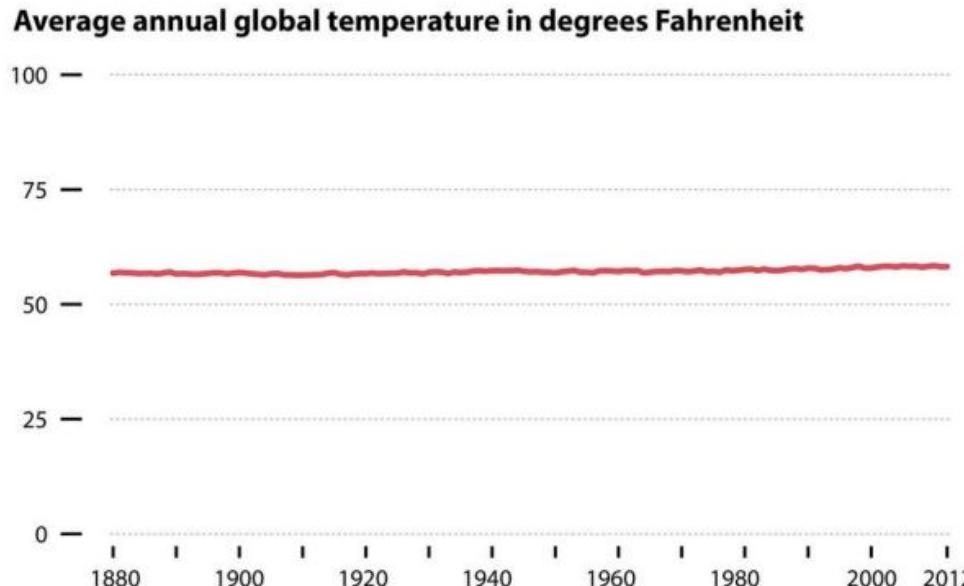
Lie factor → Ejes engañosos



... relacionados con el uso correcto de canales

Lie factor → Ejes engañosos

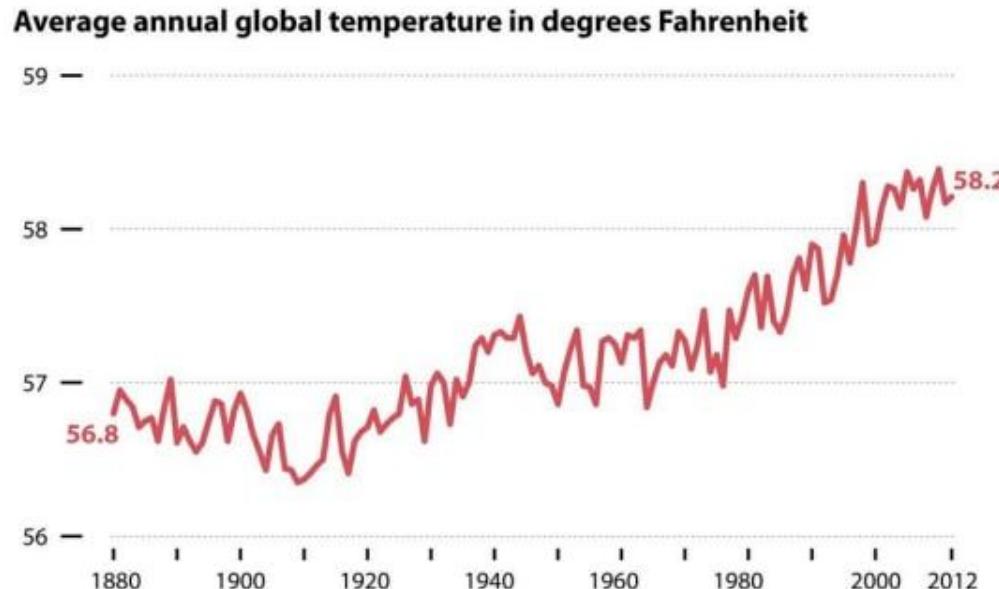
Mostrar el cero en el eje también puede ser engañoso, ya que se **intenta ocultar la tasa con la que ocurren los cambios.**



... relacionados con el uso correcto de canales

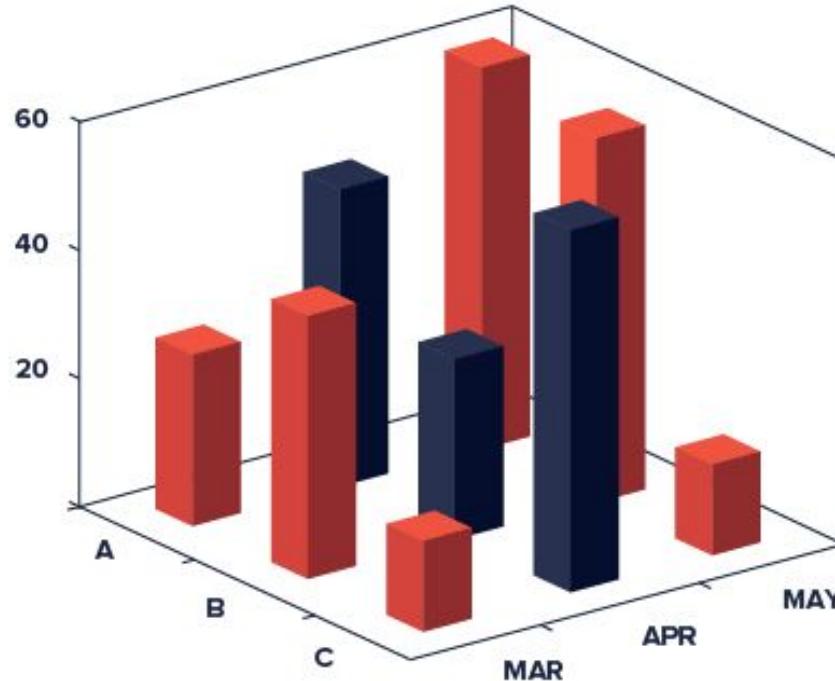
Lie factor → Ejes engañosos

Esta es una **representación más correcta** de lo que ocurre con la temperatura promedio global de nuestro planeta ya que **no es importante la magnitud, sino que el cambio.**



... relacionados con el uso correcto de canales

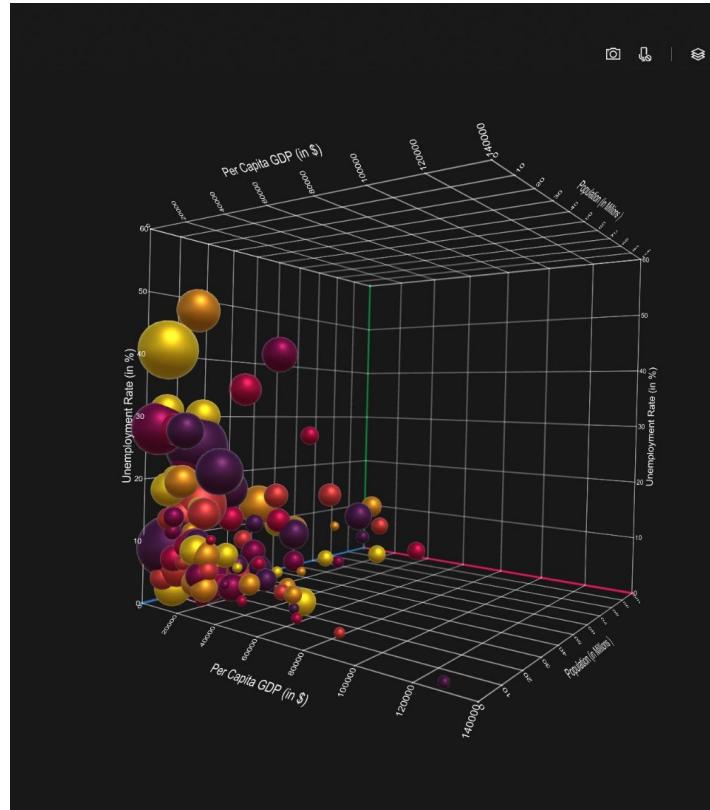
No al 3D injustificado



... relacionados con el uso correcto de canales

No al 3D injustificado

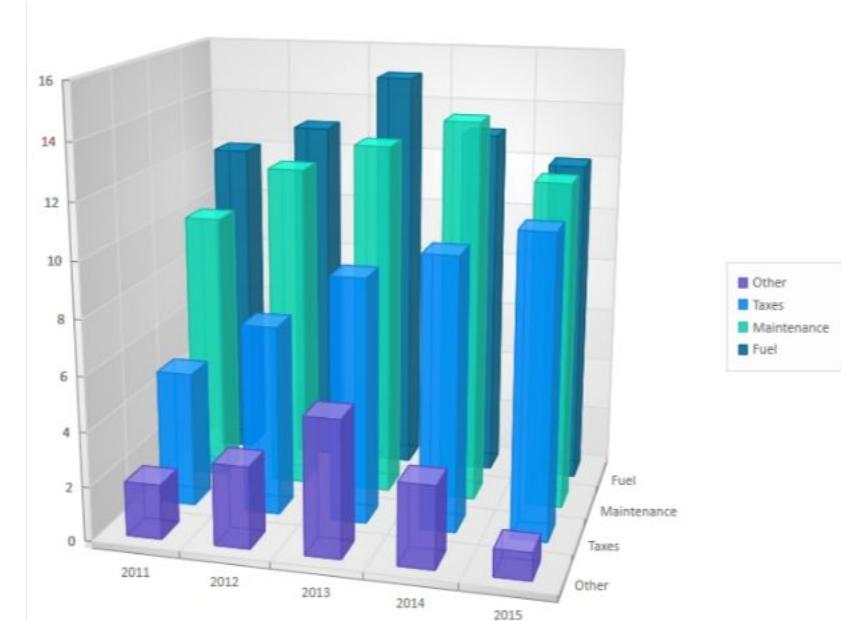
- **Oclusión:** cuando elementos quedan ocultos detrás de otros
- Si bien es posible agregar algún tipo de navegación interactiva, el costo asociado a esto es el tiempo



... relacionados con el uso correcto de canales

No al 3D injustificado

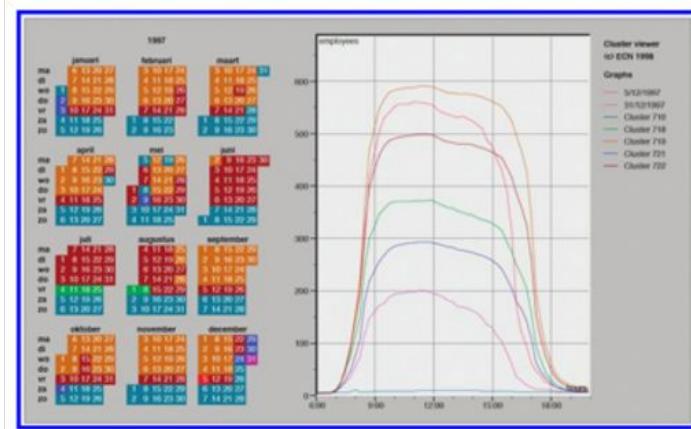
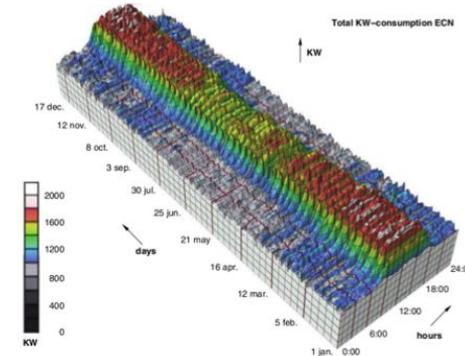
- **Distorsión por perspectiva:** cuando los objetos que están a mayor distancia se perciben más pequeños
- Por la perspectiva (y también por la oclusión), cuesta comparar los tamaños de las barras



... relacionados con el uso correcto de canales

No al 3D injustificado

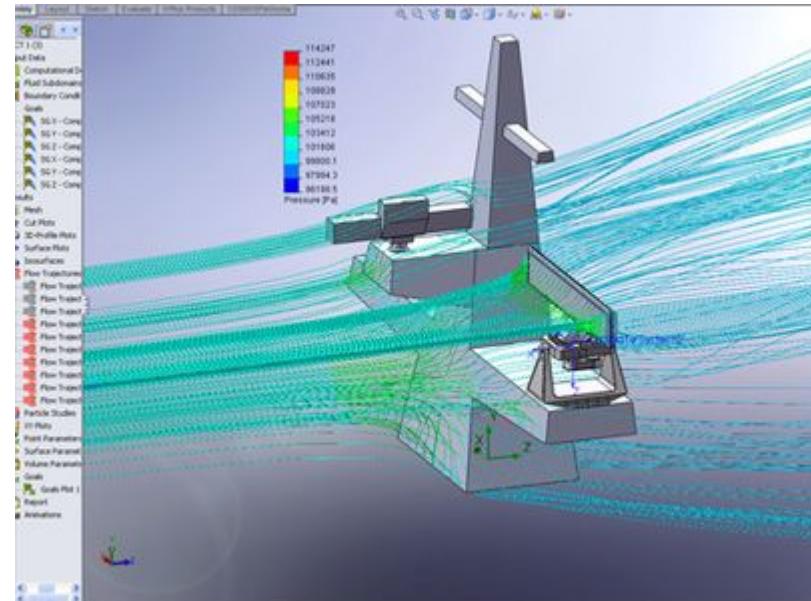
- **Distorsión por perspectiva:** cuando los objetos que están a mayor distancia se perciben más pequeños
- La idea es buscar alternativas a usar una tercera dimensión en la visualización



... relacionados con el uso correcto de canales

No al 3D injustificado

Existen situaciones donde sí se justificará usar exclusivamente 3D



... relacionados con el uso correcto de canales

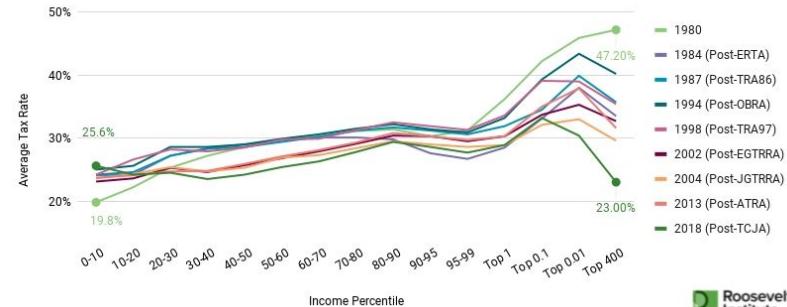
Lograrlo en blanco y negro (*Get it right in black and white*)

Algunos defienden que el aspecto más crucial de una visualización se debería transmitir incluso en blanco y negro (o escala de grises).

Se sugiere que el color sea un **canal secundario** y dentro de este, que uno varíe la saturación o luminosidad de los colores.

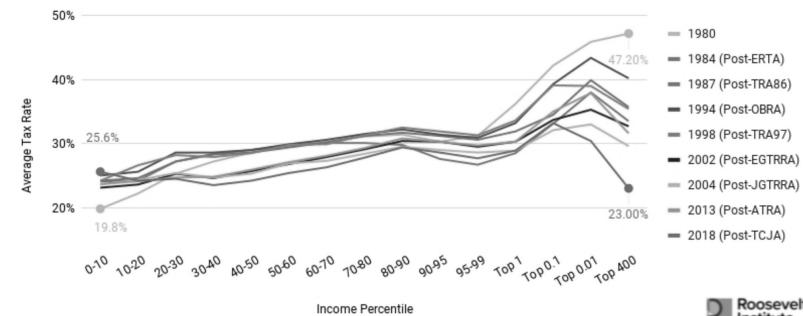
Average Tax Rates by Income Group after Tax Legislation between 1980 and 2018

While tax rates at the bottom have stayed relatively stable, or slightly increased, since 1980, tax rates for the wealthy have severely declined.



Average Tax Rates by Income Group after Tax Legislation between 1980 and 2018

While tax rates at the bottom have stayed relatively stable, or slightly increased, since 1980, tax rates for the wealthy have severely declined.



... relacionados con el uso correcto de canales

Lograrlo en blanco y negro (*Get it right in black and white*)

Algunos defienden que el aspecto más crucial de una visualización se debería transmitir incluso en blanco y negro (o escala de grises).

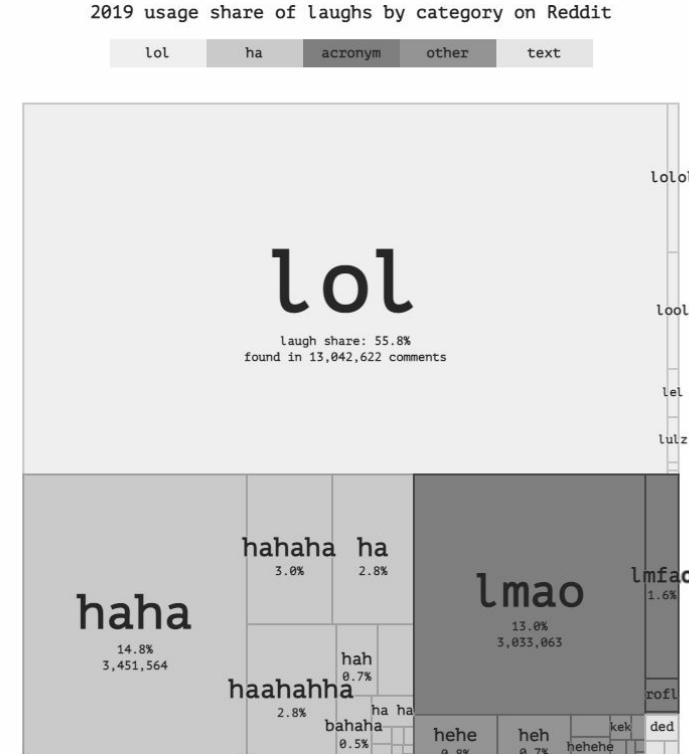
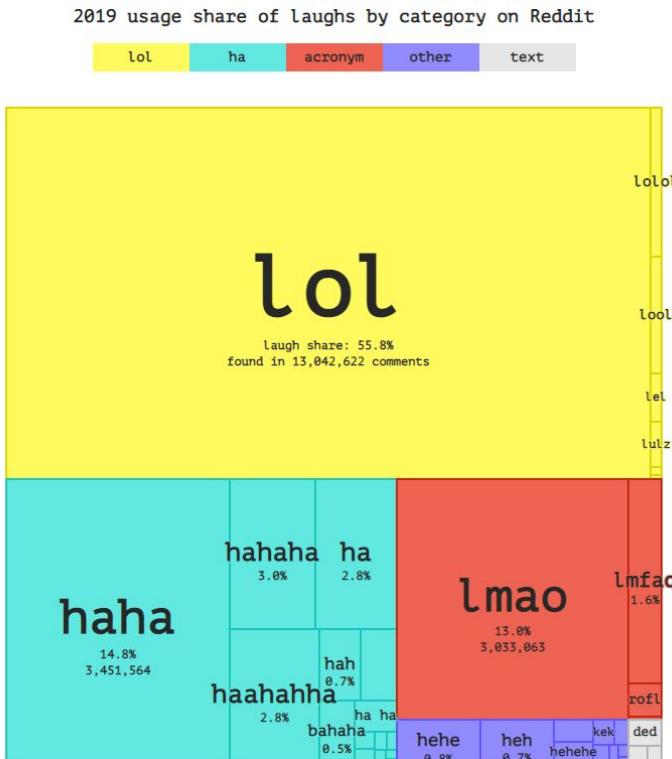
Se sugiere que el color sea un **canal secundario** y dentro de este, que uno varíe la saturación o luminosidad de los colores.

Este principio busca evitar este caso.



... relacionados con el uso correcto de canales

Lograrlo en blanco y negro (*Get it right in black and white*)



Fuente: Laughing online

... relacionados con HCI

... relacionados con HCI

- Principios de diseños relacionados con la interacción del usuario y la exposición de información por la visualización.
- Se presentan 3 principios relacionadas al HCI
 - *Overview first, details on demand.*
 - *Los ojos le ganan a la memoria (eyes beat memory).*
 - *Tiene que ser receptivo (responsive is required).*

... relacionados con HCI

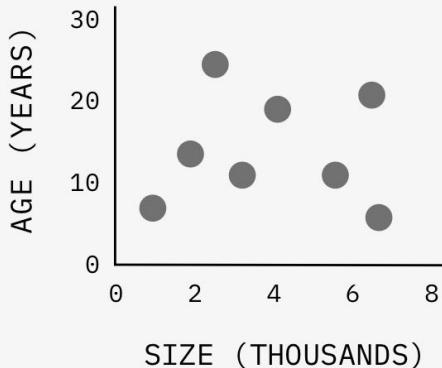
Overview first, details on demand

- Primero la visión general, los detalles a pedido
- *Guideline* escrito por Ben Shneidermann (1996) para hacer énfasis que cuando hay interacción, se debe cumplir dos requisitos: tener un **overview** y luego conocer los **detalles** a pedido del usuario.
- Lo esperado es que primero el usuario entienda la panorámica de la información antes de ir a lo particular.

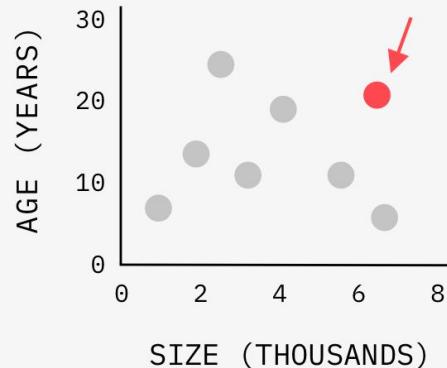
... relacionados con HCI

Overview first, details on demand

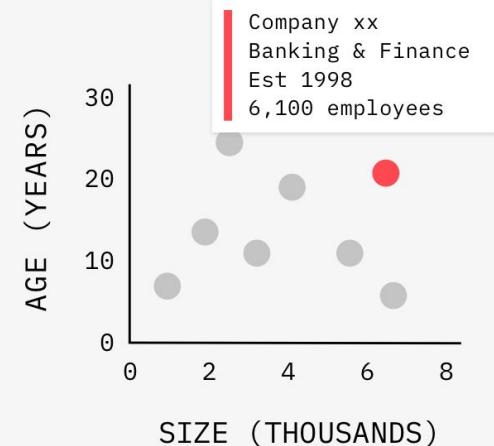
Example: companies by age and size



1. OVERVIEW FIRST



2. ZOOM AND FILTER



3. DETAILS ON DEMAND

... relacionados con HCI

Overview first, details on demand

La página parte mostrando todos los países de la base de datos. Luego se puede ir a un gráfico de burbuja que filtra ciertos países. **Cumple el principio** 🎉



... relacionados con HCI

Los ojos le ganan a la memoria (*eyes beat memory*)

Es más fácil usar la **cognición externa** que nuestra memoria interna

Por lo tanto, es más fácil comparar, moviendo nuestros ojos de lado a lado, que hacerlo tratando de recordar algo que vimos recientemente

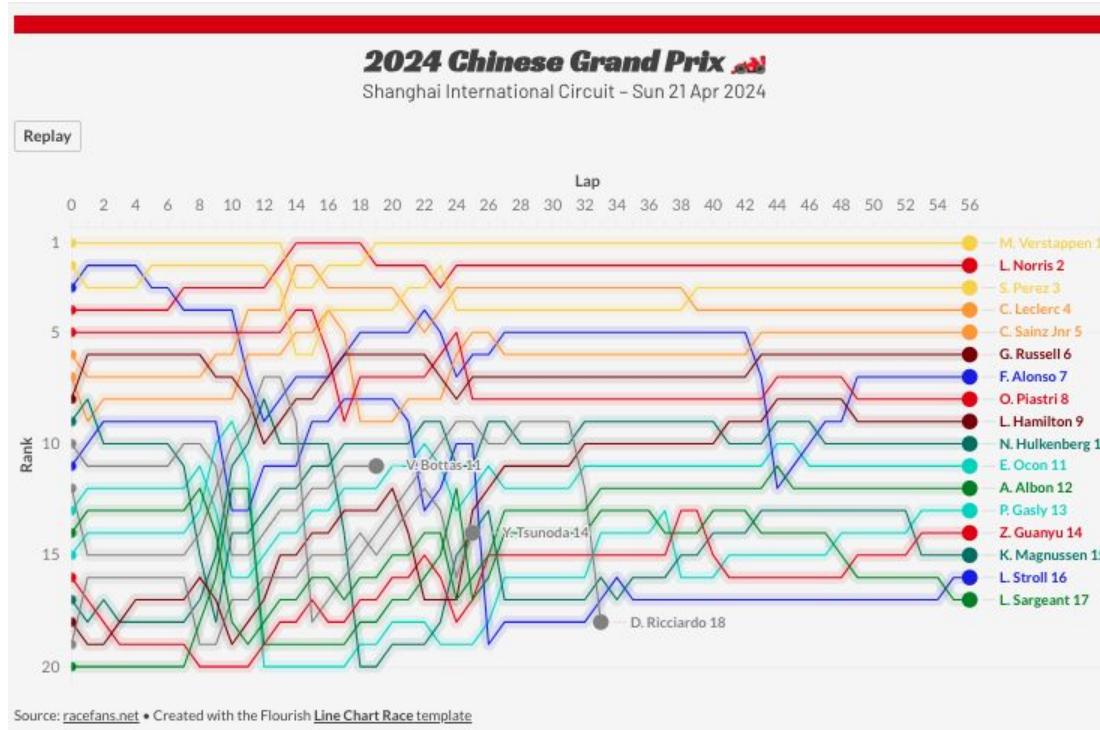
2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support



Orange and green colors correspond to states where support for vouchers was greater or less than the national average. The seven ethnicreligious categories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants. Where a category represents less than 1% of the voters of a state, the state is left blank.

... relacionados con HCI

Los ojos le ganan a la memoria (*eyes beat memory*)



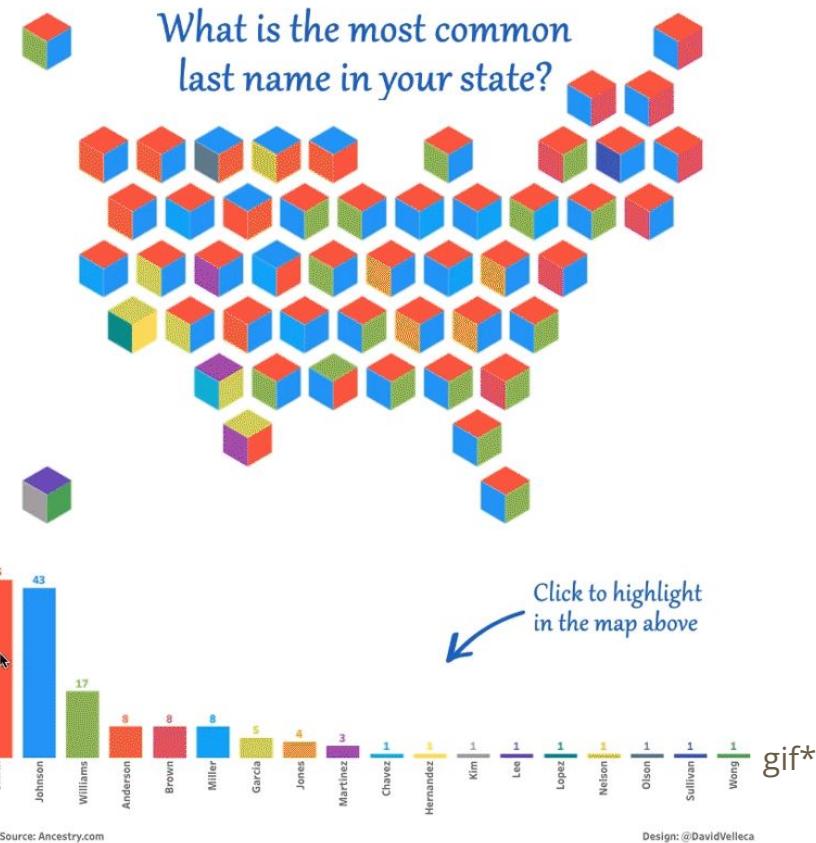
Fuente: [Ready, set, race! How to make a line chart race visualization without coding | The Flourish blog](#)

... relacionados con HCI

Tiene que ser receptivo (*responsive is required*)

Nuestro **nivel de irritación** crece mientras las operaciones van tomando más y más tiempo.

Se recomienda que la herramienta pueda **entregar un feedback ante cualquier acción del usuario**.



... relacionados con HCI

Tiene que ser receptivo (*responsive is required*)

Se recomienda que la herramienta pueda **entregar un feedback ante cualquier acción del usuario.**

Si es que una operación está tomando más tiempo de lo que el usuario esperaría, una **barra de progreso** debería ser mostrada al usuario



... relacionados con diseño gráfico

... relacionados con diseño gráfico

- Principios de diseño relacionados con el diseño de las visualizaciones y documentos que incluyan estas.
- Vamos a analizar 3 principios:
 - Tasa de tinta de datos (*Data ink ratio*) propuesto por Edward Tufte
 - Consistencia (interna y externa).
 - Autocontención.

... relacionados con diseño gráfico

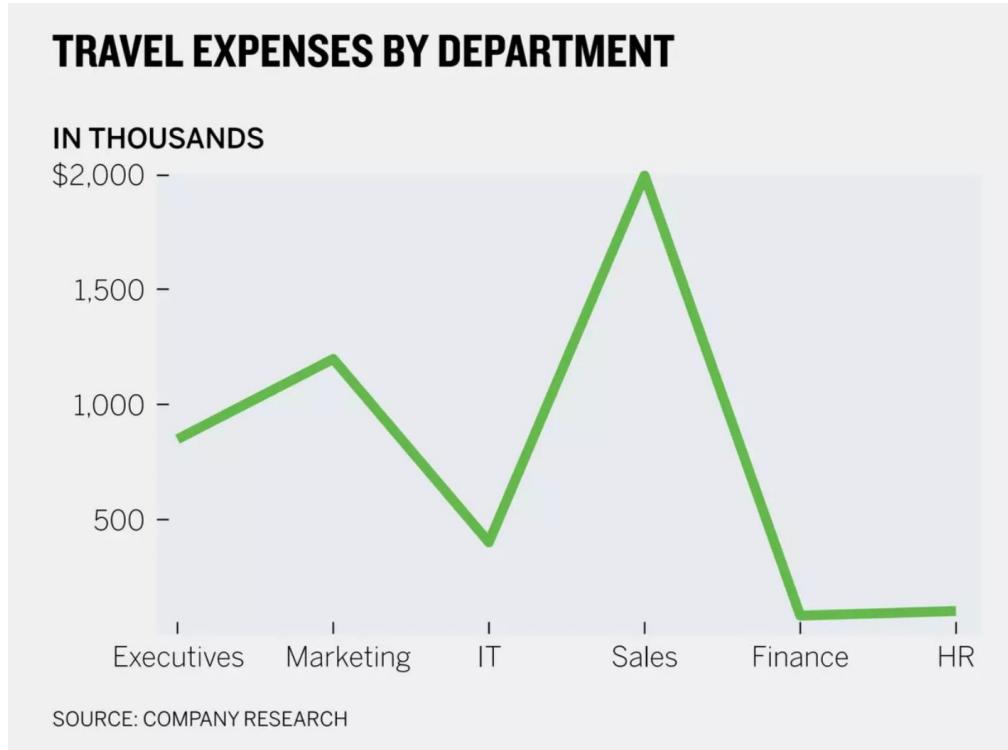
Tasa de tinta de datos (*Data ink ratio*)

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

- El espíritu de este principio es que **el uso de cada píxel en una visualización esté justificado.**
- Cada marca/canal que usemos tenga una razón por la cual se está usando.
- Si tenemos una tasa de 1, significa que hemos justificado cada pixel de nuestra visualización.

... relacionados con diseño gráfico

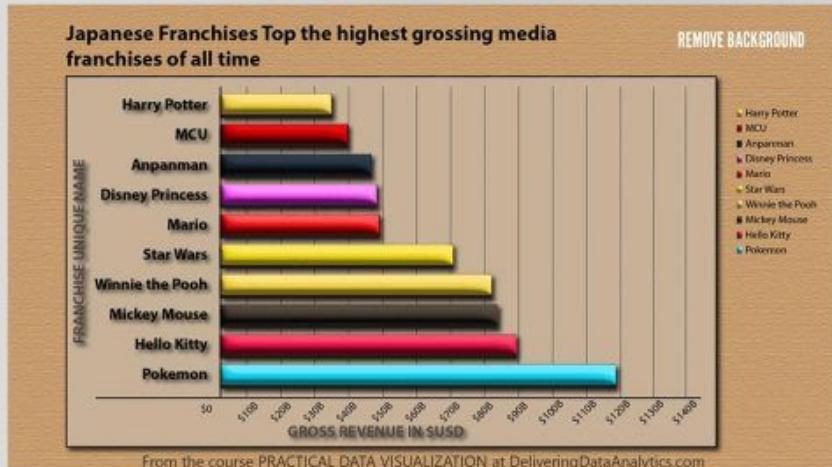
Tasa de tinta de datos (*Data ink ratio*)



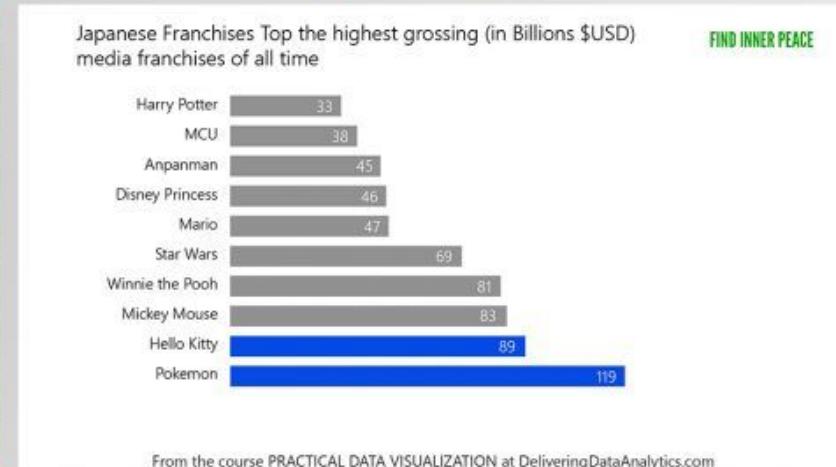
... relacionados con diseño gráfico

Tasa de tinta de datos (*Data ink ratio*)

BEFORE

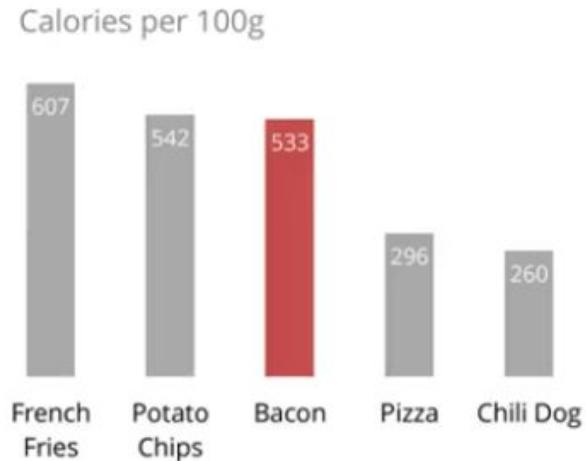
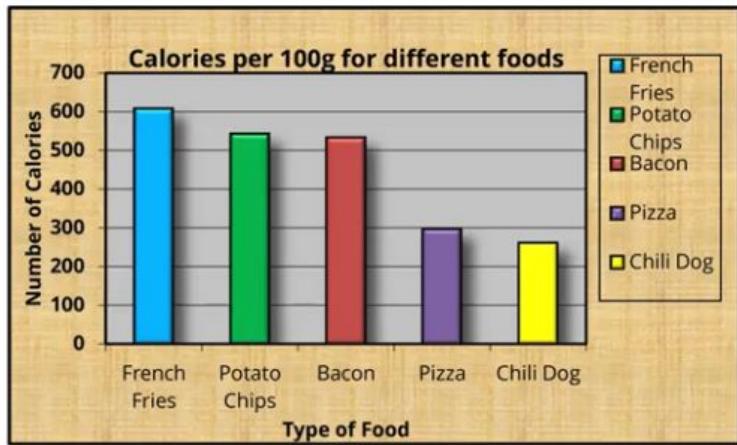


AFTER



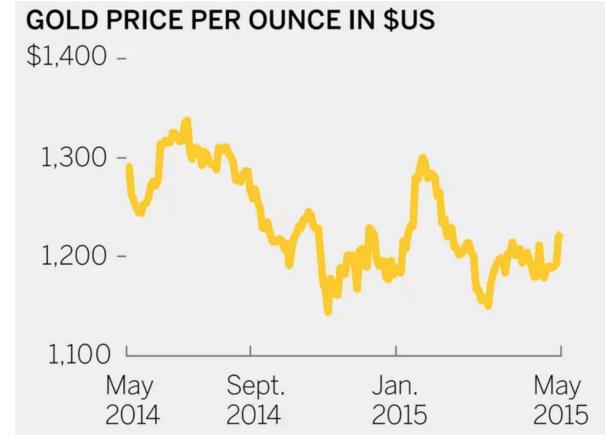
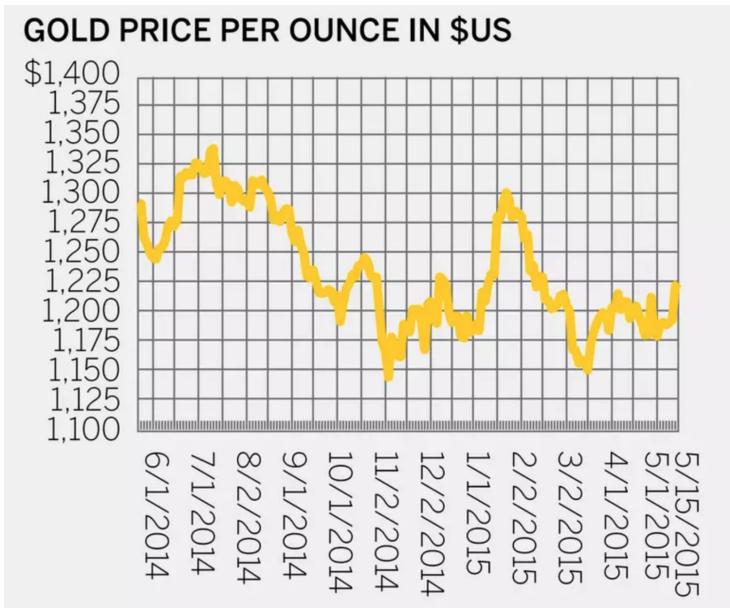
... relacionados con diseño gráfico

Tasa de tinta de datos (*Data ink ratio*)



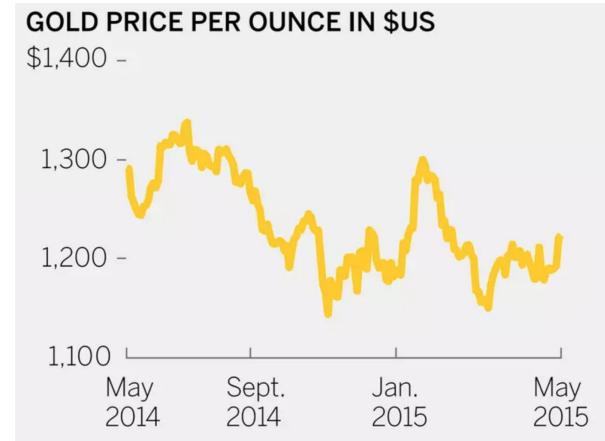
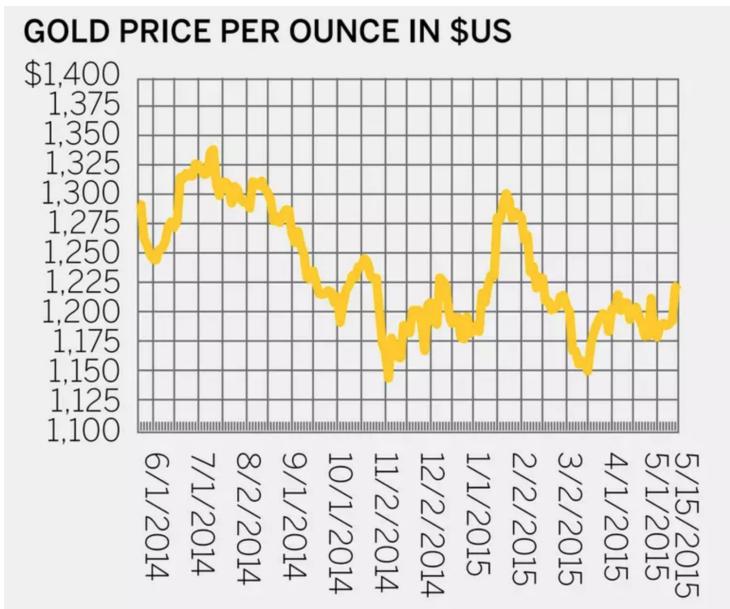
... relacionados con diseño gráfico

Tasa de tinta de datos (*Data ink ratio*)



... relacionados con diseño gráfico

Tasa de tinta de datos (*Data ink ratio*)

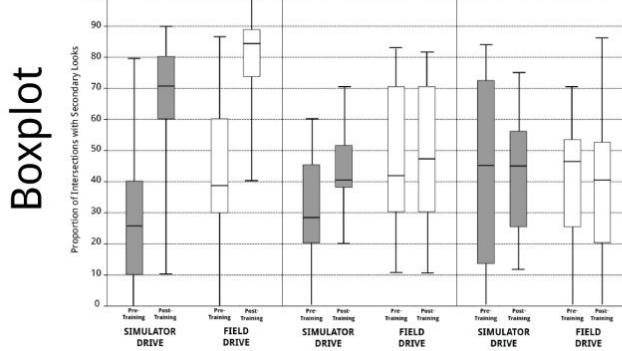
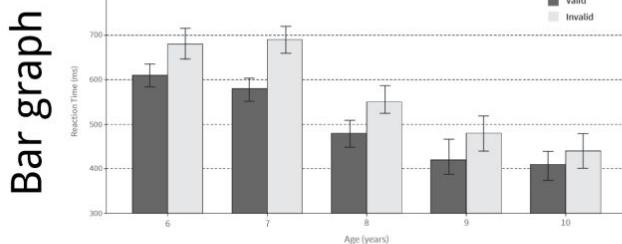


Evaluemos qué elementos de una visualización son necesarios y cuáles no.

... relacionados con diseño gráfico

Tasa de tinta de datos (*Data ink ratio*)

Low Data-Ink

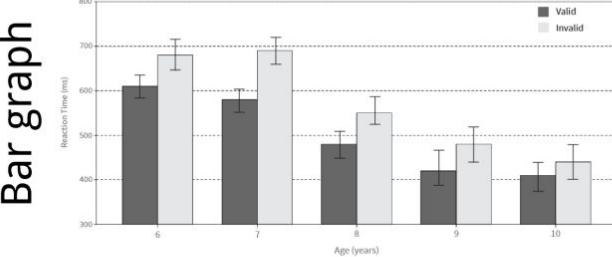


Fuente: [Experimental stimuli: low, medium, and high data-ink bar graph and boxplots.](#)

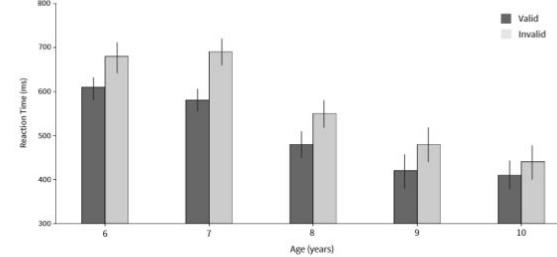
... relacionados con diseño gráfico

Tasa de tinta de datos (*Data ink ratio*)

Low Data-Ink

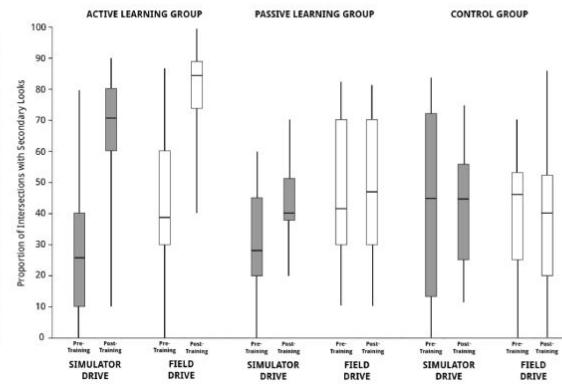
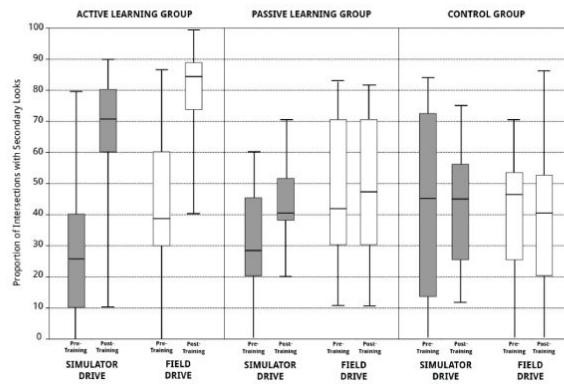


Medium Data-Ink



Bar graph

Boxplot

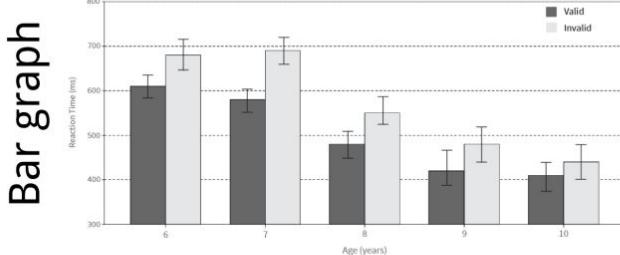


Fuente: [Experimental stimuli: low, medium, and high data-ink bar graph and boxplots.](#)

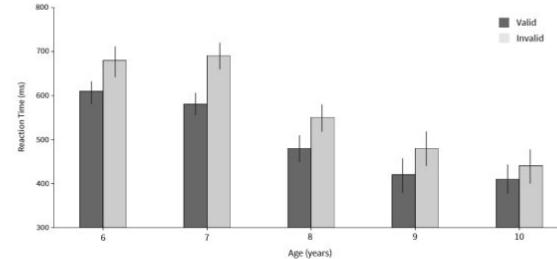
... relacionados con diseño gráfico

Tasa de tinta de datos (*Data ink ratio*)

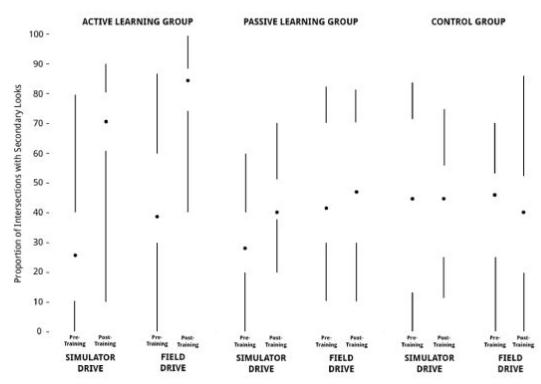
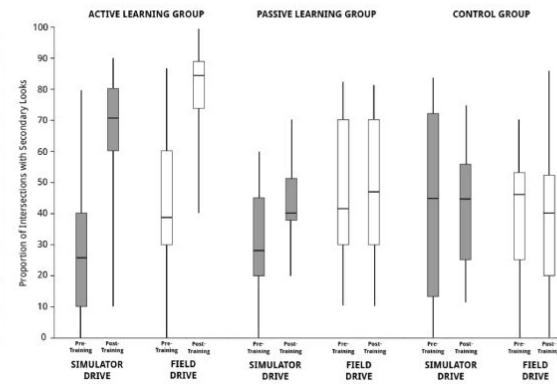
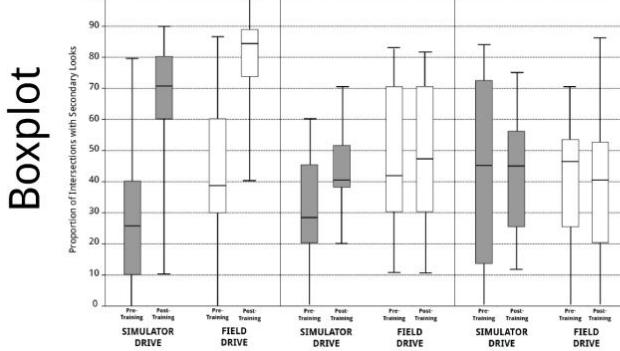
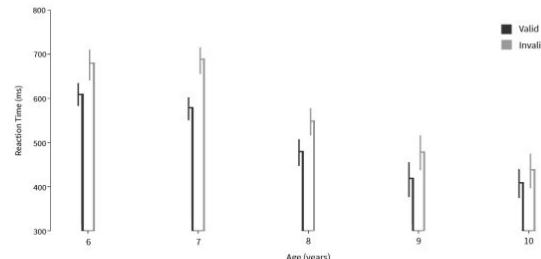
Low Data-Ink



Medium Data-Ink



High Data-Ink



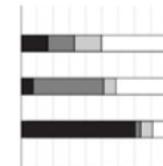
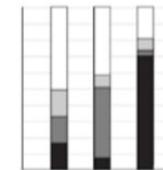
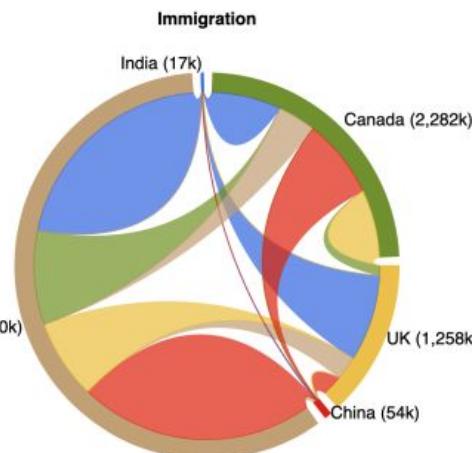
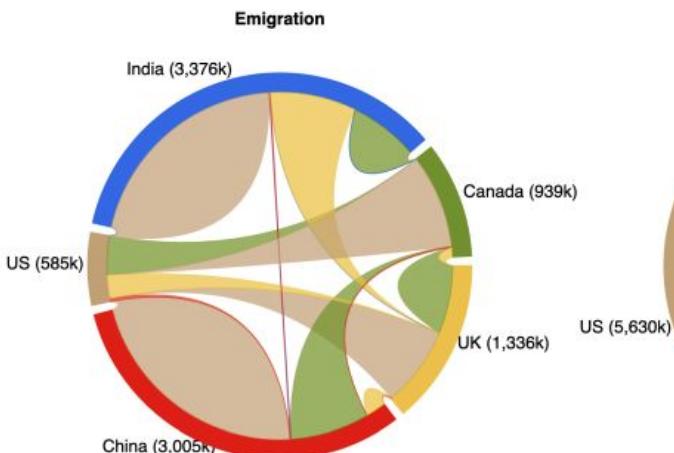
Fuente: Experimental stimuli: low, medium, and high data-ink bar graph and boxplots.

... relacionados con diseño gráfico

Consistencia - interna

Mantén las decisiones de diseño realizadas durante todo el documento.

Si decides que un país tendrá un color determinado, que se mantenga así en las demás visualizaciones.



... relacionados con diseño gráfico

Consistencia - interna

Mantén las decisiones de diseño realizadas durante todo el documento.

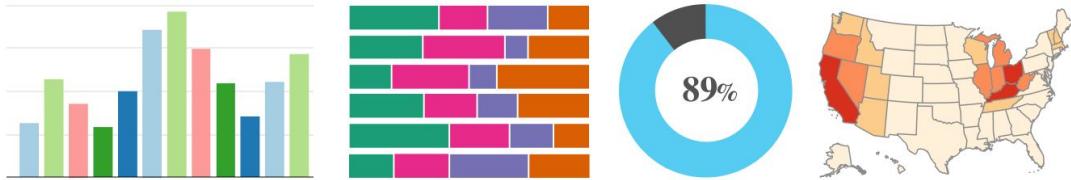


... relacionados con diseño gráfico

Consistencia - interna

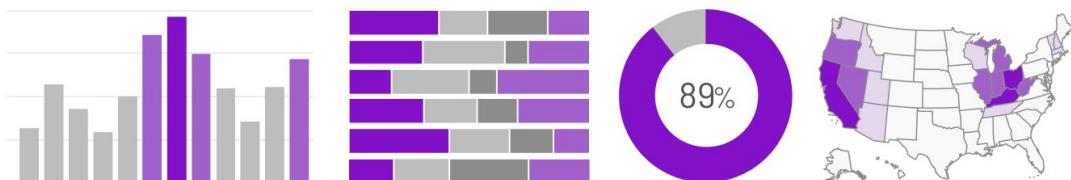
Don't ✗

Disparate color and type palettes within the same product create disharmony.



Do ✓

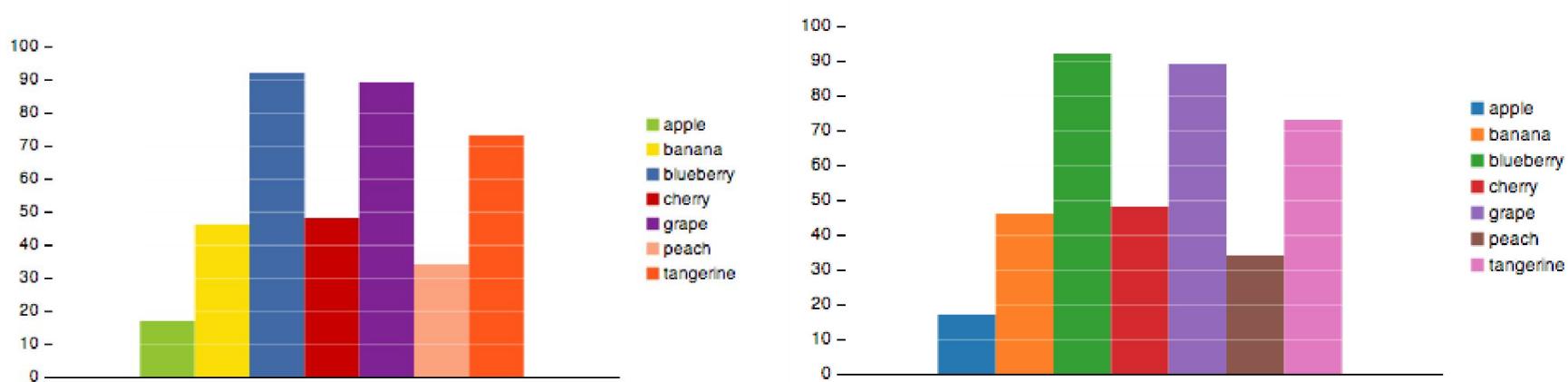
Sticking to a set style guide ensures cohesion among data visualizations.



... relacionados con diseño gráfico

Consistencia - externa

Tomar elecciones inspiradas en referentes externos al documento o visualización, y que potencialmente permiten una mejor comprensión gracias al conocimiento previo.



... relacionados con diseño gráfico

Consistencia - externa

Las fechas van de mayor a menor, situación que uno no está "acostumbrado" en un gráfico de barra, sino que vayan de menor a mayor.



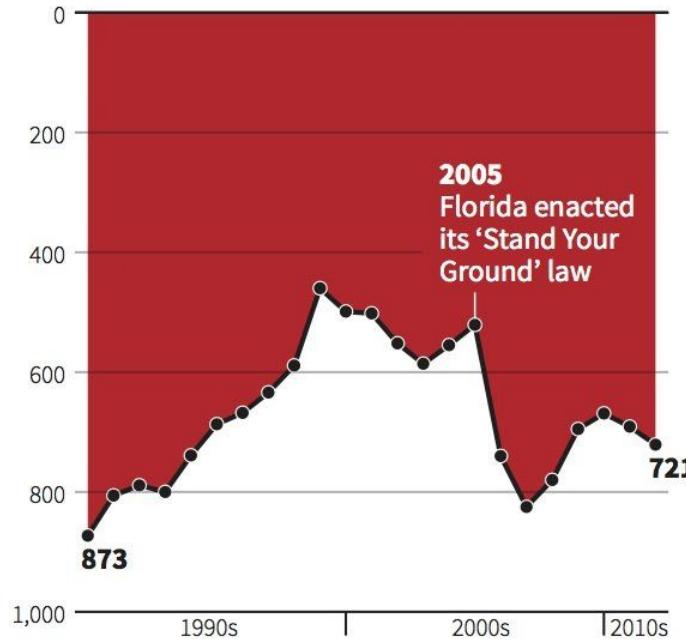
... relacionados con diseño gráfico

Consistencia - externa

Que los datos crezcan hacia abajo no es algo el cual estamos acostumbrados

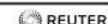
Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

C. Chan 16/02/2014



... relacionados con diseño gráfico

Autocontención

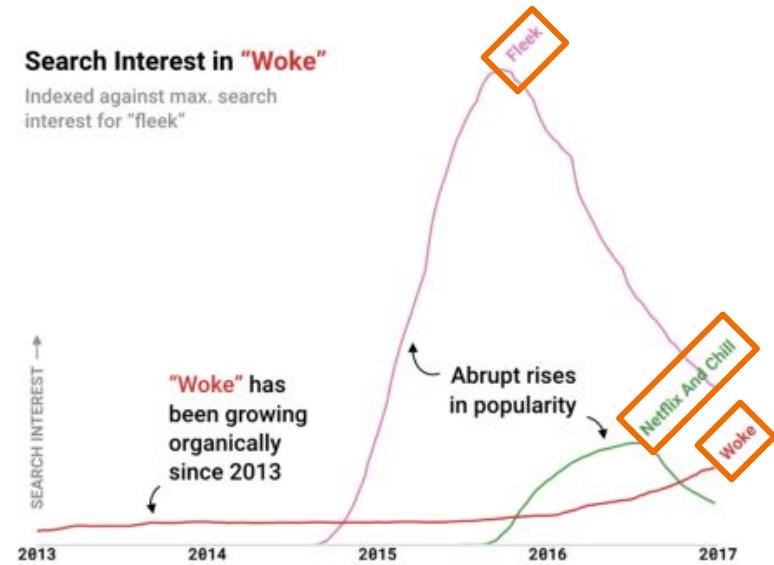
El contenido de una visualización, en conjunto a su contenedor, debe ser autoexplicativo.

... relacionados con diseño gráfico

Autocontención

El contenido de una visualización, en conjunto a su contenedor, debe ser autoexplicativo.

- Buen uso de leyendas.

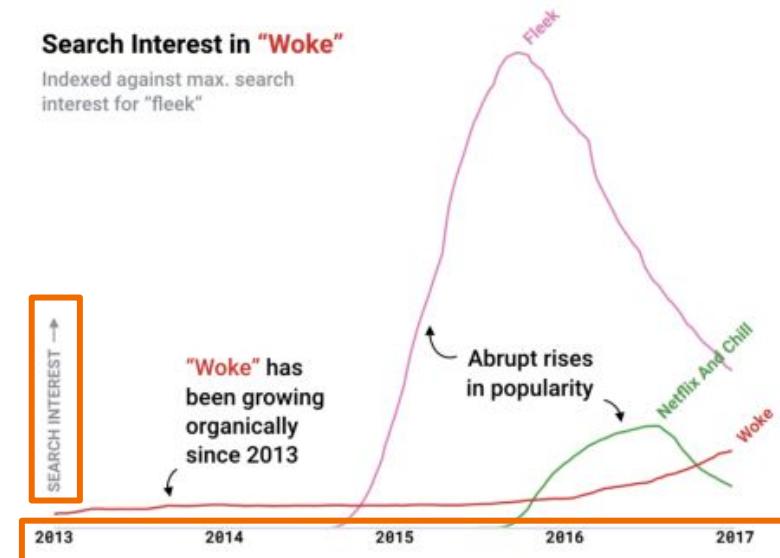


... relacionados con diseño gráfico

Autocontención

El contenido de una visualización, en conjunto a su contenedor, debe ser autoexplicativo.

- Buen uso de leyendas.
- Buen uso de ejes.

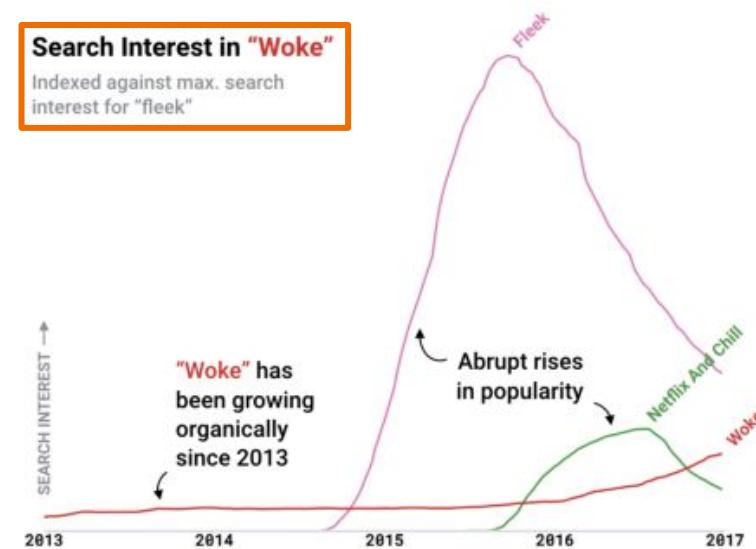


... relacionados con diseño gráfico

Autocontención

El contenido de una visualización, en conjunto a su contenedor, debe ser autoexplicativo.

- Buen uso de leyendas.
- Buen uso de ejes.
- Buen uso de títulos en una visualización.



... relacionados con diseño gráfico

Autocontención

El contenido de una visualización, en conjunto a su contenedor, debe ser autoexplicativo.

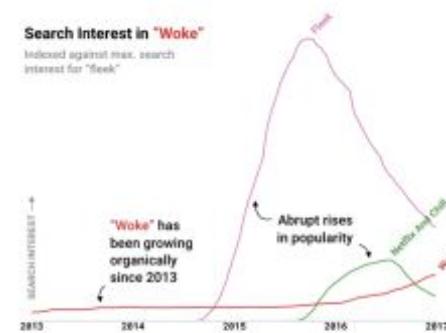
- Buen uso de leyendas.
- Buen uso de ejes.
- Buen uso de títulos en una visualización.
- Agregar contexto a la visualización (texto introductorio previo y posterior).

Catalyst 2: Politics

In the US, last year's political circus tipped five of 2016's top slang into the mainstream.

Woke (noun) - aware of (racial) social injustice

"Woke" is the one term on 2016's list that's been growing organically for at least a decade. There was no major event that propelled it into the spotlight – it just finally crossed the threshold of mainstream use.



There were some unique events in 2016 for "woke." In terms of semantics, there's the notion that 2016's US election cycle whitewashed the term. Today, I see it used broadly as "political awareness," beyond its historic racial connotations. Childish Gambino's "Redbone" helped too (~50M streams on Spotify), which dropped in November 2016. The song's hook repeats the phrase "stay woke," helping to normalize the phrase.

Principios de diseño ...

... relacionados con el uso correcto de canales

- Factor de la mentira (*Lie factor*)
- Ejes engañosos
- No al 3D injustificado
- Lograrlo en blanco y negro (*Get it right in black and white*)

... relacionados con HCI

- *Overview first, details on demand*
- Los ojos le ganan a la memoria (*Eyes beat memory*)
- Tiene que ser receptivo (*Responsive is required*)

... relacionados con el diseño de la visualización

- Tasa de tinta de datos (*Data ink ratio*)
- Consistencia interna y externa
- Autocontención

Otras referencias

- “*The Visual Display of Quantitative Information*” — Edward Tufte
- “*Visualization, Analysis and Design*” — Tamara Munzner
- [Data Is Ugly](#)
- [WTFViz](#)
- [Response Time Limits: Article by Jakob Nielsen](#)
- [Calling Bullshit: in the age of big data](#)
- [99 Rules of Data Viz \(and why you should break them\) — AddTwo](#)
- [Data vis do's & don'ts - Datawrapper Blog](#)

Presentación de diversos gráficos

Presentación de diversos gráficos

- Veremos un compilado de gráficos en donde se entregará:
 - Una descripción del gráfico.
 - En qué situaciones se utilizan generalmente.
 - Código para programarlo con Pandas, Matplotlib y/o Seaborn.
- Catálogo 100% recomendado: [Data Viz Project](#)
- Códigos para confeccionar estas visualizaciones se encuentran en:
[\[MIA\] Clase 4 - Visualizaciones Tabulares Matplotlib-Seaborn](#)
- Datos: [Heart Failure Prediction Dataset](#)

Presentación de diversos gráficos

Comparación y tendencias
Gráfico de barra y línea

Gráfico de barras

- Utilizado para representar un atributo **categórico** y otro **numérico**.
- Uno de los ejes muestran los datos **categóricos**.
- El otro muestran los datos **cuantitativos**.
- Gráfico muy efectivo para hacer **comparaciones**.

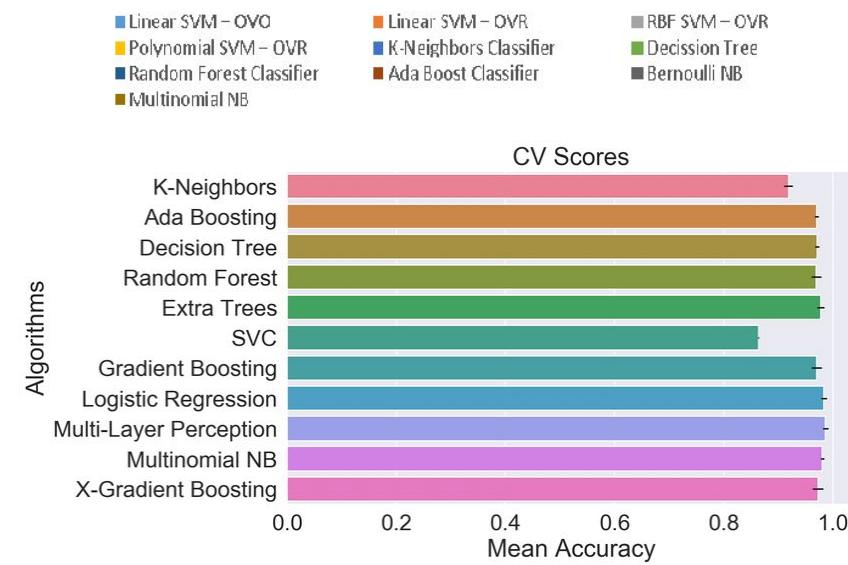
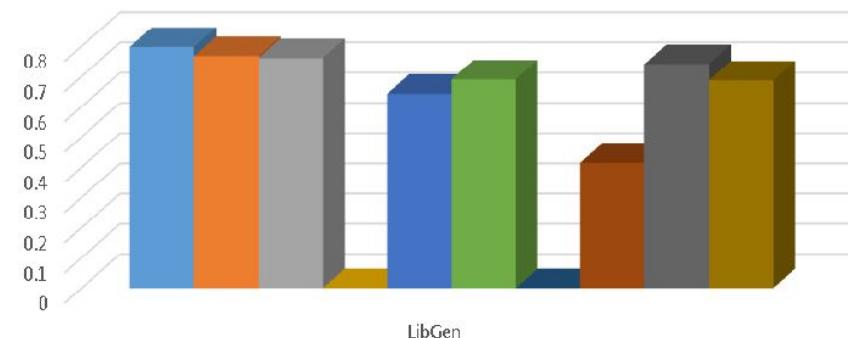
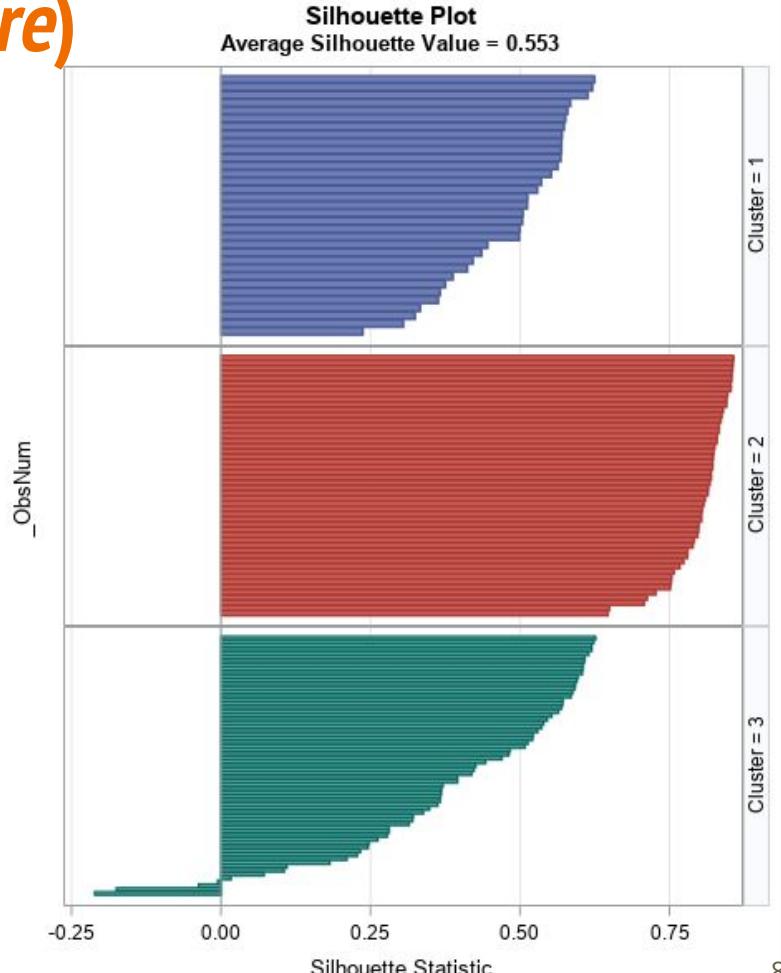
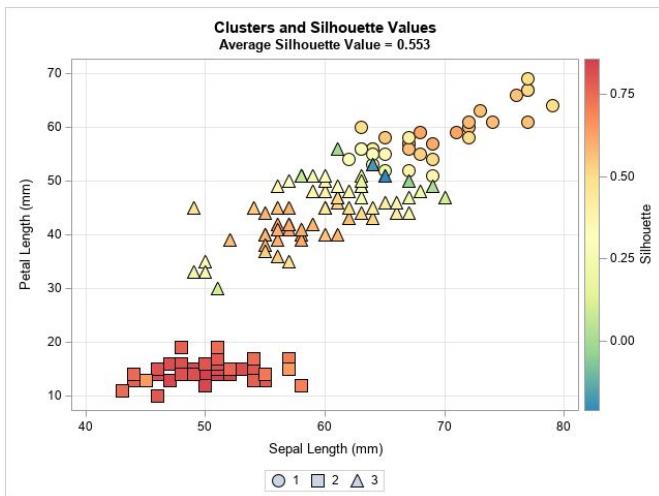


Gráfico de barras (*Silhouette Score*)

- Utilizado para evaluar la calidad de *clustering*.
- Se mide la similitud de cada punto de datos con el *cluster* al que pertenece y su diferencia con otros *clusters*



Fuente: [Compute the silhouette statistic in SAS - The DO Loop](#)

Gráfico de línea

- Utilizados para presentar **tendencias**
- En el eje **X** se pone una variable **ordenada** (típicamente datos temporales)
- En el eje **X** puede perfectamente ponerse **cientos de niveles** (valores) y el gráfico escala bien visualmente

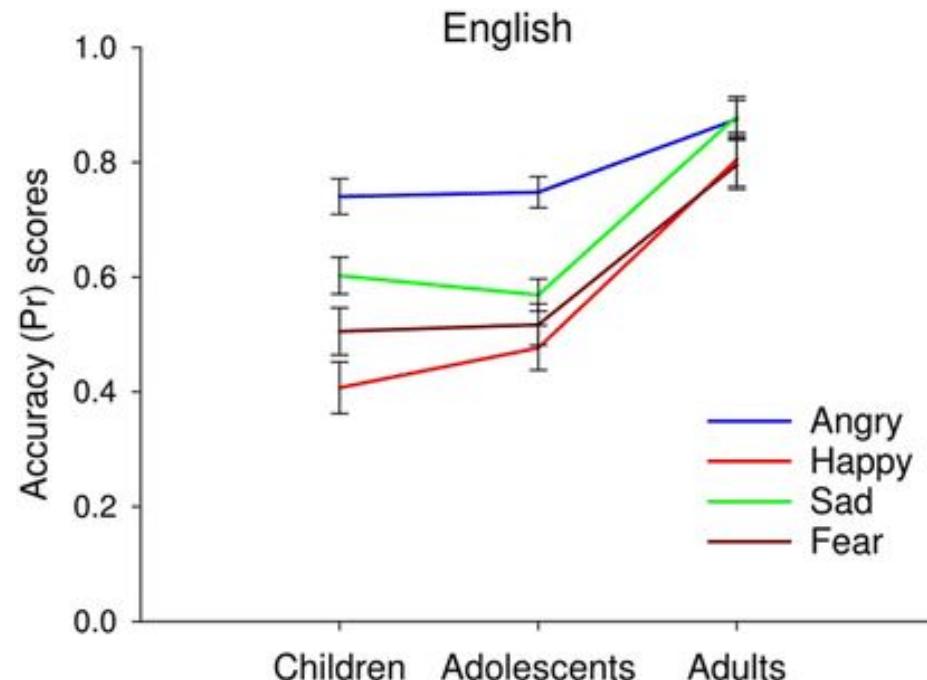


Gráfico de línea

- El uso de línea provoca una interpretación de orden en los puntos. Por este motivo, un gráfico de línea **debe presentar un eje X ordenable**

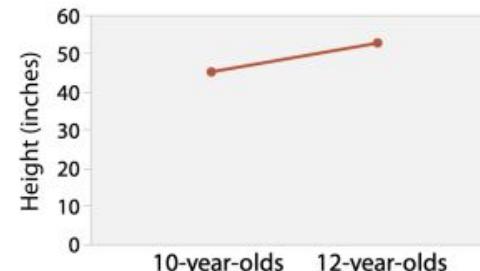
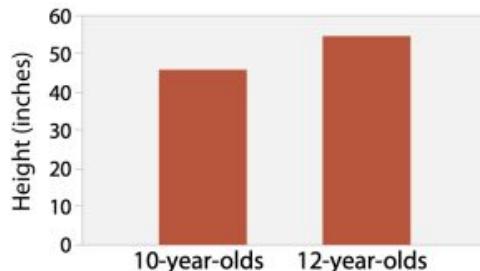
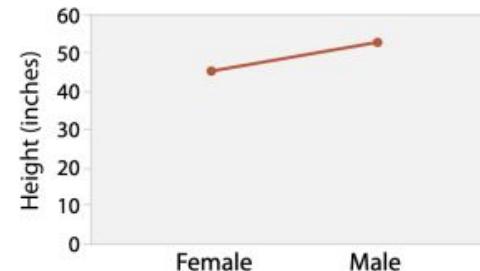
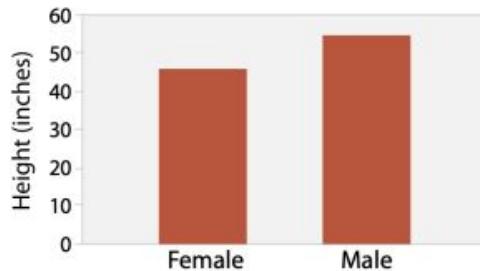


Gráfico de línea (ROC Curve)

- Dado un modelo que predice con cierta probabilidad y un resultado binario. Se determinan diferentes umbrales de probabilidad para ver cómo clasifica el modelo.
- *True Positive Rate*: El modelo predice correctamente que algo es positivo, dividido el total de datos que son positivos.
- *False Positive Rate*: El modelo predice positivamente algo que en realidad es negativo, dividido el total de datos negativos.

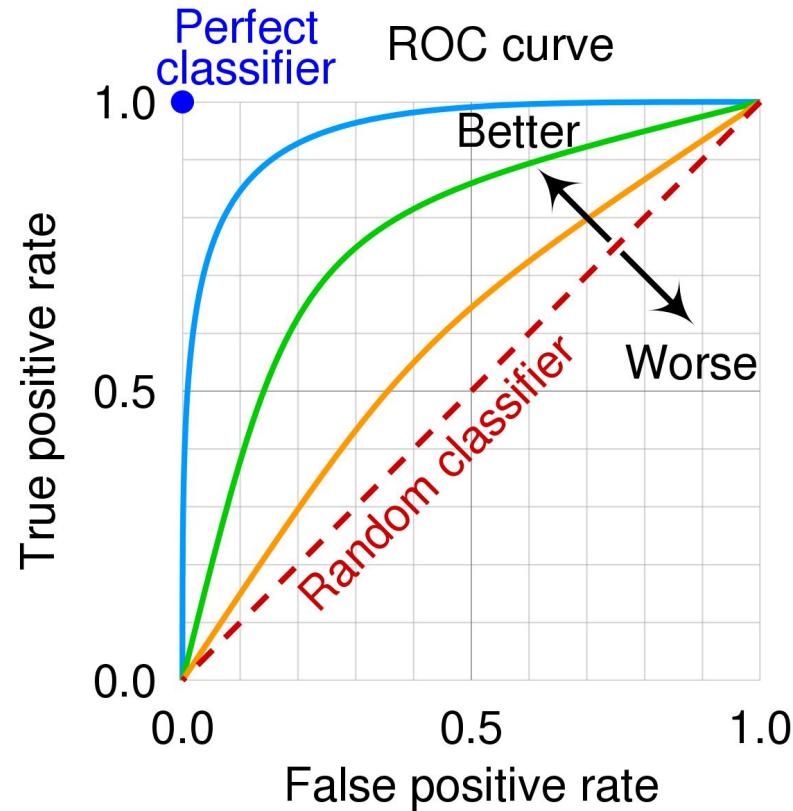


Gráfico de línea (*Elbow Curve*)

- Técnica que se utiliza para determinar el número óptimo de *clusters*.
- Se calcula la suma de las distancias al cuadrado entre cada punto de datos y su centroide asignado para diferentes valores de k .
- Se busca donde la línea ya no presenta una pendiente tan pronunciado.

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

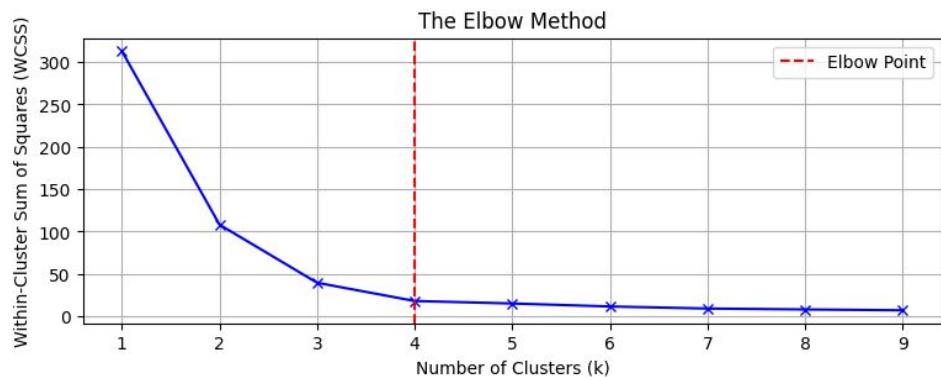
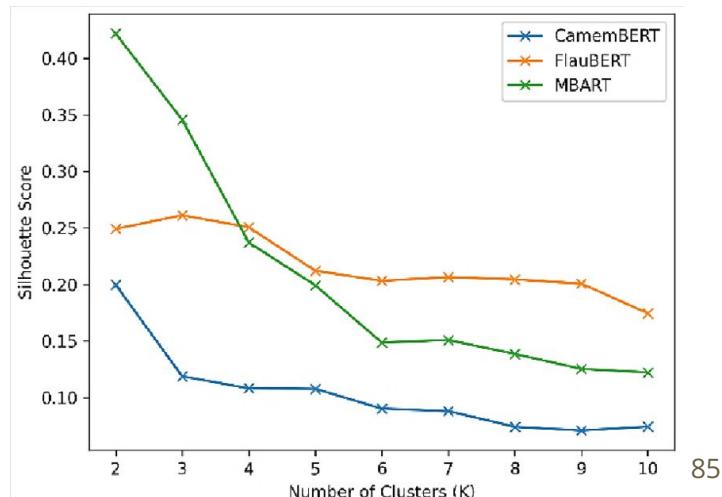
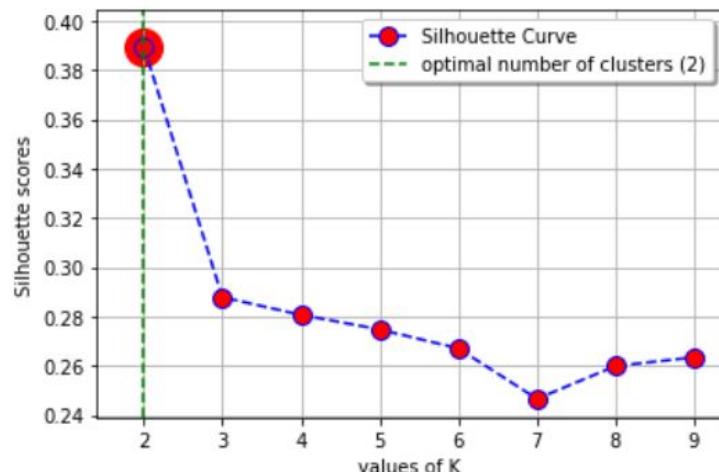


Gráfico de línea (*Silhouette Score*)

- Esta métrica también se puede usar como técnica para determinar el número óptimo de *clusters*.
- Para diferente número de *clusters* se calcula el *silhouette score* promedio.
- Se busca el punto más alto.
- Se puede utilizar para contrastar varios modelos de *clustering* simultáneamente.



Código en Python

Vamos al código



Presentación de diversos gráficos

Correlaciones

Gráfico de dispersión y de burbuja

Gráfico de dispersión

- En inglés se conoce como **scatterplot**.
- Utilizado para visualizar **relaciones entre dos atributos cuantitativos**.
- Permite visualizar fácilmente si es que existe **correlación**, tanto positiva como negativa.
- Se usa el canal de posición horizontal y vertical (X,Y) para transmitir la información de los 2 atributos cuantitativos.

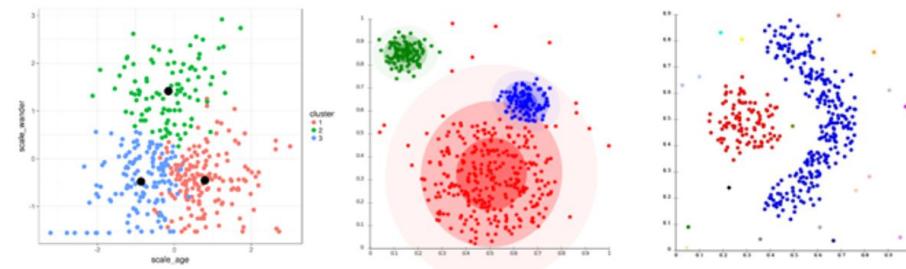
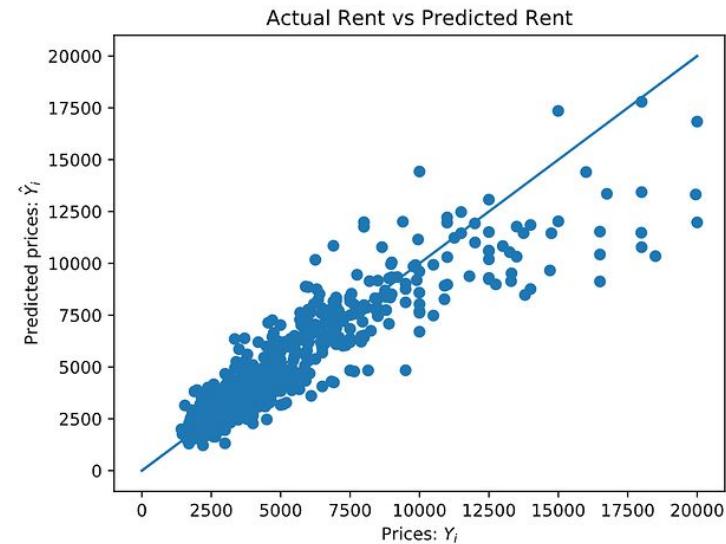


Gráfico de dispersión (ejemplos)

- Muestra una correlación **positiva**
- Ilustra que **no hay relación** entre las variables X e Y

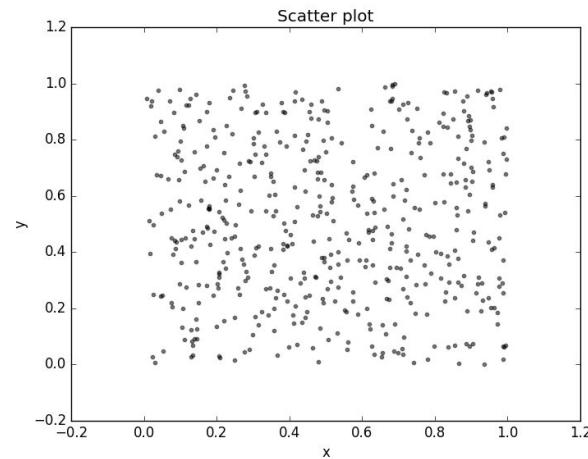
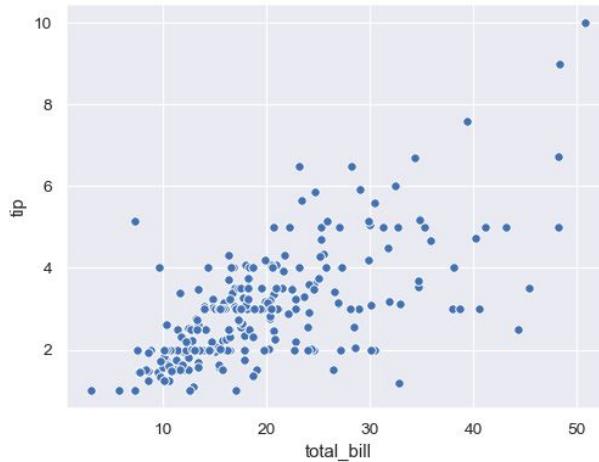


Gráfico de burbuja

- **Extensión** del gráfico de dispersión.
- Permite codificar **más atributos** usando el **color** y el **tamaño** de los puntos.
- El tamaño del círculo se utiliza para atributos **ordenados**
- El color puede usarse para atributos **categóricos u ordenados** dependiendo de la paleta de colores utilizada.

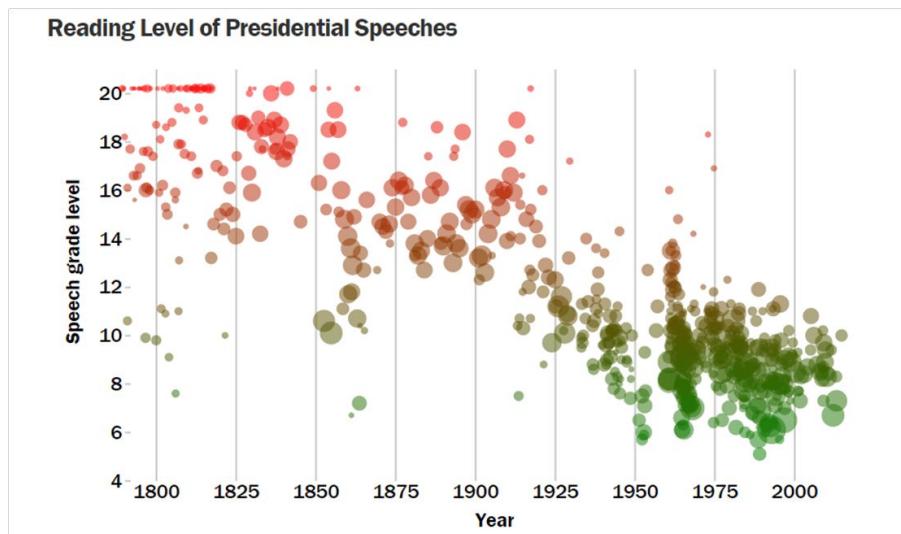
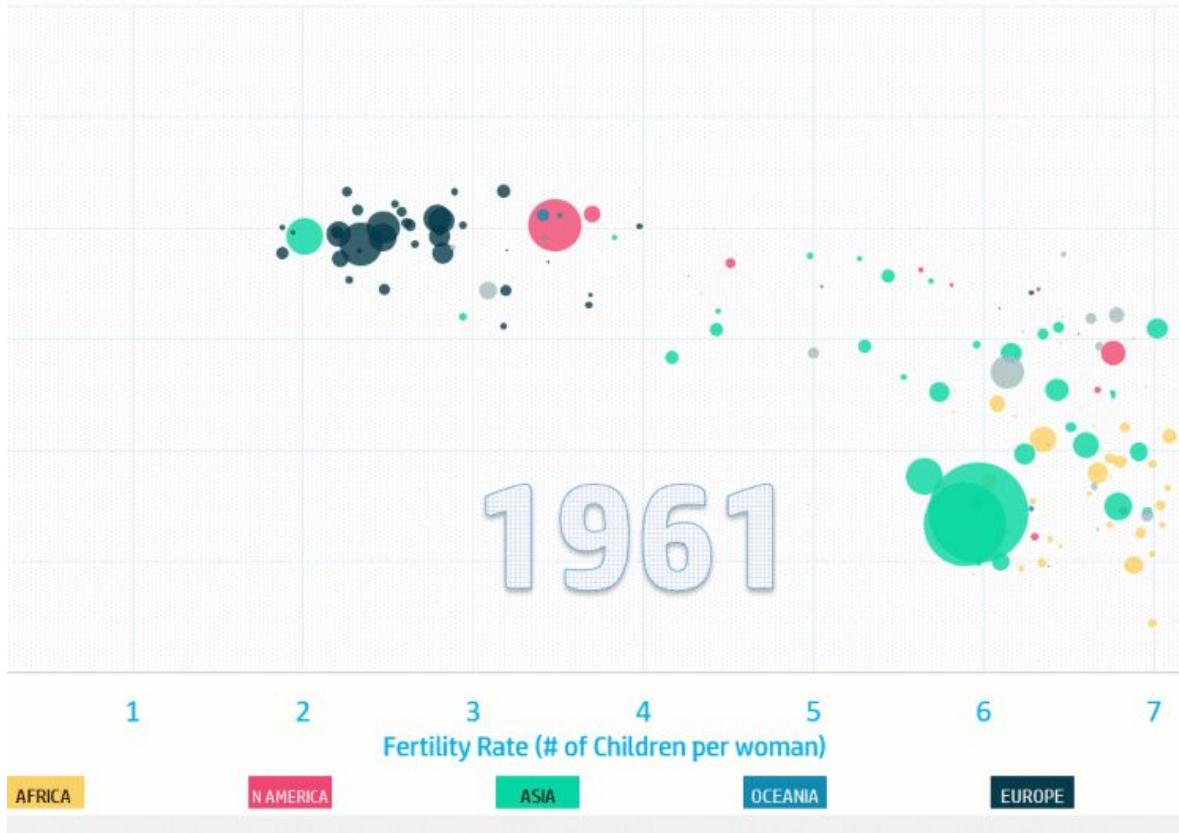


Gráfico de burbuja (ejemplo)



Código en Python

Vamos al código



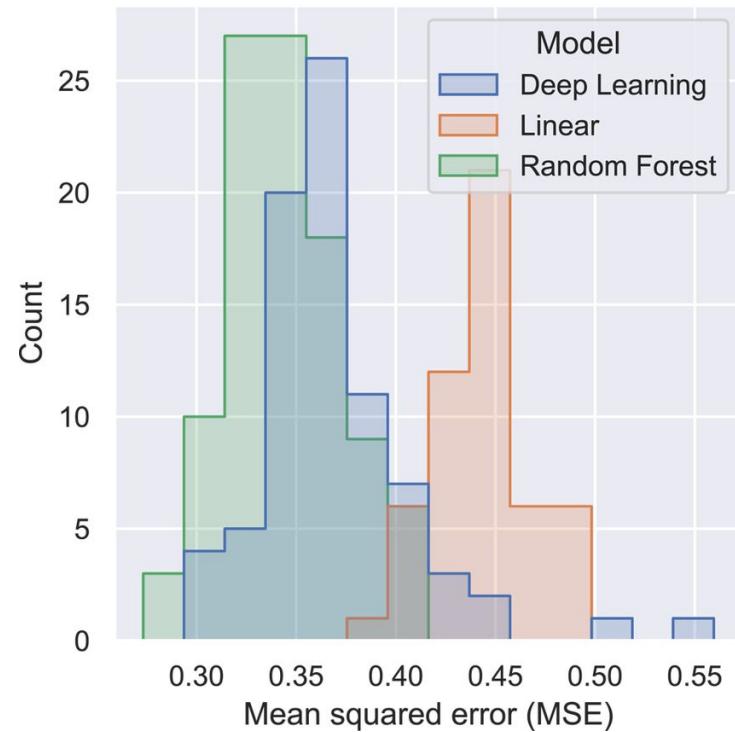
Presentación de diversos gráficos

Distribuciones

Histograma, gráfico de caja, violín,
de torta y mapa de calor

Histograma

- El término histograma fue acuñado en 1891 por el matemático y estadístico Karl Pearson.
- Gráfico muy usado en **estadística**, especialmente análisis de datos exploratorio.
- Permite conocer la **distribución** de los valores de un atributo numérico, su rango y **frecuencia** de un atributo categórico.
- Ejemplo: la distribución del *error* entre 3 modelos distintos.



Histograma

- Permite identificar rápidamente los valores **más comunes** (Ej.: edad más común).
- También permite identificar valores menos comunes y datos extraños (**outliers**).
- Similar a un gráfico de barras, pero diferente en cuanto a la **obtención de datos**.
- Las cantidad de barras, en lugar de corresponder a categorías fijas, corresponden a **rangos de valores** calculados automáticamente.

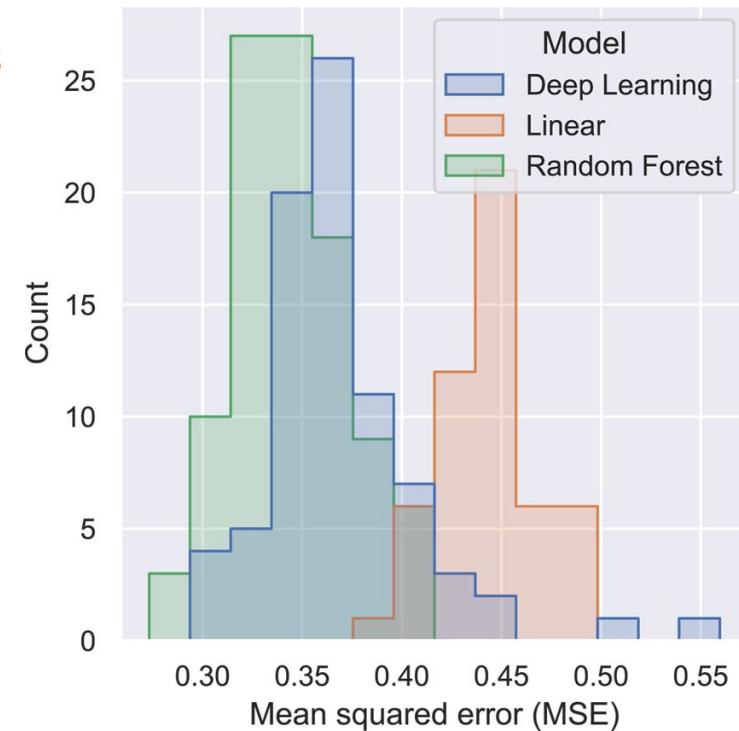


Gráfico de caja

- En inglés se denomina ***box plot***.
- Permite visualizar la **distribución de datos de una variable numérica** (cuantitativa continua), enfatizando medidas estadísticas como la mediana, cuartiles y *outliers*.
- La caja central está delimitada por el cuartil Q1 (percentil 25) y el cuartil Q3 (percentil 75). La línea dentro de la caja indica la mediana (percentil 50).

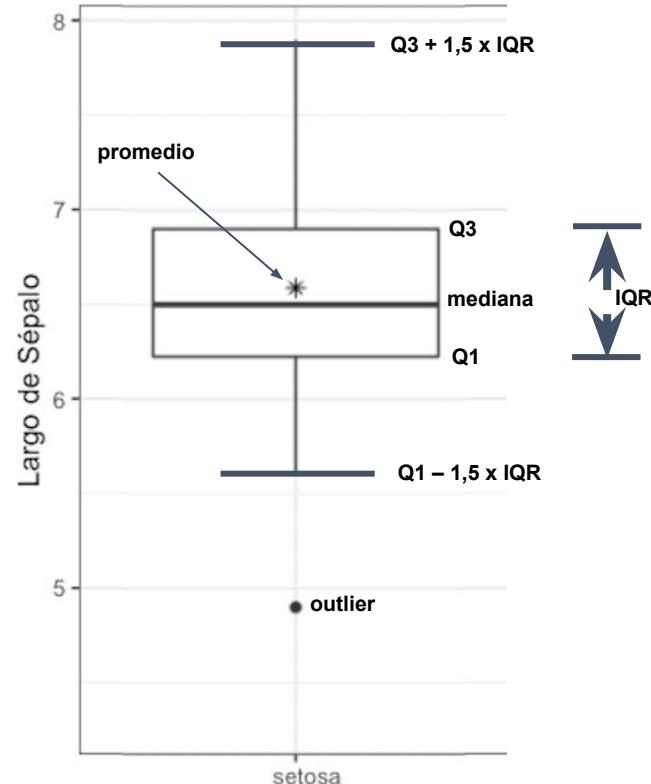


Gráfico de caja

- Se puede utilizar una marca adicional (un asterisco, por ejemplo) para representar el **promedio**.
- Se agregan dos líneas adicionales fuera de la caja que indican unos **límites mínimos y máximos esperados**, equivalentes a:
 - $\min = Q1 - 1,5 \times IQR$
 - $\max = Q3 + 1,5 \times IQR$
 - IQR significa "rango intercuartil" = $Q3 - Q1$.
- Fuera de los límites mínimo y máximo, se encuentran los **"outliers"**.

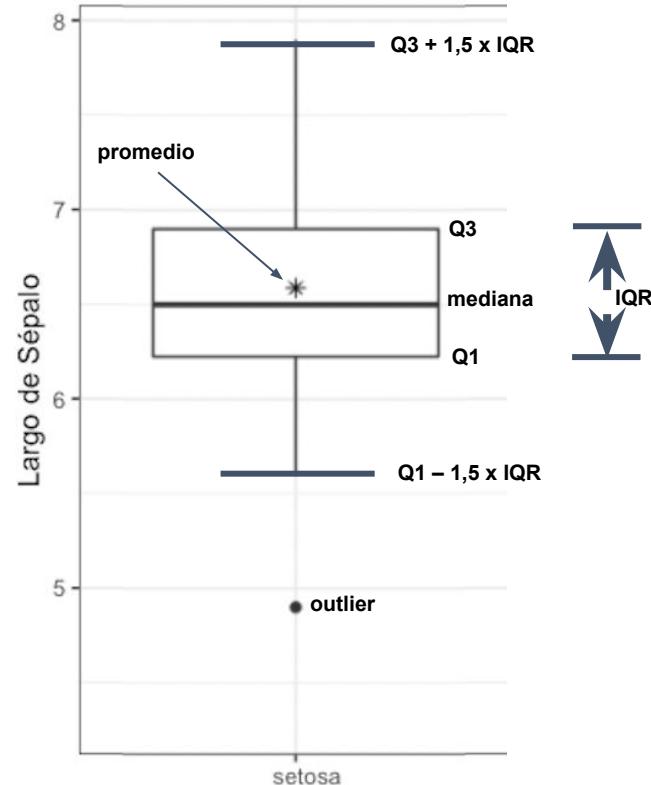


Gráfico de caja

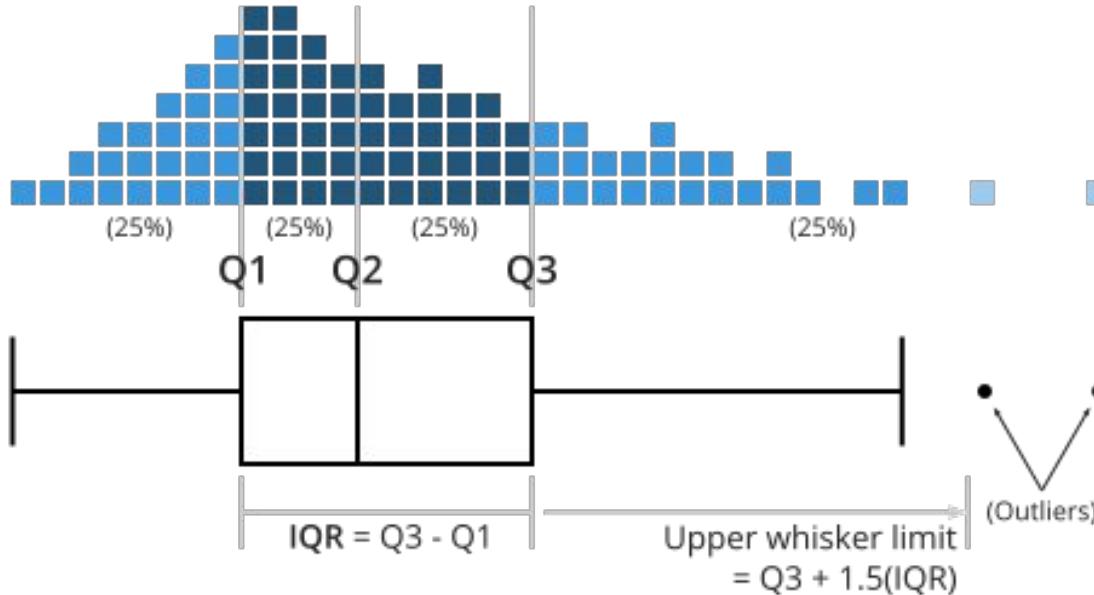


Gráfico de caja (ejemplo)

- Permite comparar fácilmente la distribución de valores de *accuracy* entre diferentes modelos para una tarea de *Machine Learning*.

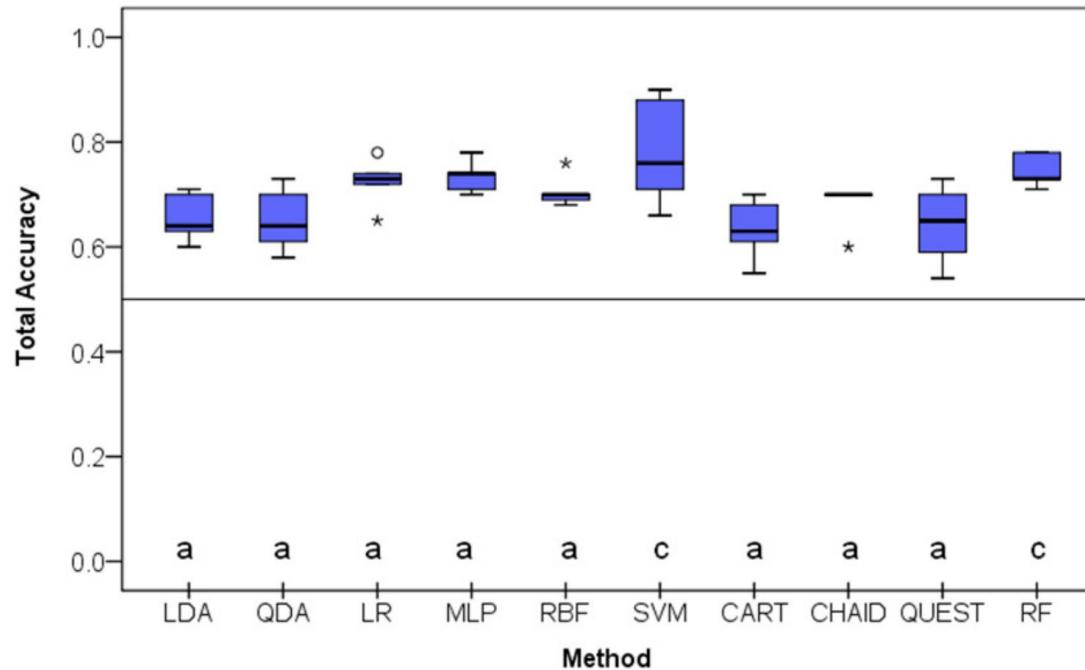


Gráfico de violín

- Combinación del gráfico de un gráfico de densidad y un gráfico de caja
- Permite visualizar **distribución estadística y densidad** de un atributo numérico.
- Paper donde fue propuesto: [Violin Plots: A Box Plot-Density Trace Synergism](#)

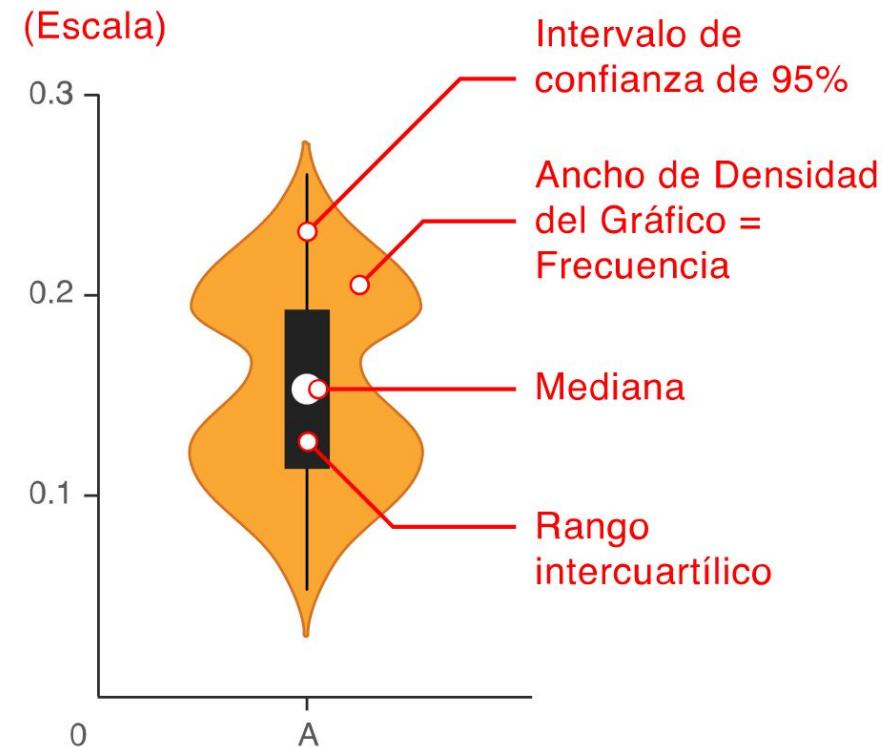


Gráfico de violín (ejemplo)

2018 daily avg temperatures by month (°F)

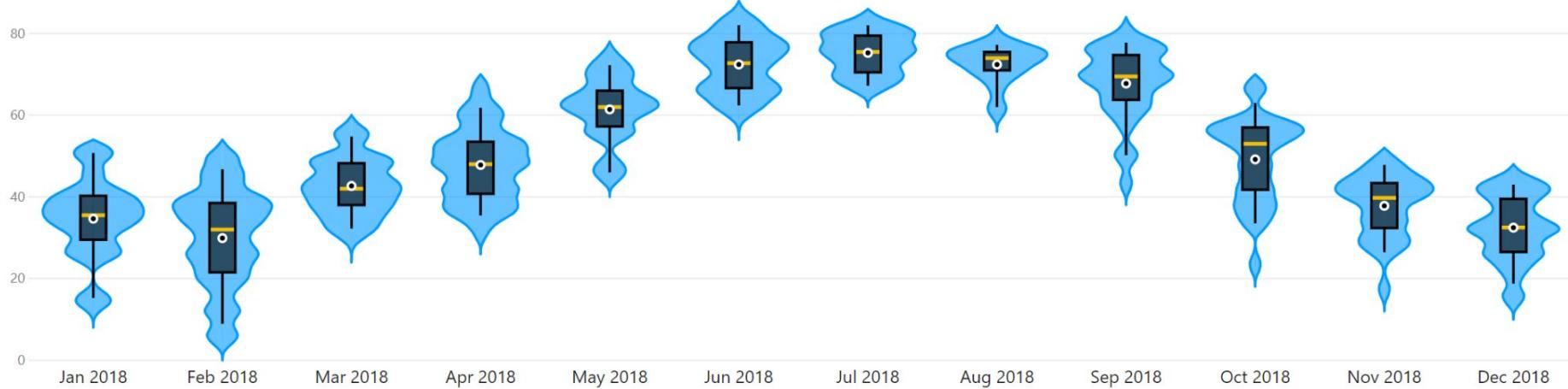
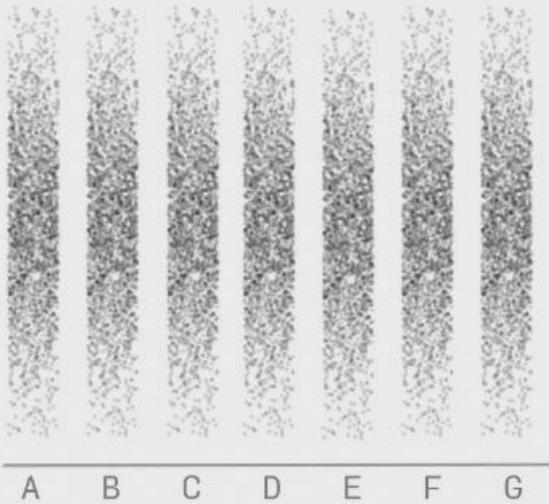
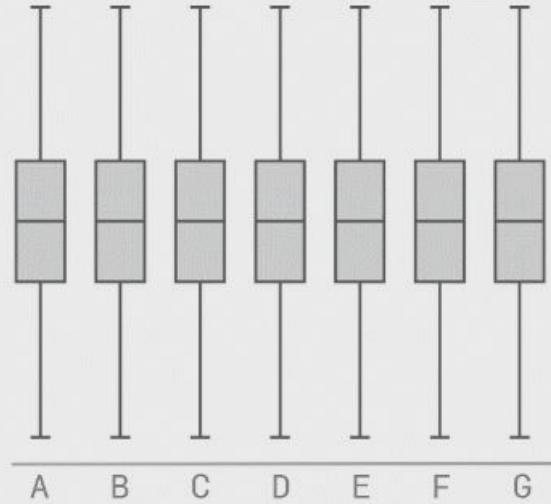


Gráfico de violín vs Gráfico de Caja

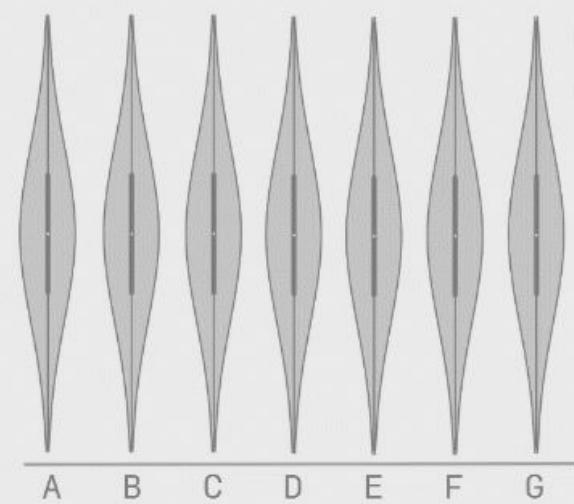
Raw Data



Box-plot of the Data



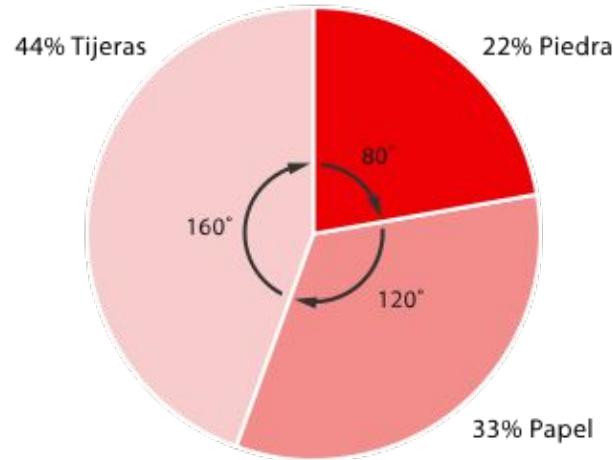
Violin-plot of the Data



*gif

Gráfico de torta

- Se utiliza una disposición circular para visualizar los datos, y se utiliza el **ángulo** para codificar la información. A mayor ángulo, el área es mayor y se entiende que el valor es más grande.
- Se utiliza principalmente para mostrar **proporciones** dentro de un total



Datos			
Piedra	Papel	Tijeras	TOTAL
2	3	4	9
Para calcular porcentajes			
2/9=22%	3/9=33%	4/9=44%	100%
Grados para cada «porción de tarta»			
(2/9) x 360 = 80°	(3/9) x 360 = 120°	(4/9) x 360 = 160°	360°

Gráfico de torta (mini debate)

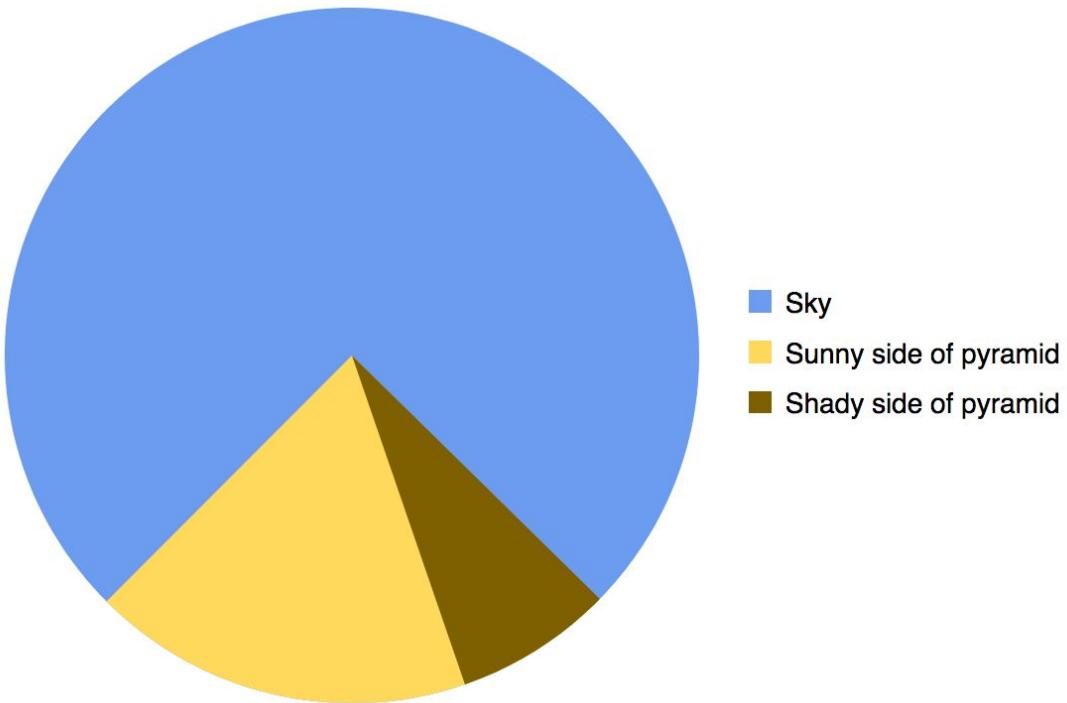


Gráfico de torta (mal ejemplo)

- Asegurar que siempre sea para visualizar una parte del total, es decir que los porcentajes suman 100%.
- No poner tantas categorías.

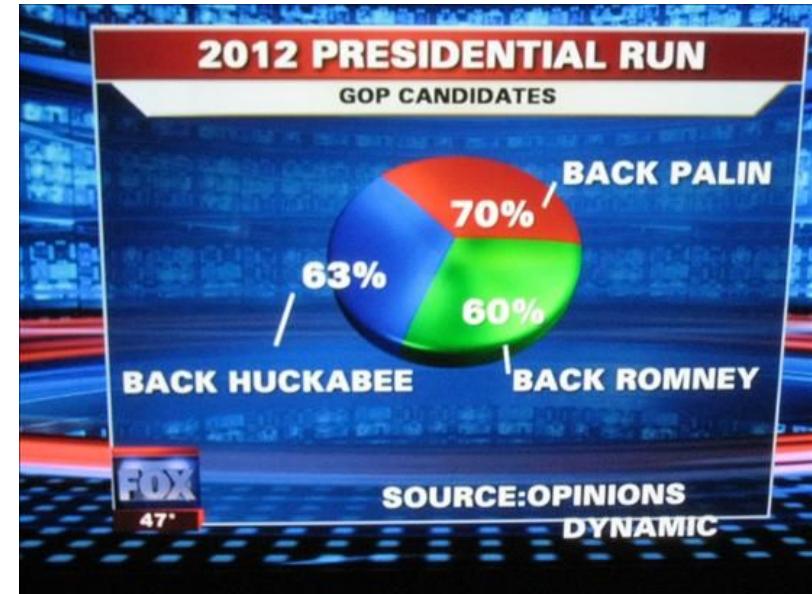
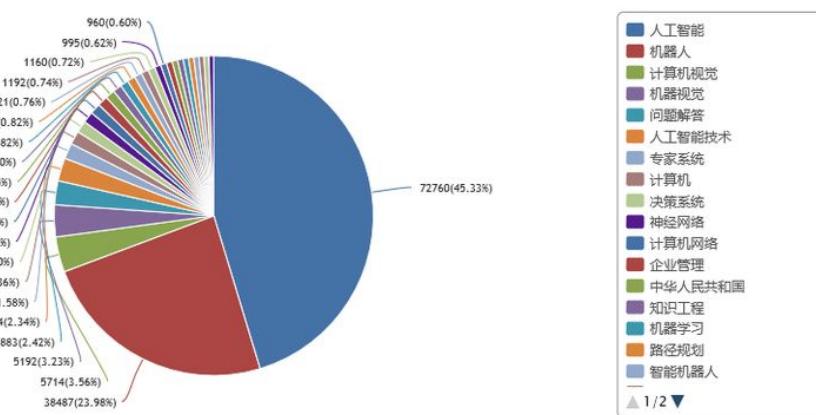
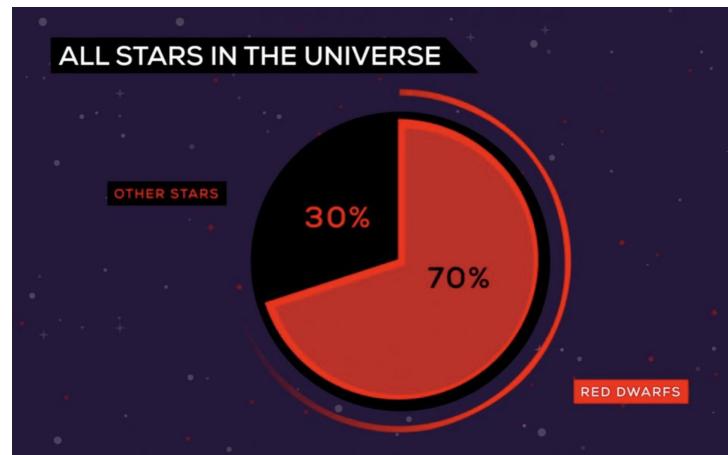


Gráfico de torta (buen ejemplo)

- Podrían existir situaciones exitosas con este gráfico.
- Por ejemplo, cuando son pocas categorías o para complementar otra visualización.

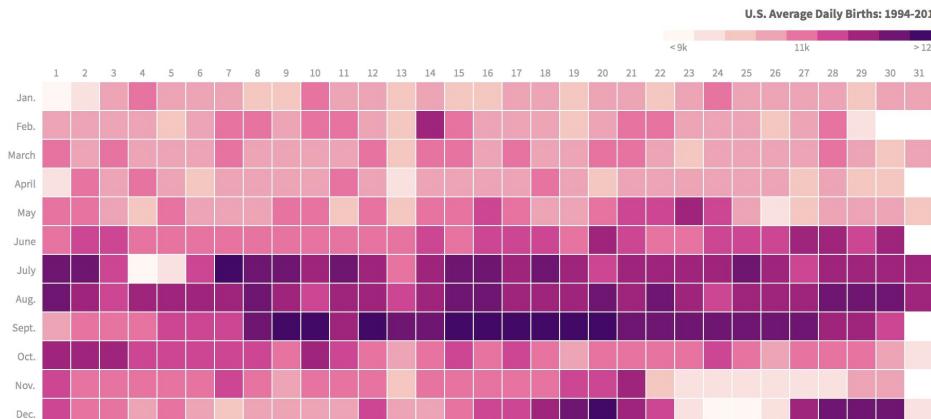


Mapa de calor

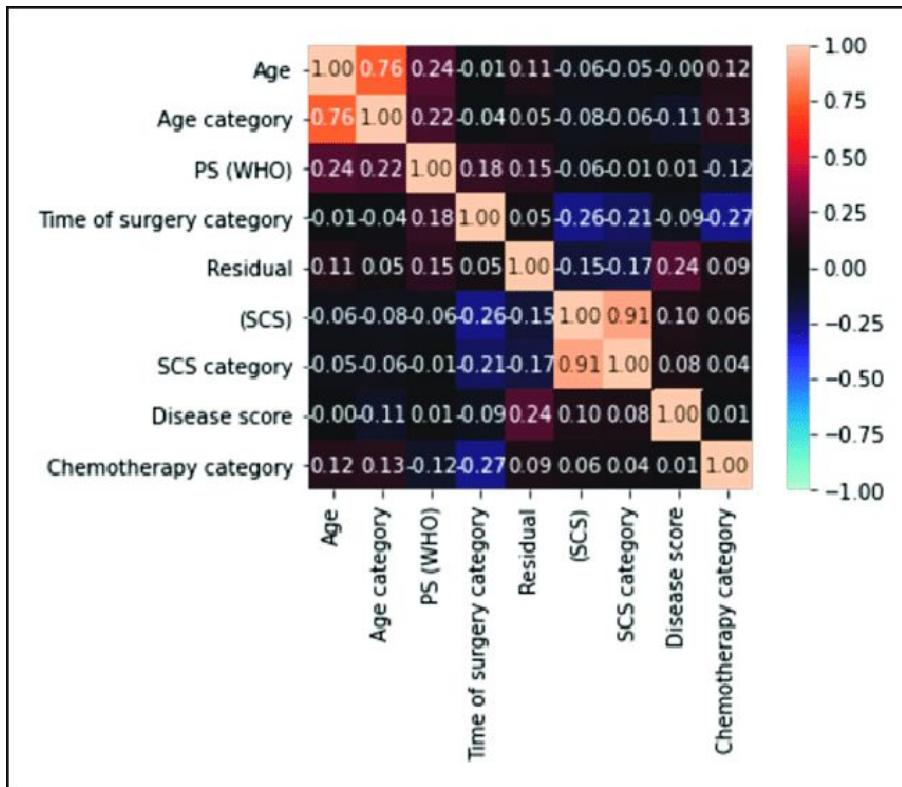
- Cada celda de la matriz codifica un **tercer atributo**.
- Este tercer atributo debe ser cuantitativo secuencial o divergente.
- Permite ver la distribución del tercer atributo a lo largo de las combinaciones de las filas y columnas.
- Eficiente en espacio.

How Popular Is Your Birthday?

Two decades of American birthdays, averaged by month and day.



Mapa de calor (ejemplo)



Percentage of men's English league games ending in a given score, for tiers 1-4, 1888 through 2013-14 season

		FINAL VISITOR SCORE								
		0	1	2	3	4	5	6	7+	
FINAL HOME SCORE	0	7.2%	6.3%	3.4%	1.4%	0.4%	0.1%	<0.1%	<0.1%	
	1	9.8%	11.6%	5.6%	2.3%	0.7%	0.2%	0.1%	<0.1%	
2	8.1%	8.9%	5.2%	1.8%	0.6%	0.2%	<0.1%	<0.1%		
3	4.8%	5.2%	2.8%	1.1%	0.3%	0.1%	<0.1%	<0.1%		
4	2.3%	2.5%	1.4%	0.5%	0.2%	<0.1%	<0.1%	<0.1%		
5	1.0%	1.1%	0.6%	0.2%	0.1%	<0.1%	<0.1%	<0.1%		
6	0.4%	0.4%	0.2%	0.1%	<0.1%	<0.1%	<0.1%	0.0%		
7+	0.2%	0.2%	0.1%	<0.1%	<0.1%	<0.1%	<0.1%	0.0%		

Código en Python

Vamos al código



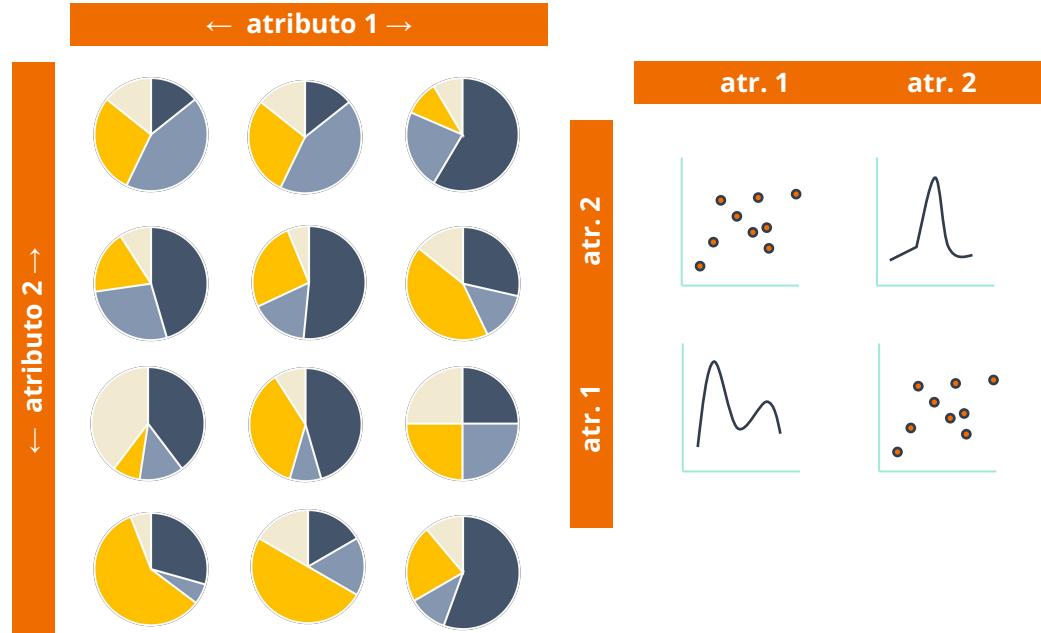
Presentación de diversos gráficos

Múltiples gráficos

Matriz de gráfico y pequeños
múltiples

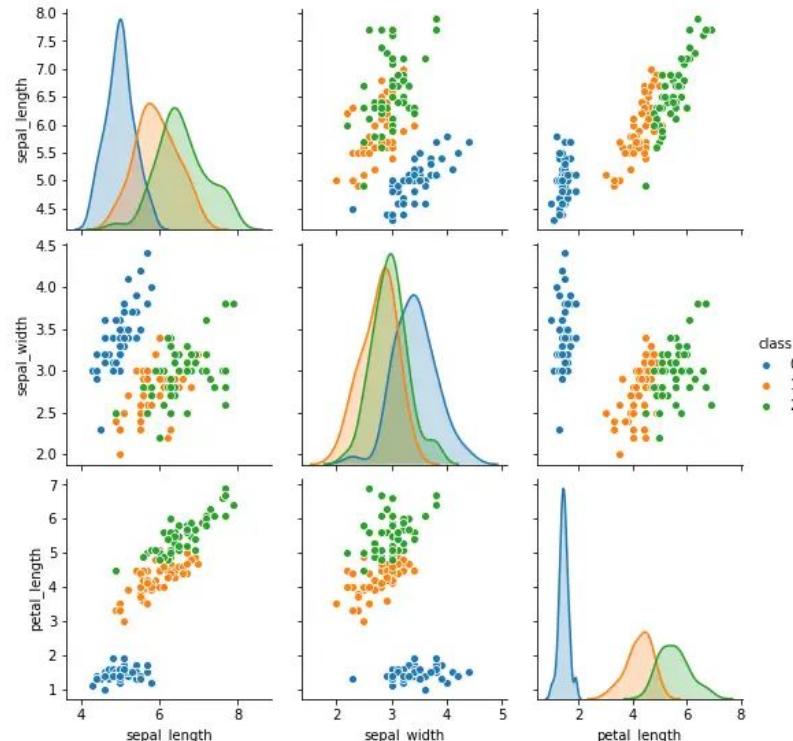
Matriz de gráficos

- Cada celda de la matriz es otra codificación por sí sola.
- Cada codificación no necesariamente debe ser la misma.
- *Datasets* con múltiples atributos suelen ser compatibles.
- Forma eficiente de mostrar la relación entre variadas combinaciones de atributos.



Matriz de *scatterplots*

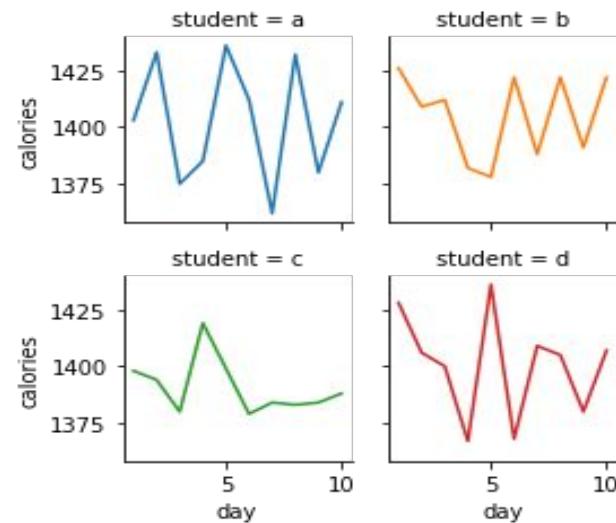
- Un caso común de matriz de gráficos.
- Un *scatterplot* permite apreciar una posible relación entre dos atributos cuantitativos.
- Un *dataset* con más de dos atributos cuantitativos es muy compatible con esta codificación.
- En la diagonal a veces se incluye otro tipo de visualización



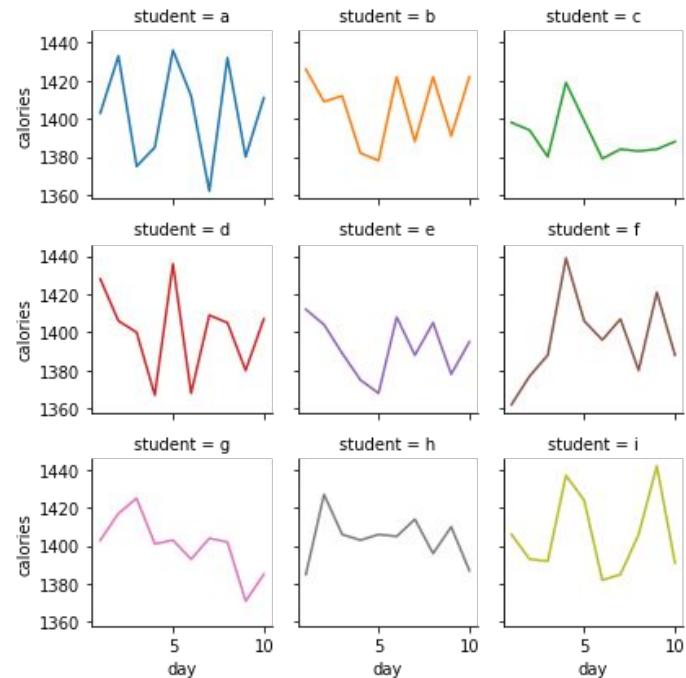
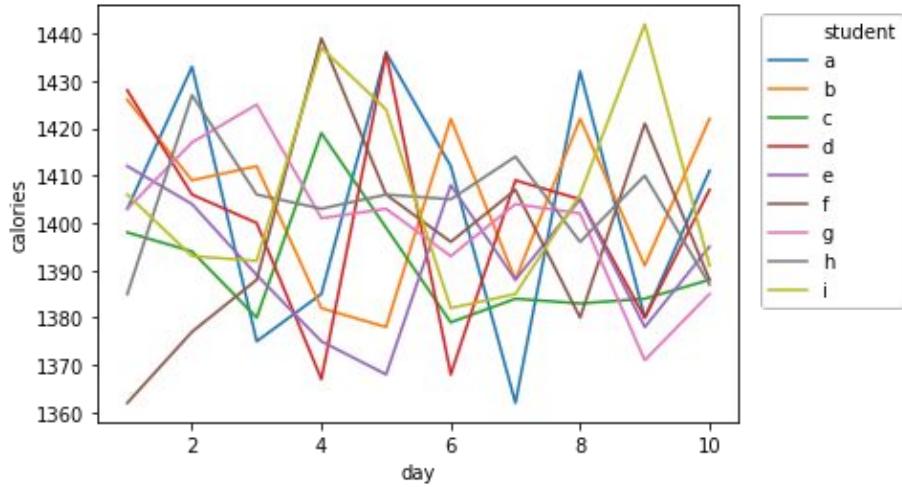
Pequeños múltiples o *Small Multiples*

- Un caso **particular** de matriz de gráficos.
- Cada celda presenta la **misma codificación**.
- Todos los gráficos comparten los **mismos atributos**.
- Cada gráfico posee sólo **1 categoría** posible respecto al atributo definido.
- Utilizado comúnmente cuando hay un atributo con **demasiadas categorías**.

day	calories	student
1	1403	a
5	1378	b
6	1422	b
10	1422	b
6	1379	c



Pequeños múltiples o *Small Multiples*



Pequeños múltiples o *Small Multiples*

FOOD CONSUMED BY MONTH
IN 2010



FOOD ORDER: CHICKEN BEEF PORK TURKEY HAM SEAFOOD PIZZA PASTA SOUP SALAD FRIES SUSHI VEGGIE BURGER RICE GREEK YOGURT FRUIT VEGGIES

Código en Python

Vamos al código





Visualización de Información y Analítica Visual

Hernán Valdivieso López (hfvaldivieso@uc.cl)
