Visualización de datos extraídos de los medios sociales

PID_00278308

Jordi Morales i Gras

Tiempo mínimo de dedicación recomendado: 3 horas

Universitat Oberta de Catalunya

Jordi Morales i Gras

Doctor en Sociología por la Universidad del País Vasco; profesor de Análisis de redes, Aprendizaje automático y Datos masivos, y socio director de Network Outsight, empresa especializada en el análisis sociológico de macrodatos.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por la profesora: Andrea Rosales Climent

Primera edición: septiembre 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Jordi Morales i Gras
Producción: FUOC

Índice

Introducción

En los dos módulos anteriores hemos visto aspectos clave de la minería de datos provenientes de los medios sociales, el propio concepto de *big data* o *macrodatos* y su cadena de valor, y hemos introducido las técnicas de análisis principales, como el *machine learning* (aprendizaje automático) y la programación del lenguaje natural. Este último módulo lo dedicaremos a uno de los aspectos más importantes del análisis de datos y una de las herramientas principales que facilitan la interpretación y la toma de decisiones a partir de datos: la visualización de datos.

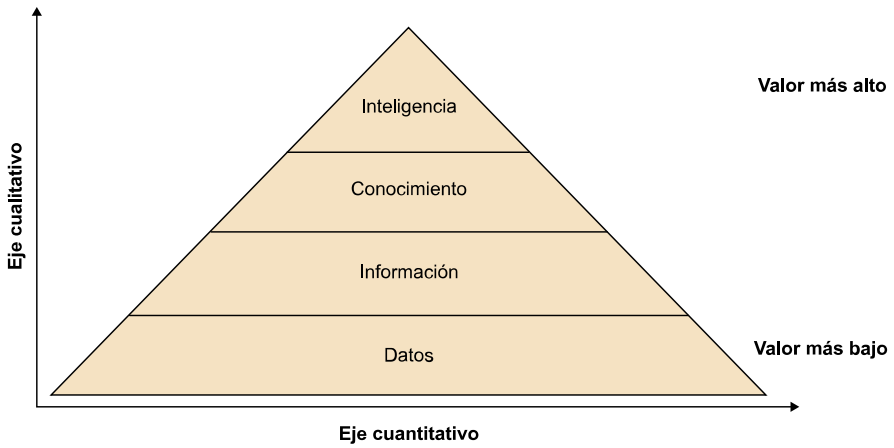
Revisaremos los principios básicos de toda visualización funcional y veremos las técnicas más utilizadas para representar gráficamente las propiedades de las diferentes variables y sus relaciones. También veremos algunas técnicas de visualización menos habituales y más complejas, especialmente indicadas para representar ciertos elementos presentes en las bases de macrodatos y los datos provenientes de los medios sociales. No solamente veremos una serie de buenas prácticas de visualización, sino que también veremos algunas de no tan buenas o directamente malas y que es importante evitar en una visualización funcional. También veremos las diferentes herramientas de que disponemos actualmente para llevar a cabo proyectos de visualización de datos, con especial énfasis en las herramientas de tipo interactivo y que permiten una exploración por parte del lector o visualizador.

Finalmente, y para cerrar el ciclo que empezamos en el primer módulo, veremos cuáles son los aspectos clave a considerar en un proyecto de investigación a partir de macrodatos de los medios sociales. Para hacerlo, deberemos tener en cuenta que cualquier investigación exitosa parte de la explotación de un ciclo virtuoso y que requiere cierto grado de retroalimentación: entre unas preguntas de investigación muy formuladas (por ejemplo, relativas a los contenidos de las conversaciones, a su calidad o a sus liderazgos), una metodología apropiada y muy desarrollada, una visualización de datos clarificadora y una interpretación de los resultados ajustada a los datos y al universo de posibilidades que nos han abierto las mismas preguntas de investigación y la metodología aplicada.

1. Fundamentos de la visualización de datos

En el primer módulo introdujimos la pirámide informacional (figura 1) como modelo teórico que permite explicar la transformación de los datos en inteligencia, mediante una serie de operaciones orientadas a aportar o añadir valor (Gloria Ponjuán Dante, 1998). Junto con el análisis de datos, la visualización de datos es una de las estrategias fundamentales para llevar a cabo el primer y el segundo paso del proceso: la transformación de los datos en información, y la información en conocimiento. En un escenario orientado a la creación de inteligencia, la visualización de datos permite reducir la incertidumbre sobre el tipo de decisiones que finalmente derivarán del análisis y la interpretación de los mismos datos (Imogen Robinson, 2016).

Figura 1. La pirámide informacional



Fuente: elaboración propia a partir de Gloria Ponjuán Dante (1998).

El objetivo de la visualización de datos es facilitar la comprensión de un fenómeno particular y su relación con una serie de aspectos o procesos que son de interés (Alexandru Telea, 2014). En el contexto de una visualización de datos, comprender un fenómeno puede significar dos cuestiones diferentes: responder a una pregunta específica sobre un problema determinado (mediante un razonamiento deductivo), o descubrir facetas o dimensiones desconocidas sobre el mismo fenómeno (mediante un razonamiento inductivo).

1) En el primer caso (visualización orientada a la respuesta a preguntas o al contraste de hipótesis), partiremos de una hipótesis o pregunta explícita sobre el fenómeno o sobre su relación con alguna o algunas variables. Estas hipótesis o preguntas pueden incluir tanto variables cualitativas (describen las calidades, circunstancias o características de un objeto o persona haciendo uso de categorías; por ejemplo: creo que mi comercio electrónico lo visitan más

mujeres que hombres) como cuantitativas (describen las características de un objeto o persona haciendo uso de números; por ejemplo: quiero saber cuál es la tasa de conversión en mi comercio electrónico).

2) En el segundo caso (visualización orientada a la exploración empírica), intentaremos obtener una panorámica general sobre los datos sin partir de una hipótesis o pregunta explícita, y descubrir alguna propiedad desconocida, y tal vez sorpresiva, sobre el fenómeno que analizamos (por ejemplo: he descubierto que el consumo de pañales se correlaciona con el consumo de *sushi* en mi comercio electrónico).

Desde el punto de vista del diseño, y tanto si nos encontramos en un escenario de respuesta a preguntas concretas como en un escenario de exploración visual, debemos tener en cuenta que una visualización de datos es una forma de comunicación y, como tal, cuenta con un contenido difundido por un canal que se ajusta a un código determinado y con un usuario que lo interpreta, lo recrea y lo puede poner en circulación por medio de otros canales. Esto se traduce en seis principios aplicables a la elaboración de cualquier gráfica (Manuel Lima, 2017):

1) Honestidad en la representación de datos. La claridad y la transparencia son más importantes que los aspectos estéticos de una visualización. Hay que proporcionar todos los elementos necesarios porque una visualización permita comprender un fenómeno: etiquetas y esos claros, leyendas e informaciones emergentes (indicadores de función), fuentes de datos explicitados, etc.

2) Facilidad de lectura de las gráficas. Hay que partir del código que conocen los usuarios. Es importante diseñar visualizaciones teniendo en cuenta la manera como los usuarios están acostumbrados a leer la información gráfica. En el caso de las visualizaciones interactivas, es importante que el lector pueda indagar y descubrir las correlaciones y asociaciones en los datos por sí mismo: selecciones interactivas, *zoom* interactivo, filtros interactivos, etc.

3) La experiencia del usuario en el centro. Cuanto más agradable sea una visualización, más poderosa será la capacidad informativa y de generar comprensión sobre el fenómeno o los fenómenos representados. En las visualizaciones de tipo interactivo, aspectos como la rapidez o la narrativa incrustada en los datos son tan importantes como la excelencia gráfica.

4) Claridad en el enfoque gráfico. Cualquier elemento de una gráfica (por ejemplo, los colores, el tamaño de los elementos, su distribución, etc.) tiene que estar orientado a la comprensión de las jerarquías, las tipologías y las relaciones representadas. Es importante evitar cualquier estímulo innecesario que pueda distraer al lector o visualizador de la comprensión de lo que es esencial en una visualización.

5) Sensibilidad al dispositivo y escalabilidad. Hay que atender, también, el canal de difusión del mensaje visual, sus particularidades y posibilidades. Del mismo modo que es importante tener en cuenta aspectos como el color o el tamaño en una gráfica impresa, es importante tener en cuenta la accesibilidad de los diferentes dispositivos que accederán a una visualización interactiva: paletas de color, filtros, resoluciones de pantalla, etc.

6) Estructura y consistencia gráfica. Finalmente, es importante dotar la visualización o el conjunto de visualizaciones de una consistencia y coherencia gráficas general: el tamaño de los elementos, la paleta de colores, la tipografía, la orientación de los ejes, etc. Cuanto más intuitiva y familiar resulte una visualización, más fácil será que esta permita la comprensión del fenómeno representado.

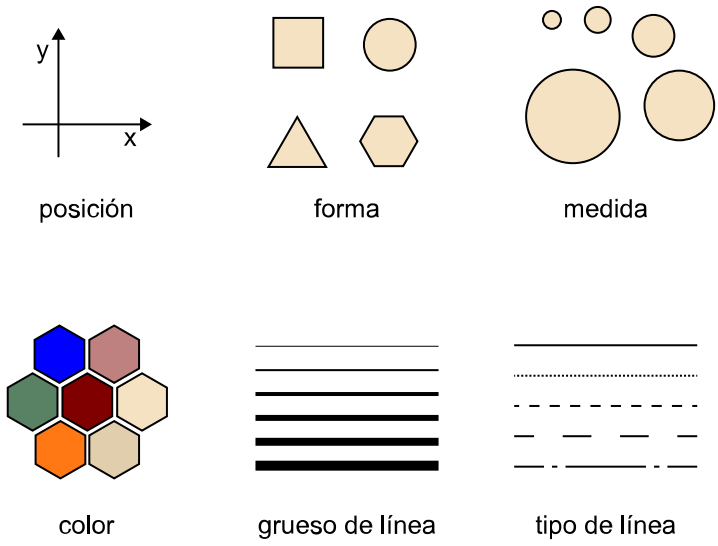
Desde la perspectiva enunciada, un **diseño funcional** puede definirse como una representación visual de datos honesta, fácil de leer, agradable, clara, sensible al medio y gráficamente consistente.

Como veremos a continuación, no hay una sola manera de aplicar correctamente estos seis principios y hará falta que conozcamos muy bien tanto nuestro mensaje como nuestra audiencia para generar las mejores condiciones posibles que permitan al lector comprender el fenómeno representado.

2. Técnicas de visualización de datos

Cualquier visualización de datos tiene como objetivo generar condiciones para que el lector o visualizador pueda comprender un fenómeno, ya sea respondiendo a alguna pregunta clave o descubriendo aspectos desconocidos. Actualmente, hay una gran cantidad de técnicas para visualizar datos que representan propiedades y relaciones entre diferentes tipos de variables (por ejemplo, cifras, categorías, fechas o textos) mediante una serie de elementos gráficos como son posiciones, formas, tamaños, colores y líneas de diferentes gruesos y tipos (figura 2; Claus O. Wilke, 2019).

Figura 2. Los elementos gráficos de la visualización de datos

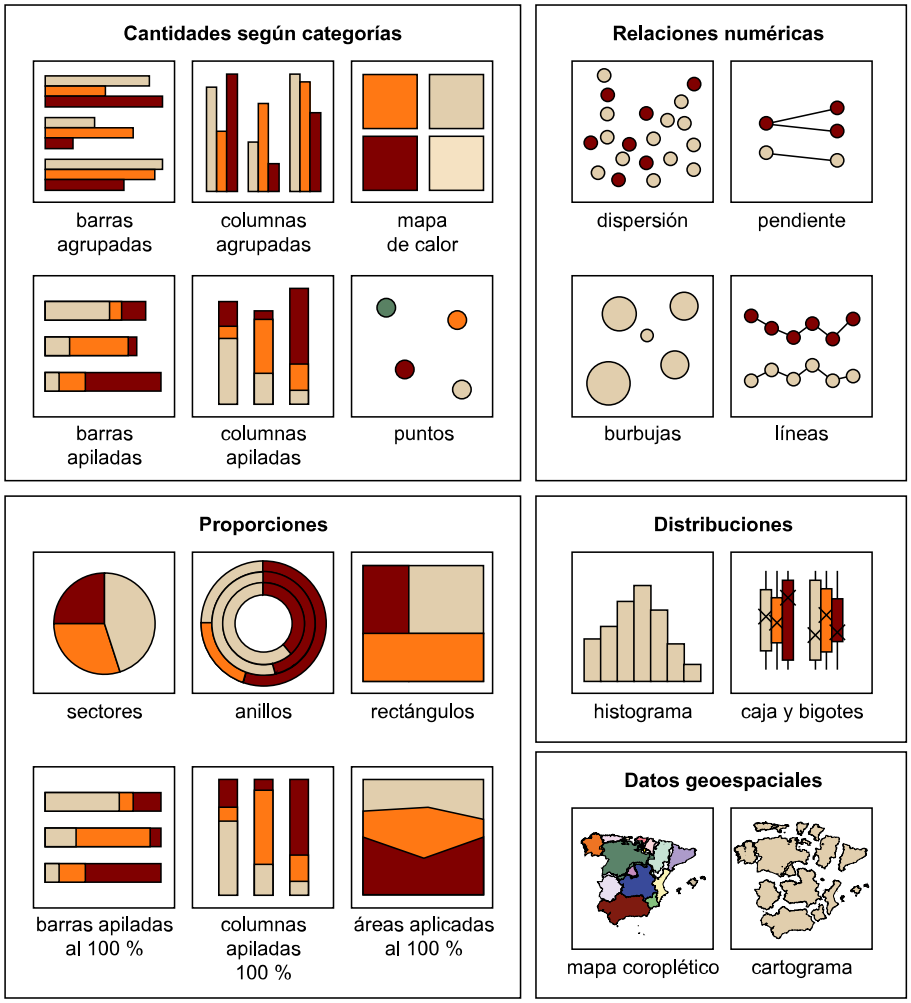


Fuente: elaboración propia a partir de Claus O. Wilke (2019).

Las diferentes visualizaciones de datos de que disponemos utilizan y combinan los elementos anteriores de diferentes maneras para representar diferentes aspectos clave de los datos. En función del tipo de datos de que se dispone (variables numéricas, categóricas, temporales, geoespaciales, etc.) y del tipo de comprensión que se quiere facilitar[1], el analista ha de escoger la visualización más funcional para un conjunto de datos. Las técnicas que veremos a continuación, y que resumimos en la imagen siguiente (figura 3), constituyen algunas de las técnicas de visualización de datos más habituales.

(1) Ya sea comparar la misma variable numérica para dos categorías diferentes, observar la correlación entre dos variables numéricas, observar la distribución interna, etc.

Figura 3. Las visualizaciones de datos más habituales

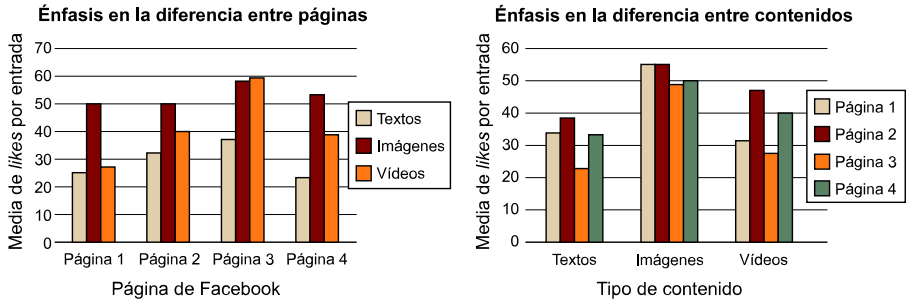


Fuente: elaboración propia (datos aleatorios).

2.1. Visualización de cantidades según categorías

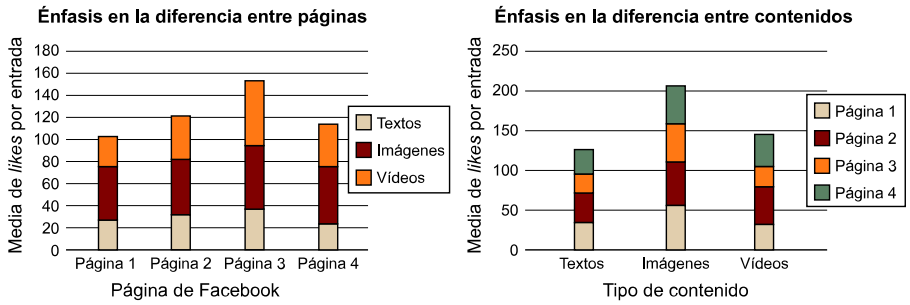
Uno de los escenarios más habituales en la visualización de datos es la representación de magnitudes (variables cuantitativas) por medio de categorías o clases (variables cualitativas). En términos generales, querremos comparar casos e identificar su orden o jerarquía. Por ejemplo, podemos estar interesados en desglosar el número de *likes* a los contenidos de una o más páginas de Facebook en función de si se trata de entradas con texto, imágenes o vídeos: el número de *likes* es la variable cuantitativa, y tanto el tipo de contenido como el nombre de la página son las variables cualitativas. En este tipo de casos, lo más habitual y recomendable será utilizar una gráfica de barras o de columnas, y representarlas de manera agrupada en caso de visualizar simultáneamente más de una variable cualitativa. En función de la disposición de las variables en los ejes o en la leyenda de colores, se enfatizará una propiedad de los datos u otra (figura 4). Si lo que se quiere es enfatizar todavía más una de las dos variables cualitativas, puede optarse por una gráfica de columnas o barras apiladas (figura 5). Estas gráficas incorporan elementos de proporcionalidad que veremos más adelante y que van más allá de la representación de magnitudes.

Figura 4. *Likes* para contenidos en diferentes páginas de Facebook



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

Figura 5. *Likes* para contenidos en diferentes páginas de Facebook

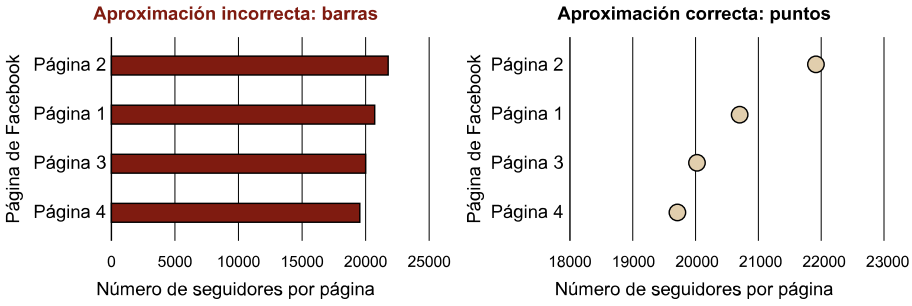


Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

Uno de los principios de las **gráficas de barras y columnas** es que la interpretación de los datos se basa en la longitud del elemento central: la barra o la columna. Ello implica que la base de la columna o de la barra siempre tiene que ser cero.

Tal como veremos más adelante, violar este principio es una de las formas de manipulación visual más extendidas, ya sea por ignorancia o por mala fe. Aun así, en algunas ocasiones, este principio puede resultar limitador para la representación de diferencias numéricas entre categorías como, por ejemplo, cuando las diferencias entre estas son más o menos pequeñas. En estas ocasiones, puede ser una buena alternativa la gráfica de puntos, puesto que la interpretación de este tipo de representación no se basa en la longitud del elemento central sino en su posición, lo que permite utilizar bases diferentes de cero, siempre que se haga de manera explícita (figura 6).

Figura 6. Número de seguidores de cuatro páginas de Facebook



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

Cuando contamos con una cantidad elevada de magnitudes a representar en función de una serie de variables categóricas, una buena alternativa a las barras, las columnas y los puntos son los colores. Los **mapas de calor** permiten llevar a cabo este tipo de visualizaciones.

Como en cualquier visualización de cantidades según categorías, es fundamental ordenar los casos para enfatizar la jerarquía.

Figura 7. Evolución del número de seguidores de diez usuarios de Twitter

2897	2927	2943	2957	2974	3001	3054	3118	3103	3164	3121	3104	@usuario1
2784	2822	2868	2912	2944	3015	3024	2988	2942	2932	3003	3062	@usuario2
2426	2433	2435	2483	2460	2531	2575	2526	2601	2632	2707	2729	@usuario3
2317	2337	2291	2283	2318	2355	2305	2277	2346	2394	2422	2476	@usuario4
1942	1961	2036	2096	2129	2205	2237	2295	2291	2337	2301	2329	@usuario5
1519	1582	1590	1632	1706	1764	1833	1824	1788	1767	1842	1817	@usuario6
943	948	978	1053	1059	1011	1020	1035	1016	981	967	957	@usuario7
511	537	369	689	654	707	686	668	655	730	736	782	@usuario8
484	510	498	488	462	450	519	482	457	445	444	425	@usuario9
201	245	273	259	337	358	354	341	302	323	283	273	@usuario10
Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	@usuario11

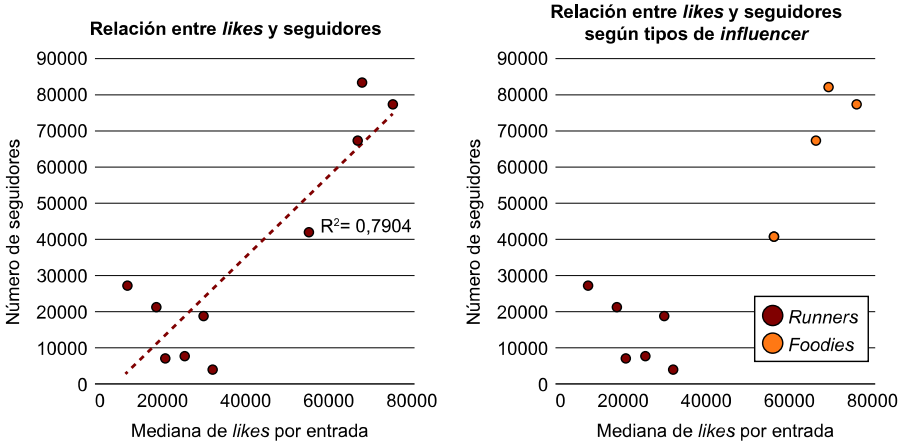
Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

2.2. Visualización de relaciones numéricas entre variables

Otro tipo de escenario de representación de datos muy común es la visualización de la relación entre dos variables numéricas. La manera más sencilla e intuitiva de visualizar una relación de este tipo es el **diagrama de dispersión**, que utiliza las coordenadas cartesianas para mostrar la posición de los casos en función de las variables.

Por ejemplo, si queremos estudiar la relación entre el número de seguidores y la media de *likes* por entrada de una serie de *influencers* de Instagram, podremos representar esta relación con un diagrama de dispersión (figura 8). Esto nos permitirá trazar la línea de tendencia y conocer el coeficiente de determinación entre las dos variables (como vimos en el módulo 2, la estadística R^2 nos informa de la capacidad de un modelo para replicar los resultados). Mediante elementos como el color también podremos representar variables cualitativas en un diagrama de dispersión, por ejemplo, para identificar que los *influencers* de *running* tienden a tener menos seguidores y *likes* que los *foodies*.

Figura 8. Relación entre *likes* y seguidores para los *influencers* de Instagram

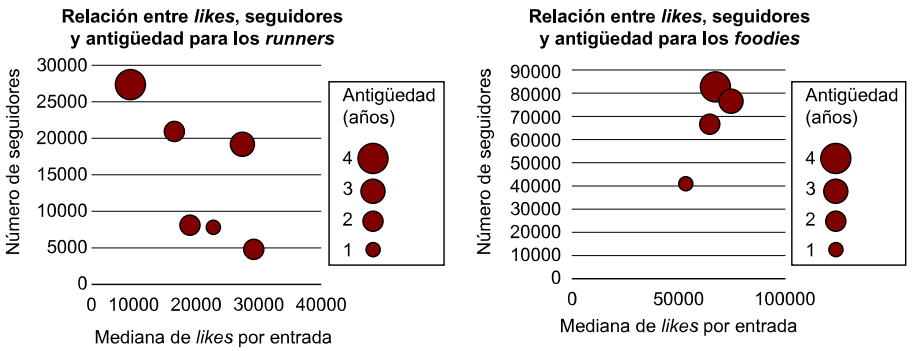


Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

En algunas ocasiones, puede ser interesante añadir alguna dimensión adicional a la visualización de datos como, por ejemplo, los años de antigüedad del usuario de Instagram en cuestión. En estos casos, una buena opción es la **gráfica de burbujas**.

Se trata de una visualización preparada para la representación de tres o más variables; aun así, es recomendable no excederse en el número de variables para no sobrecargar la visualización. Una buena estrategia puede ser crear diferentes gráficas en función de la cuarta o quinta variable, en lugar de continuar explotando recursos gráficos. Siguiendo con nuestro ejemplo, podemos generar dos gráficas para observar la importancia de la antigüedad en función del tipo de *influencer* (figura 9). Cuando representamos una gráfica de burbujas, es importante tener en cuenta que las diferencias entre las posiciones de los ejes adquieren más énfasis que las diferencias entre el tamaño de las burbujas, que son más difíciles de percibir visualmente.

Figura 9. Relación entre *likes*, seguidores y antigüedad por tipo de *influencer*

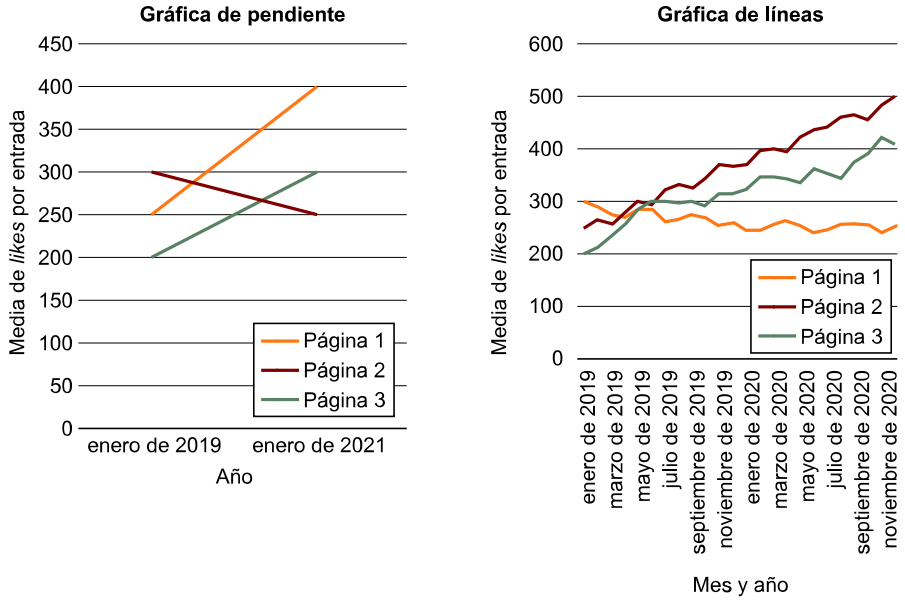


Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

Las coordenadas cartesianas también pueden utilizarse para representar la relación entre una variable numérica y una serie de fechas, ordenadas cronológicamente. Este tipo de representaciones sirve para mostrar la evolución de una variable numérica en función del tiempo (por ejemplo, el número medio de *li-*

kes por entrada de una página de Facebook) y muestra varias categorías simultáneamente (por ejemplo, la misma métrica para varias páginas de Facebook), hecho que facilita el análisis comparativo. En este tipo de representaciones es clave el número de categorías cualitativas a representar –es conveniente no saturar la visualización con demasiadas categorías: cinco o seis puede ser un buen número– y la granularidad o la sensibilidad de la variable temporal: no es lo mismo representar décadas que años, o trimestres, o días, u horas, etc. Si lo que se quiere es establecer una comparación entre dos periodos temporales, lo mejor será optar por una gráfica de pendiente (figura 10); en cambio, si la granularidad es más grande, lo más conveniente será optar por una gráfica de líneas (figura 10). La gráfica de áreas es una versión de la gráfica de líneas con cierto refinamiento estético, pero que no aporta elementos adicionales. En cambio, la gráfica de áreas apiladas al 100 % sí que incorpora elementos de proporcionalidad que veremos más adelante.

Figura 10. Evolución temporal del compromiso medio en tres páginas de Facebook



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

2.3. Visualización de proporciones

Muchas veces, el aspecto clave a representar en una visualización de datos puede no ser una magnitud absoluta sino relativa, que se puede expresar con un porcentaje sobre un total de casos.

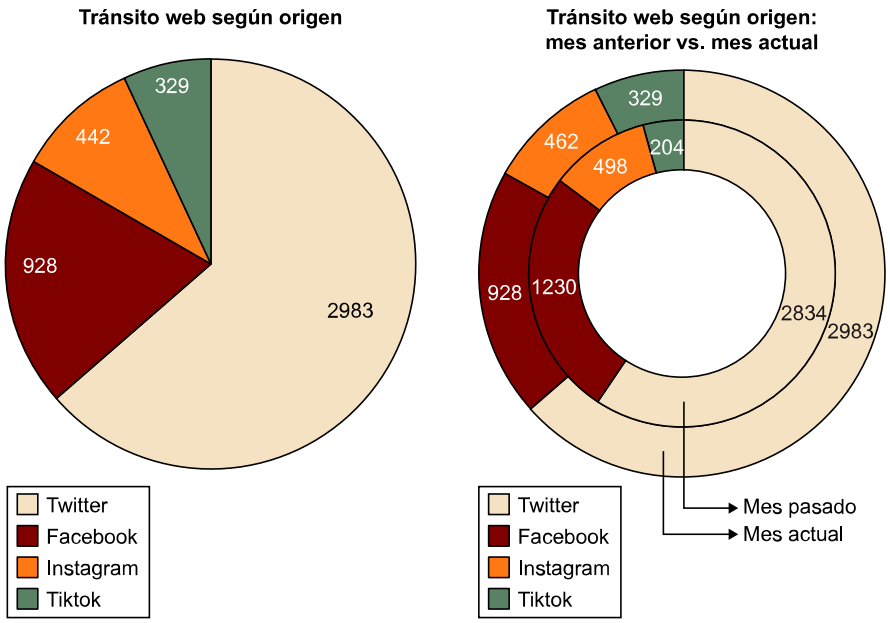
La visualización de proporciones más habitual es el **diagrama de sectores**, también llamado *gráfico circular* o *diagrama de pastel*.

Se trata de una representación que permite mostrar claramente las partes del total según una serie de categorías, y que resulta especialmente pertinente cuando estas partes responden a fracciones simples (por ejemplo, 1/2, 1/4,1/8, etc.).

La gráfica **de anillas** es una representación similar que puede incorporar más niveles de complejidad (por ejemplo, comparaciones entre periodos) mediante la representación de varias magnitudes relativas en una sola visualización (figura 11).

Actualmente, son muchas las voces críticas con las gráficas circulares o de anillas, puesto que tienen un número elevado de limitaciones que tienen que ver con su dificultad interpretativa cuando tenemos muchas fracciones, especialmente por las más pequeñas, y también cuando se trata de comparar distribuciones entre variables con magnitudes muy diferentes (Claus O. Wilke, 2019).

Figura 11. Tráfico web proveniente de redes sociales



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

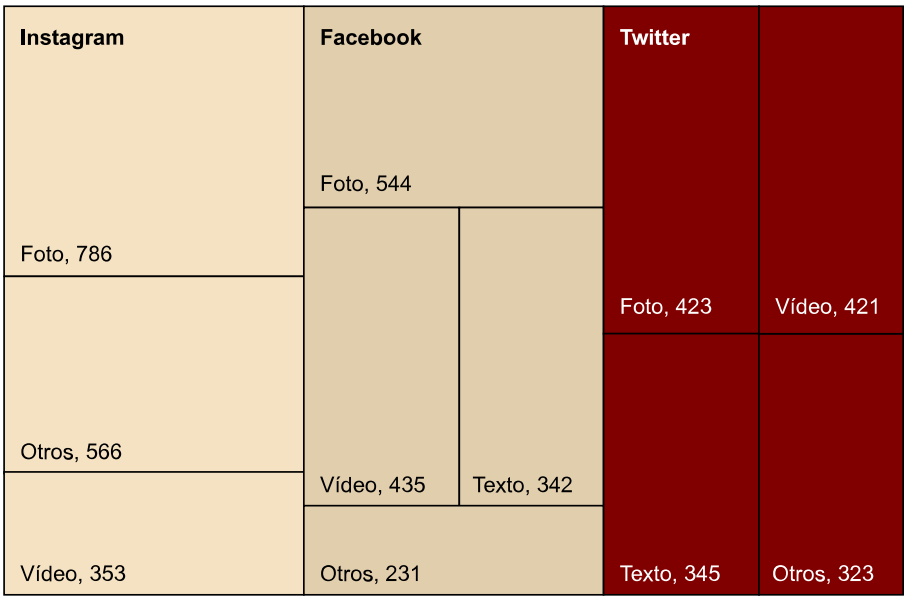
A la hora de decidir si utilizar una gráfica circular o de anillas o no, es bueno pensar en sus limitaciones, pero hace falta también considerar sus virtudes. Además de las mencionadas anteriormente, hay otra que puede ser determinante en muchas ocasiones y que tiene que ver con el principio de facilidad de lectura: hay muchos casos en que la audiencia a quien se dirige una visualización está mucho más avezada a interpretar este tipo de representaciones que sus alternativas, que implican necesariamente más complejidad. Una vez más, es el analista el que tiene que valorar todos estos aspectos y escoger una técnica u otra de representación.

Las soluciones alternativas a las gráficas circulares o de anillas basan su propuesta tanto en aspectos relativos a la forma representada (por ejemplo, rectángulos en lugar de ángulos en una circunferencia, columnas apiladas o agrupadas) como a cuestiones más complejas, vinculadas a las representaciones temporales.

Las **gráficas de rectángulos** son una de las alternativas más potentes a las gráficas circulares, puesto que permiten comparar proporciones en función de variables cualitativas mucho más fraccionadas y con diferencias más grandes de magnitud entre ellas.

Por regla general, permiten un mejor ajuste de etiquetas y facilitan visualizaciones más agradables y con más capacidad comunicativa, según una serie de estudios clásicos de visualización de datos elaborados hace casi noventa años (Frederick E. Croxton y Harold Stein, 1932). Además, se trata de visualizaciones particularmente eficientes a la hora de representar propiedades imbricadas (por ejemplo, el compromiso en función del tipo de contenido y, a la vez, en función de la red social; figura 12), lo cual se consigue jugando con los colores, las posiciones y las medidas de los distintos rectángulos.

Figura 12. Compromiso según tipo de contenido y red social: rectángulos

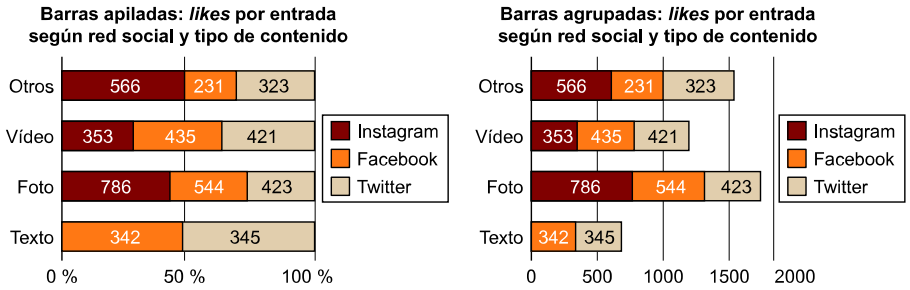


Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

Otra alternativa a las gráficas circulares son las **gráficas de columnas apiladas** al 100 %, que enfatizan diferencias relativas entre dos o más categorías, es decir, las diferencias entre el rendimiento de los diferentes contenidos en las diferentes redes sociales, mediante elementos de color (figura 13).

Esta representación puede invisibilizar aspectos como el volumen de las variables segmentadas. Para resolver este problema, es posible representar simultáneamente proporciones y magnitudes o cantidades (por ejemplo, las diferencias de compromiso entre tipo de contenido y su presencia absoluta en las diferentes redes sociales) mediante una gráfica de barras o columnas agrupadas (figura 13). Se trata de visualizaciones multivariables muy eficientes, pero que requieren cierta competencia interpretativa.

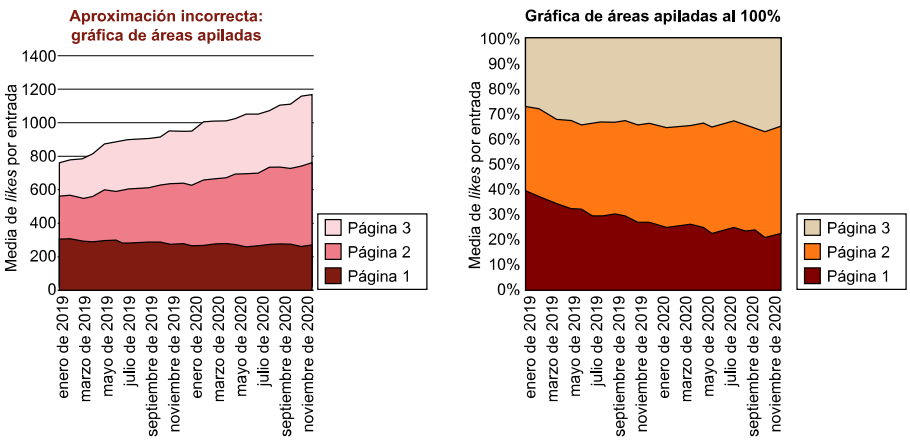
Figura 13. Compromiso según tipo de contenido y red social: columnas



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

Cuando se trata de visualizar la evolución temporal de una proporción, es posible que la gráfica de anillas múltiples también quede muy limitada si quieren representarse varios puntos de tiempo. Una buena solución a este problema es la gráfica de áreas apiladas al 100 %, que es una versión de la gráfica de líneas orientada a la proporcionalidad: representa la distribución entre varias variables cualitativas en varios puntos del tiempo (por ejemplo, la cuota de compromiso de varias páginas de un medio como Facebook). Las limitaciones de esta visualización tienen que ver con el hecho de que puede contribuir a invisibilizar magnitudes. Aun así, no es muy recomendable optar por una representación de áreas apiladas simple (figura 14), puesto que es una visualización que, por regla general, no permite observar nítidamente ni las magnitudes totales ni la relación proporcional de las variables que representa. Para representar magnitudes es preferible optar por representaciones más sencillas como, por ejemplo, una gráfica de líneas múltiples.

Figura 14. Evolución de la cuota de compromiso en tres páginas de Facebook



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

2.4. Visualización de distribuciones

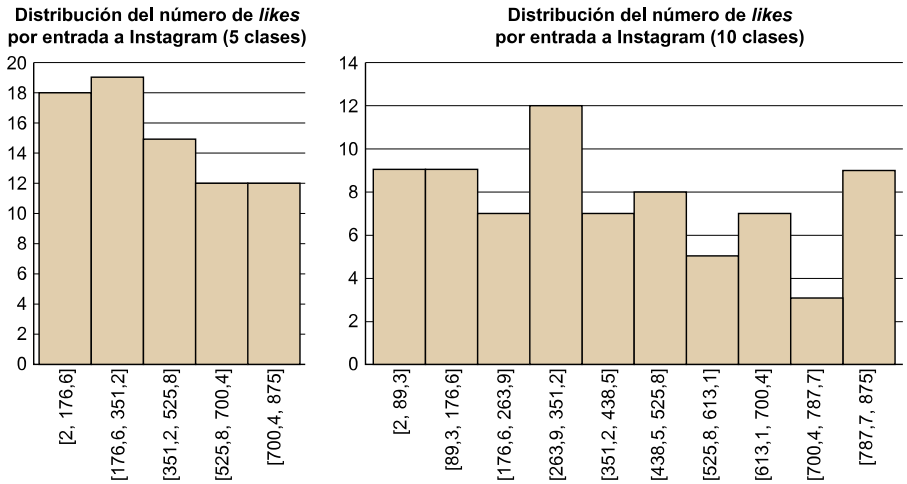
A menudo nos encontraremos con la necesidad de representar la distribución interna de una variable numérica. Se trata de poder visualizar como se distribuye la variabilidad dentro de una variable y, por lo tanto, sin tener que recurrir a representar la relación con alguna otra variable numérica o cualitativa en un diagrama de columnas o de dispersión. En los medios sociales, por ejemplo, nos puede interesar visualizar la distribución de los *likes* que reciben las entradas de un usuario para cuantificar y entender el compromiso que provoca.

Una aproximación visual a esta necesidad son los **histogramas**, que agrupan los valores que toma una variable en una serie de intervalos continuos denominados *clases*[2].

[(2)]El número de *likes* que recibe un usuario según cinco tramos: entre 0 y 199 *likes*, entre 200 y 399, entre 400 y 599, entre 600 y 799, y entre 800 y 999.

Una característica de los histogramas es que, en función del número de clases que escogemos representar, un mismo fenómeno puede tomar apariencias bastante diferentes: cuantas más clases tenga un histograma, más exhaustiva será la representación de datos y, a la vez, más compleja será su interpretación. En un histograma, el eje vertical indica el número de casos de cada clase y el eje horizontal indica el número de clases, que podemos determinar manualmente (figura 15). De este modo, y por regla general, cuando aumentamos manualmente el número de clases, bajan las magnitudes del eje vertical.

Figura 15. Número de *likes* por entrada en Instagram



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

Otra manera de dirigir la visualización de la distribución interna de una variable numérica son los **diagramas de cajas y bigotes**. Se trata de una representación de datos cuantitativos enormemente informativa y relativamente sencilla de interpretar, pero que muchas veces no se utiliza porque requiere ciertos conocimientos matemáticos por parte del lector o visualizador.

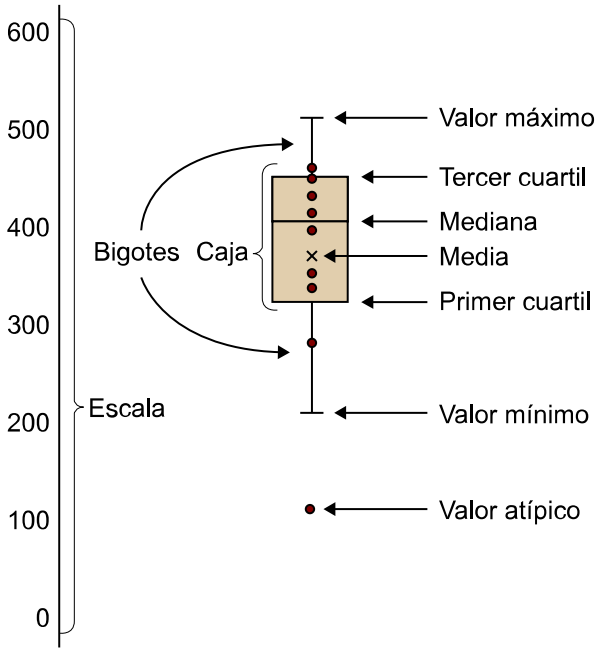
Mediante una serie de elementos gráficos muy sencillos (puntos, cajas, bigotes, una línea horizontal y una X), la gráfica permite representar las siguientes propiedades de los datos (figura 16):

- La **media aritmética** de la distribución: la suma de todos los valores dividida por el número total de casos.
- La **mediana** o el **segundo cuartil**, en el que se encuentran el 50 % de los datos.
- El **tercer cuartil**, donde se encuentran el 75 % de los datos.
- El **primer cuartil**, en el que se encuentran el 25 % de los datos.
- La **caja** representa **la amplitud intercuartílica** (IQR): la diferencia entre el tercer y el primer cuartil.
- Los **bigotes** se dibujan desde la caja hasta los **límites superior e inferior**, que se calculan restando 1,5*IQR al primer cuartil y sumando 1,5*IQR al tercer cuartil. El valor más bajo y el valor más elevado dentro de los límites superior e inferior serán los **valores máximo y mínimo** de los bigotes.
- Los **valores atípicos** son los que quedan fuera de los límites superior e inferior de los bigotes.

Los valores más relevantes del diagrama de cajas y bigotes son la media y la mediana. Si la distribución de la variable es equitativa, la media y la mediana son iguales. Por eso, las comparaciones entre la media y la mediana sirven especialmente para analizar las desigualdades internas de una variable.

Por ejemplo, un *youtuber* puede tener una media de 1.000 reproducciones por vídeo, pero una mediana de 0, porque tiene cuatro vídeos con 0 reproducciones y un vídeo con 5.000 reproducciones.

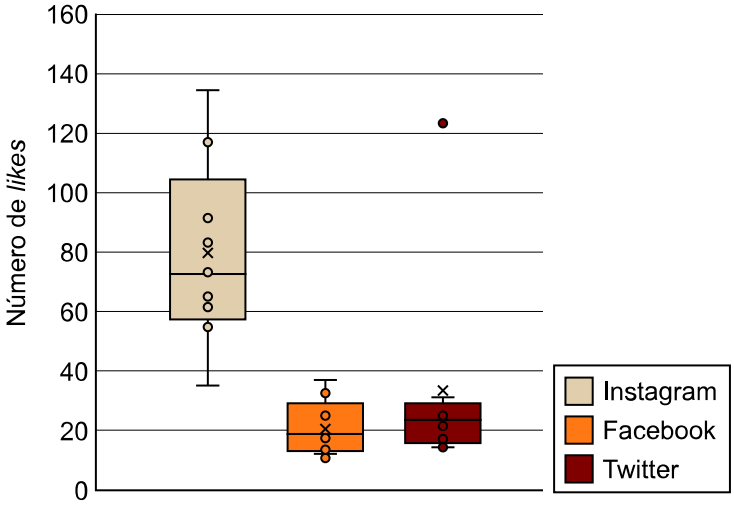
Figura 16. Elementos del diagrama de cajas y bigotes



Fuente: elaboración propia.

El diagrama de cajas y bigotes es enormemente informativo respecto a la distribución interna de una variable cuantitativa. Además, en contraste con el histograma, permite un gran número de configuraciones comparativas, tanto entre casos (por ejemplo, la distribución de *likes* por entrada en diferentes redes sociales, figura 17) como entre diferentes periodos temporales.

Figura 17. Distribución de *likes* por entrada en tres redes: cajas y bigotes



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

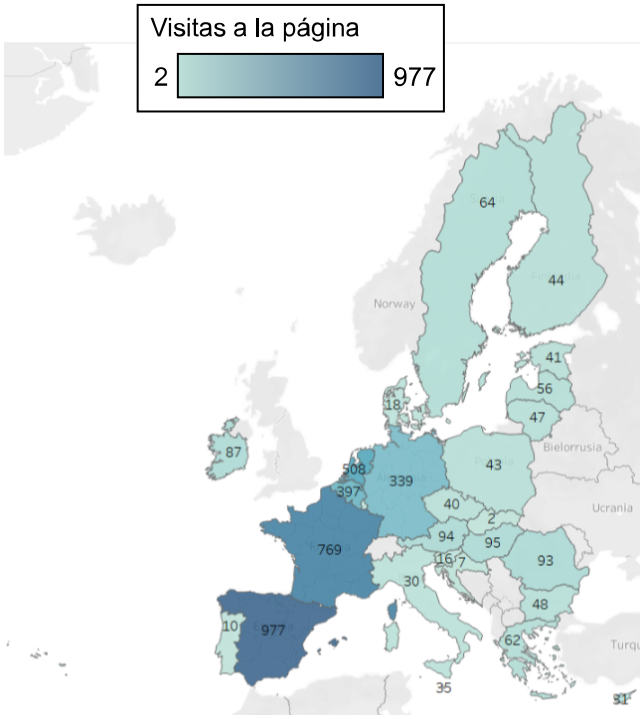
2.5. Visualización geoespacial

Entre las técnicas más habituales de visualización de datos también constan las visualizaciones geoespaciales. La más habitual y conocida de las representaciones geoespaciales es el mapa.

Los mapas que muestran variaciones de una variable cuantitativa en un territorio a partir de colores se denominan **mapas coropléticos**.

Estos mapas resultan muy intuitivos cuando el lector tiene un conocimiento más o menos profundo sobre el territorio y sobre la variable representada, pero también es fácil que induzcan a una serie de errores. En primer lugar, es importante dejar claro al lector o visualizador si nos encontramos con una variable estandarizada o no estandarizada. Cuando el objetivo de una visualización sea establecer comparaciones entre territorios, tendremos que estandarizar las variables representadas, en función del tipo de datos, dividiendo la métrica por los kilómetros cuadrados del territorio, por el total de población o por cada 100.000 habitantes. En cambio, cuando el objetivo de una visualización sea establecer una jerarquía en función del rendimiento de la misma variable (por ejemplo, visualizar los territorios que han aportado más tráfico a una página de Facebook, figura 18), nos tendremos que servir de métricas absolutas y no estandarizadas.

Figura 18. Total de visitas a una página de Facebook por países



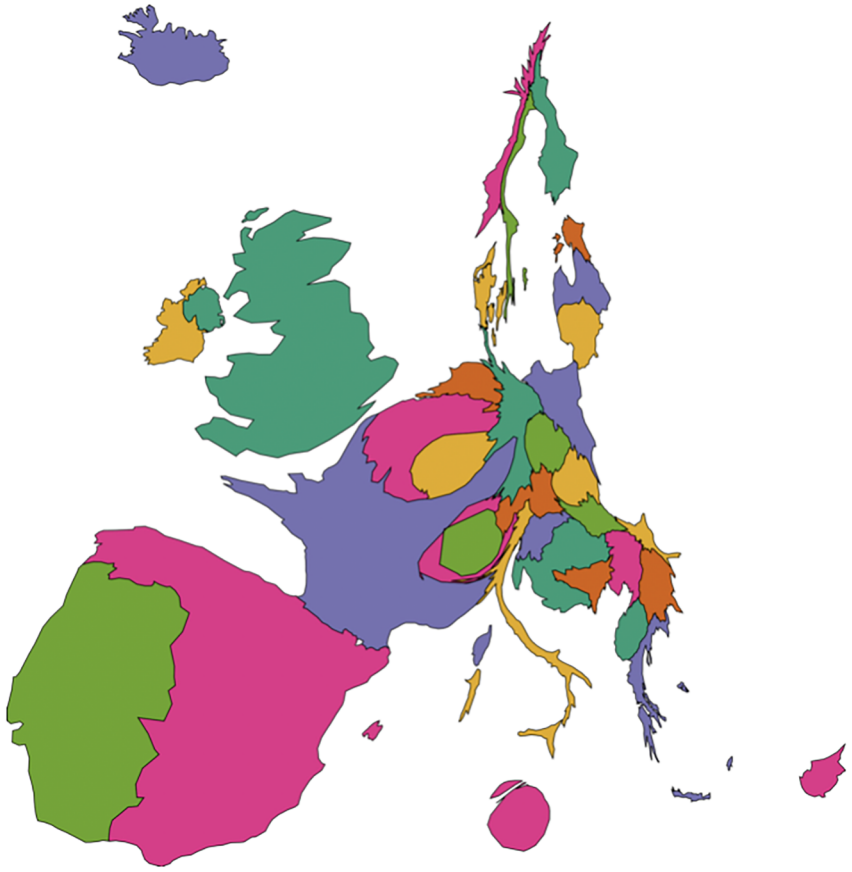
Fuente: elaboración propia con Tableau Public (datos aleatorios).

Una manera diferente de representar datos geoespaciales es ponderar los territorios en función de una variable cuantitativa.

Este tipo de mapas, que se denominan **cartogramas**, renuncian de este modo al principio de precisión tradicionalmente asociado a los buenos mapas, y priorizan la visualización de la variable cuantitativa en cuestión (por ejemplo, la aportación de tráfico a una página de Facebook; figura 19).

Los cartogramas incorporan una deformación espacial implementada algorítmicamente que magnifica los territorios con un mejor rendimiento de la variable cuantitativa en cuestión, y dan lugar a configuraciones que a veces pueden ser difíciles de reconocer, pero que pueden ser herramientas eficaces para la comprensión de un fenómeno.

Figura 19. Total de visitas a una página de Facebook por países



Fuente: elaboración propia con go-cart.io (Michael T. Gastner, Vivien Seguy y Pratyush More, 2018; datos aleatorios)

Los cartogramas pueden ser instrumentos muy útiles para ayudar a comprender un fenómeno, por ejemplo, si lo que se busca es captar la atención del lector o visualizador e incentivar su interés para ayudarlo a pensar. Aun así, resulta evidente que son visualizaciones complejas y que no facilitan interpretaciones precisas de los datos. Es trabajo del analista evaluar el lugar, el momento y la manera adecuada para utilizar ese tipo de visualizaciones disruptivas.

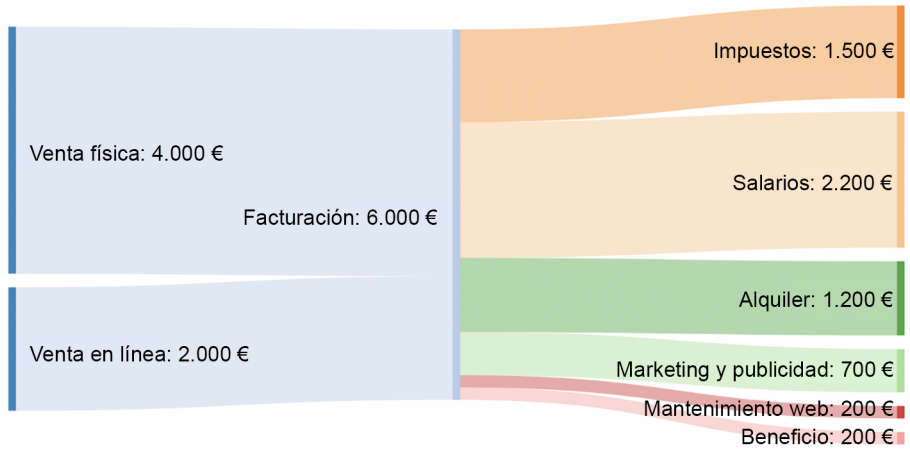
2.6. Otras técnicas de visualización de datos

Los adelantos tecnológicos de los últimos años, y sobre todo la ciencia computacional y el paradigma de los datos masivos, han dado lugar a muchas innovaciones en el campo de la visualización de datos, a partir de la generación de nuevas necesidades, retos y posibilidades técnicas. Actualmente, hay disponibles muchas herramientas que permiten representar datos mediante técnicas complejas que, en buena medida, desbordan las categorías y las agrupaciones que acabamos de ver, al incorporar elementos híbridos que no podríamos clasificar en ellas.

Una de estas visualizaciones es el **diagrama de Sankey** (figura 20), que incorpora elementos de proporcionalidad y que nos permite visualizar flujos de datos.

Cada eje vertical representa una variable cualitativa, la altura de los bloques representan la medida de los clústeres de datos y la medida del flujo representa el número de casos que comparten clústeres de varias variables. Se trata de diagramas que permiten visualizar comparaciones entre variables, correlaciones, distribuciones y tendencias temporales.

Figura 20. Diagrama de Sankey sobre el flujo de caja de una tienda en línea



Fuente: elaboración propia con sankeymatic.com (datos aleatorios).

Otro ejemplo es el **diagrama radial** (figura 21), que permite representar el rendimiento de varias variables cuantitativas en función de varias variables cualitativas en un plano bidimensional con ejes circulares.

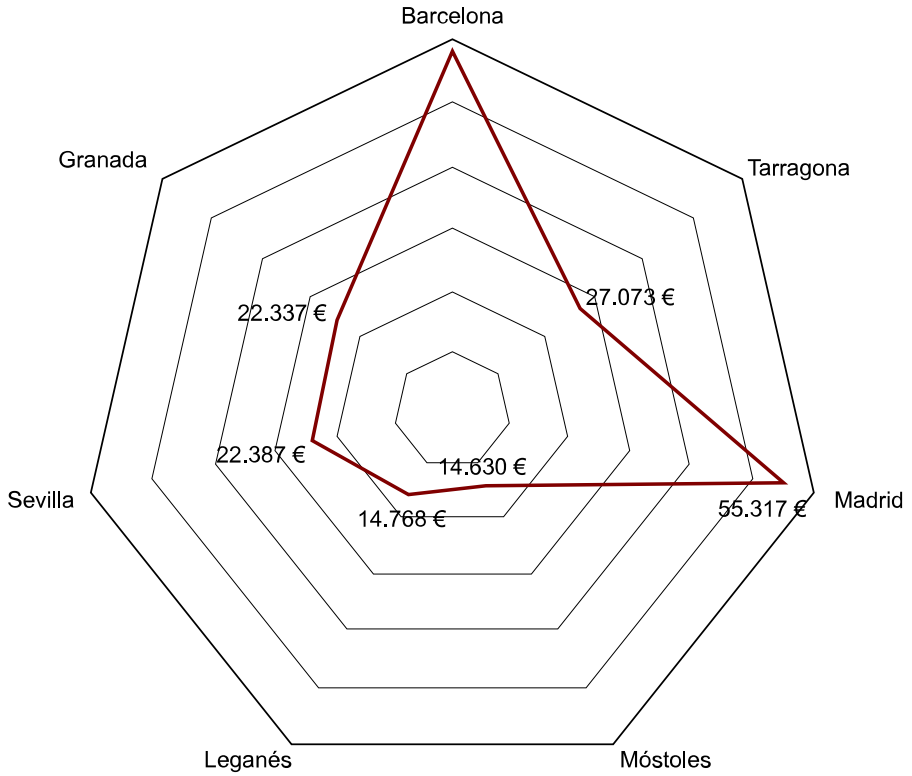
Es particularmente útil para comparar variables cuantitativas y para mostrar su distribución.

También el **diagrama de cuerdas** (figura 22) se basa en una disposición circular y permite representar las relaciones entre casos o variables de diferentes volúmenes y dimensiones en un plano bidimensional.

Es útil para representar distribuciones de datos y correlaciones. Un caso especial, y que mantiene una estrecha relación con el diagrama de cuerdas, es el grafo o red.

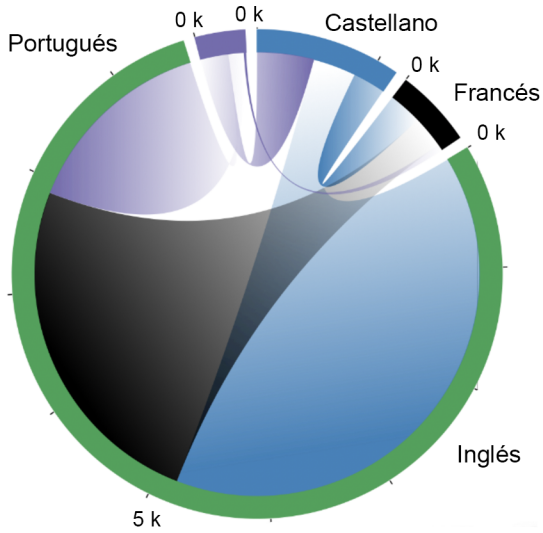
Los **grafos** (figura 23) son simultáneamente herramientas analíticas y visualizaciones de datos que permiten representar relaciones entre casos o variables mediante algoritmos complejos que enfatizan diferentes propiedades de los nodos y sus enlaces.

Figura 21. Diagrama radial sobre las ventas de una empresa por ciudades



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

Figura 22. Diagrama de cuerdas sobre las traducciones en la Wikipedia



Fuente: Wikicommons (imagen libre de derechos).

Figura 23. Grafo de etiquetas utilizadas conjuntamente en Twitter

Fuente: elaboración propia con Gephi (datos aleatorios).

2.7. Errores y malas prácticas en la visualización de datos

Las visualizaciones de datos son representaciones que nos ayudan a pensar y a comprender un fenómeno mediante la explicación de una historia o de un relato. Hasta ahora, hemos visto una serie de técnicas que nos ayudan a visualizar fenómenos a partir de variables cuantitativas, cualitativas, temporales y geoespaciales, y nos hemos centrado en las buenas prácticas y en las condiciones que permiten maximizar la funcionalidad de las diferentes representaciones. En esta sección nos fijaremos en lo contrario, en algunos errores y malas prácticas que pueden comprometer el sentido y el objetivo de una vi-

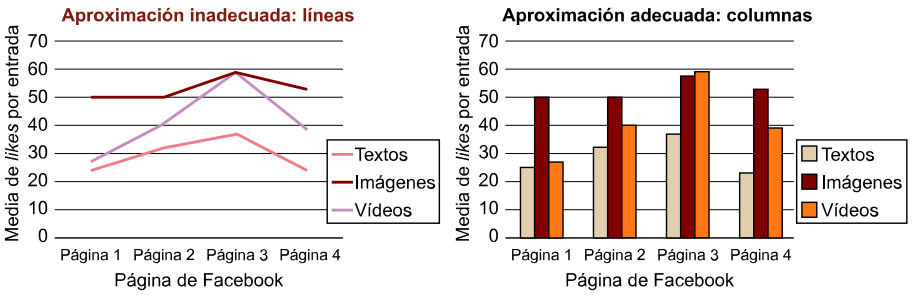
sualización de datos. Tanto por errores técnicos como por errores humanos –y también debido a manipulaciones malintencionadas–, hay un gran número de factores que pueden comprometer la interpretación de un conjunto de datos verídicos y válidos. En este sentido, es importante distinguir los problemas interpretativos derivados de una visualización disfuncional de los problemas asociados a la veracidad y la validez de los mismos datos que abordamos en el módulo 1.

Así, pues, aquí entenderemos que los errores de visualización son los que, partiendo de datos correctos, inducen a conclusiones incorrectas en el lector o visualizador como consecuencia de un mal uso de los elementos gráficos.

1) Un primer elemento sobre el cual hay que poner atención es, simplemente, la disposición **correcta de los datos en una gráfica**: distinguir los modos de representación de una variable, usar órdenes cronológicos en las representaciones temporales, agrupar casos de manera significativa, etc. Uno de los errores de este tipo más habituales es la confusión entre una variable continua y una variable discreta, plasmada en una elección incorrecta de líneas o columnas.

Por ejemplo, cuando establecemos comparaciones entre casos diferentes (por ejemplo, el compromiso de diferentes páginas de Facebook según varios tipos de contenido) no es adecuado recurrir a elementos como la línea, que transmiten continuidad, progreso o evolución de una categoría a la siguiente. Las columnas son mucho más adecuadas para este tipo de casos (figura 24).

Figura 24. Confusión entre variables continuas y discretas



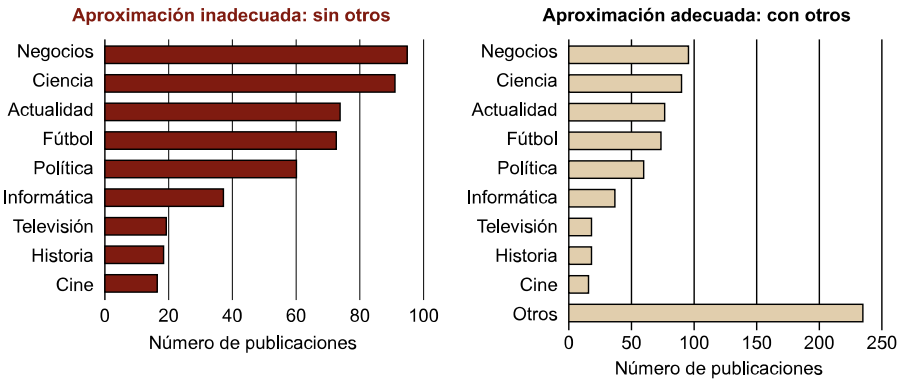
Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

2) Otro error clásico es la **invisibilización de casos**, que puede derivar de una mala agrupación de datos y de una clasificación mal establecida.

A menudo, los analistas elaboran clasificaciones con los «5 mejores» o los «10 mejores» casos, para resumir los datos, centrándose en los casos más importantes de una distribución. Esta estrategia es correcta si lo que se quiere es destacar la cifra absoluta que toma una categoría sin importar su presencia relativa; pero, en ciertas ocasiones, esta clasificación puede inducir a error (por ejemplo, si quiere elaborarse un *ranking* de temas de conversación y la categoría «otros» resulta ser más voluminosa que el tema clasificado más habitual;

figura 25). Para evitar la sobreponderación de las categorías principales de una clasificación, el analista tendrá que valorar la inclusión de elementos adicionales en un «top 10», como puede ser la categoría «otros»: visualmente puede resultar menos atractivo, pero la funcionalidad de la visualización puede ser mucho más elevada según el tipo de pregunta al que se quiera responder.

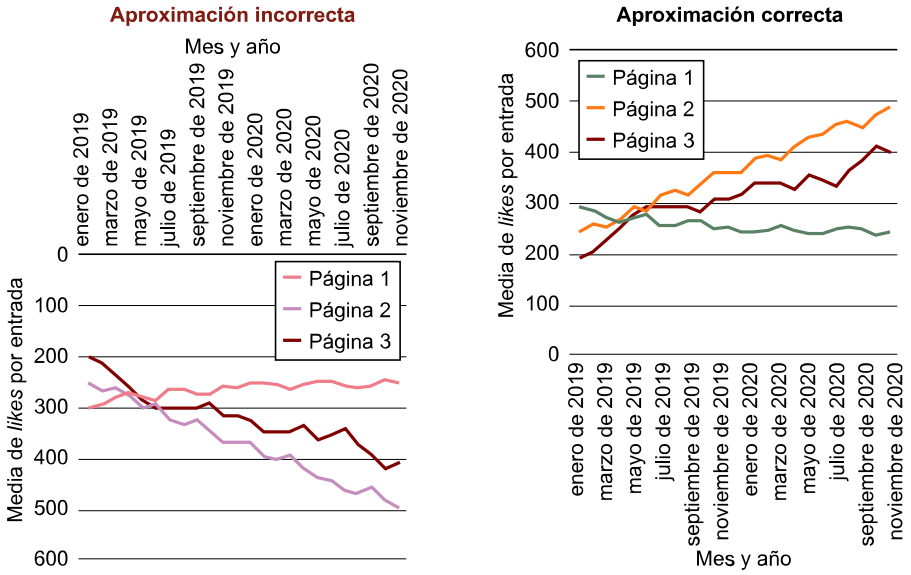
Figura 25. Invisibilización y sobreponderación



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

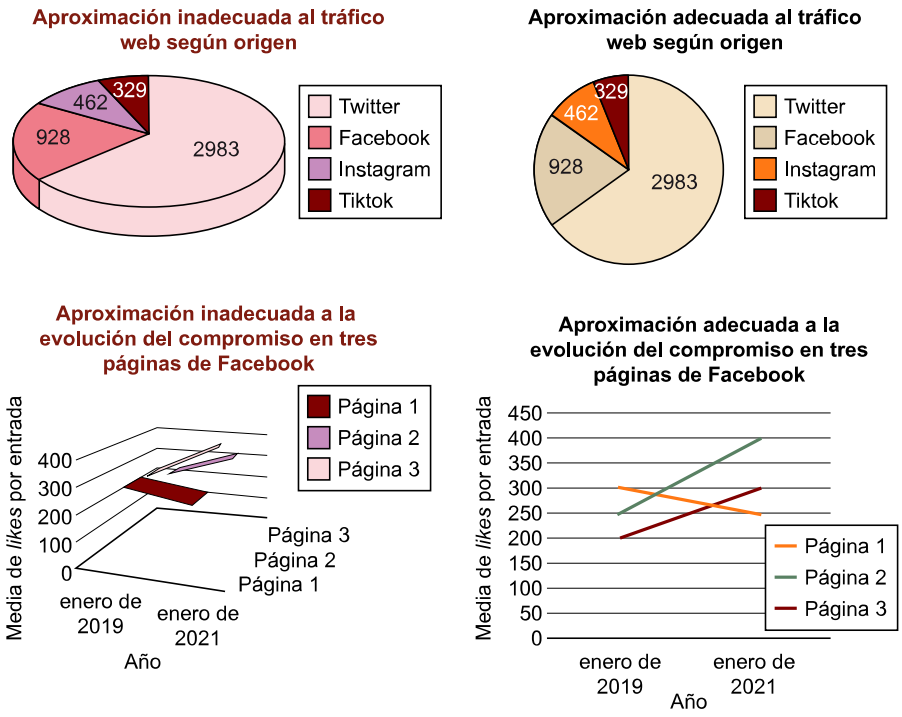
Muchos errores y malas prácticas derivan de usos de los elementos gráficos que contravienen las convenciones más básicas, como por ejemplo la inversión del eje de las ordenadas (el eje *y*) sin ninguna buena razón para hacerlo (figura 26). Una gráfica tiene que ser fácil de leer y, por lo tanto, es bueno que establezca puntos de conexión con los conocimientos y los hábitos de su audiencia. La disrupción puede ser una herramienta para llamar la atención del lector o visualizador en algunas ocasiones (por ejemplo, cuando quiere causarse sorpresa con un cartograma), pero hay que ser muy cuidadoso con la distorsión de los elementos de una gráfica que puedan pasar desapercibidos y provocar, de ese modo, una interpretación errónea del fenómeno representado. De manera similar, un factor de distorsión de los elementos gráficos muy utilizado es la representación en tres dimensiones. Por regla general, las tres dimensiones no aportan ningún valor adicional a una representación y complican la interpretación de fenómenos que pueden visualizarse mucho mejor mediante los recursos disponibles para las técnicas en dos dimensiones (figura 27).

Figura 26. Ruptura de las convenciones básicas



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

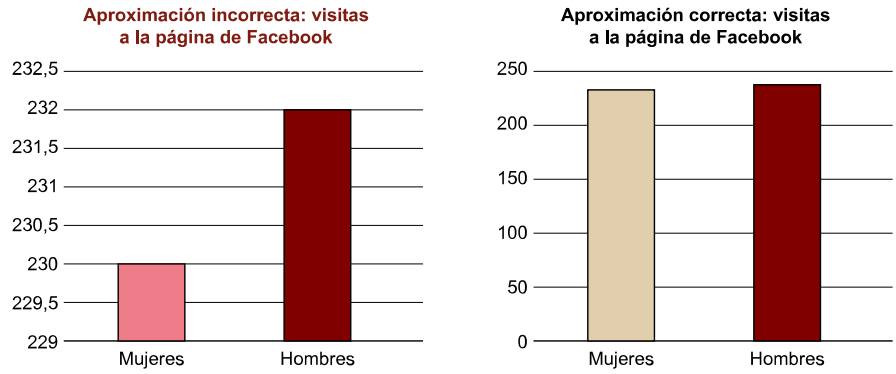
Figura 27. Las tres dimensiones gratuitas e innecesarias



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

3) De entre los errores derivados de la violación de las convenciones básicas, hay uno que es especialmente grave y que, por desgracia, es muy habitual. Se trata de los **ejes recortados y con bases diferentes de cero** en gráficas de columnas o barras. La gravedad de este error consiste en el enorme poder de sugestión de la representación, que es capaz de transmitir diferencias allí donde no las hay (figura 28). Recortar los ejes de una gráfica es una de las maneras más eficientes de manipular un conjunto de datos y de hacer que transmitan incorrecciones o mentiras. En consecuencia, es una práctica inaceptable.

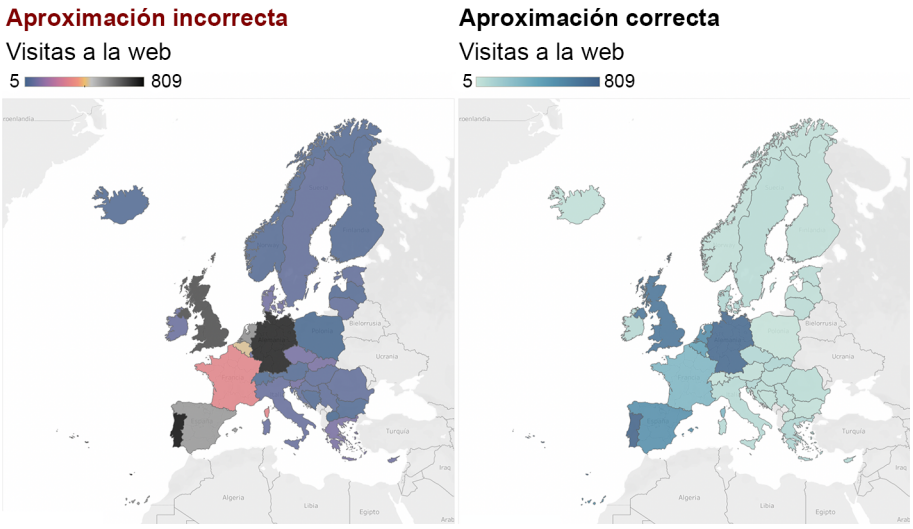
Figura 28. Ejes recortados



Fuente: elaboración propia con Microsoft Excel (datos aleatorios).

4) Para acabar, hay que destacar una serie de problemas derivados de una **mala utilización de los colores** en la visualización de datos. Los colores pueden usarse tanto para distinguir variables cualitativas (por ejemplo, hombres y mujeres, páginas de Facebook, etc.) como para indicar la intensidad de un valor cuantitativo, especialmente en los mapas coropléticos (por ejemplo, número de visitas a una web según el territorio; figura 29). En el primer caso, es aconsejable utilizar paletas multicolor o monocromáticas que permitan distinguir los casos correctamente; en el segundo caso, además, hace falta que la utilización del color no conduzca a malas interpretaciones, por ejemplo, con una escalera sin sentido cromático o mal centrada.

Figura 29. Colores mal configurados



Fuente: elaboración propia con Tableau Public (datos aleatorios).

3. Herramientas de visualización de datos

La mayoría de programas y paquetes ofimáticos utilizados por millones de usuarios desde los años noventa del siglo XX[3] incorporan herramientas nativas para la visualización de datos en los procesadores de textos, en las hojas de cálculo y también en los editores de diapositivas. Con estas herramientas se pueden llevar a cabo muchas de las técnicas de visualización que hemos visto en la sección anterior e incrustarlas en informes, presentaciones u otros tipos de soportes digitales o impresos. En términos generales, se trata de herramientas enfocadas a una comunicación unidireccional y no interactiva, que es la que permiten los soportes como el papel o los documentos en PDF. Por otro lado, hay herramientas de visualización de datos vinculadas al universo de los lenguajes programables[4] que sí están mucho más orientadas a la interactividad con el usuario, al estar pensadas para ser visualizadas en entornos web o similares. Además de estas dos grandes categorías, hay que destacar dos tipos de herramientas más, que han cobrado importancia en los últimos años en la visualización de datos. Por un lado, las infografías y, por otro lado, el software de inteligencia empresarial (*business intelligence* o BI).

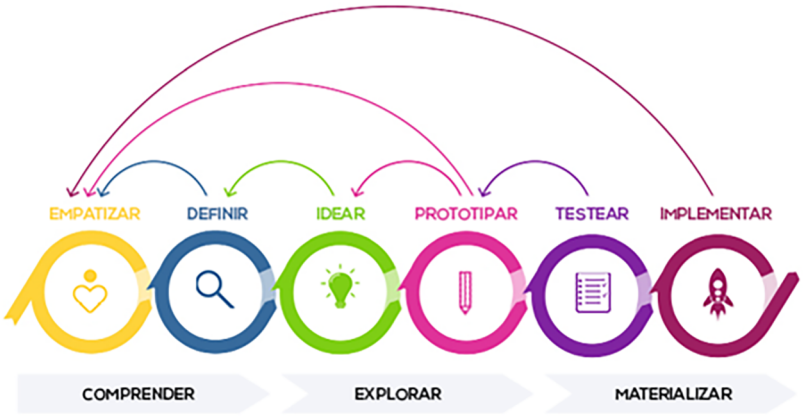
(3) Nos referimos a Microsoft Office, WordPerfect Office Suite, Google Docs, LibreOffice, etc.

(4) Por ejemplo, Chart.js o D3.js para Javascript, qplot o ggplot2 para R, Plotly o Altair para Python, entre otras muchas.

Las **infografías** consisten en representaciones que combinan elementos visuales, textos cortos, visualizaciones de datos y otros recursos estilísticos que comparten objetivos con las otras técnicas de visualización de datos: describir o explicar un fenómeno de manera atractiva, agradable y fácil de entender (Mark Smiciklas, 2012).

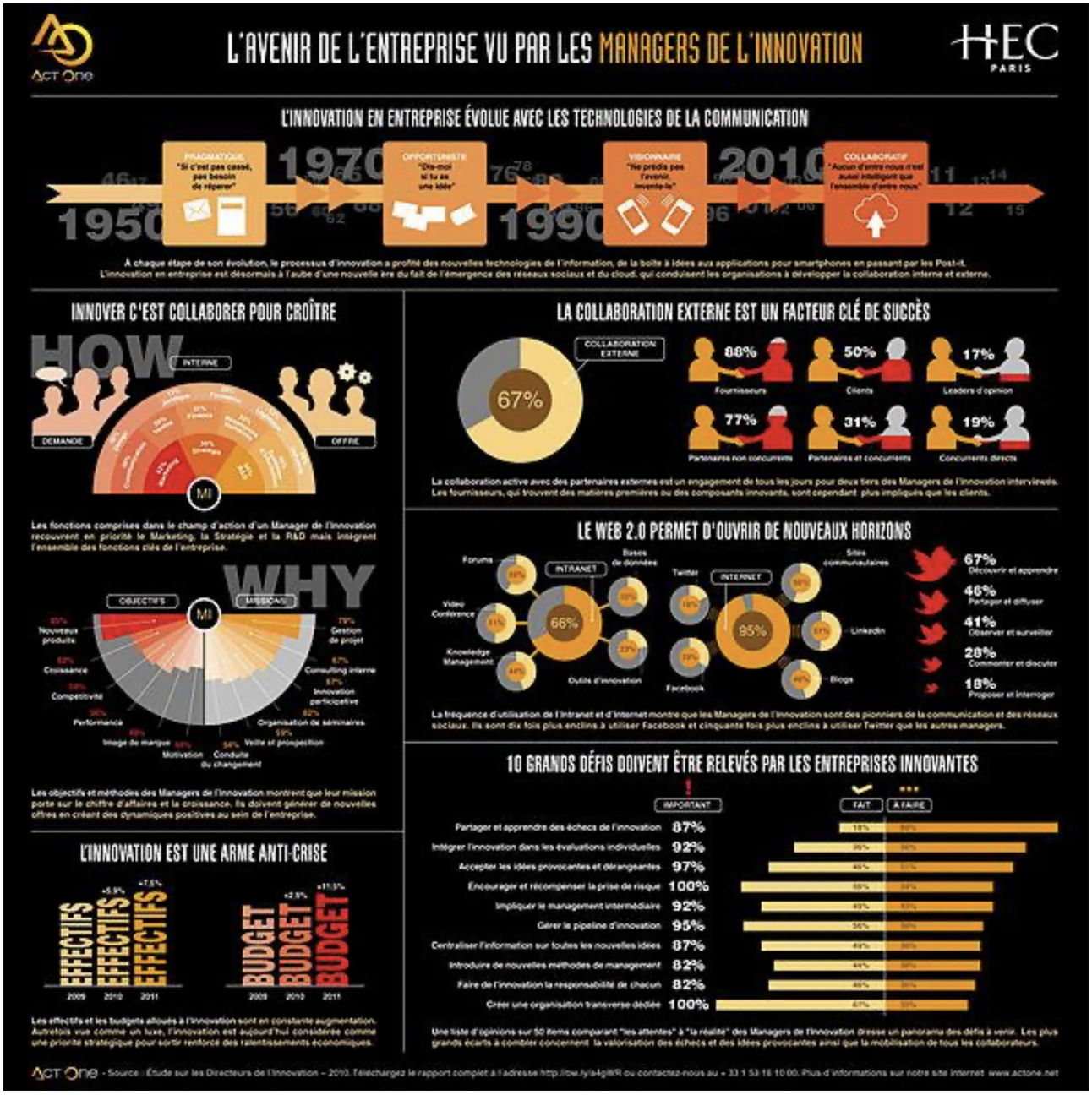
No puede afirmarse que la infografía sea una técnica de visualización de datos en sí misma, puesto que la visualización de datos no es un elemento imprescindible para una infografía (por ejemplo, una infografía puede describir una idea de manera esquemática; figura 30). Aun así, y por regla general, en las infografías se hace un uso muy amplio y variado de varios recursos y técnicas de visualización de datos para facilitar la comprensión de los fenómenos que se describen (por ejemplo, el fenómeno de la innovación en la empresa privada; figura 31). Las infografías son, por lo tanto, soportes para la visualización de datos en los cuales se tiende a la utilización de varias técnicas de manera simultánea, acompañadas de textos, ideogramas y otros recursos gráficos que acompañan al lector o visualizador.

Figura 30. Infografía sobre el pensamiento de diseño o *design thinking*



Fuente: Wikicommons (imagen libre de derechos).

Figura 31. Infografía sobre la innovación empresarial



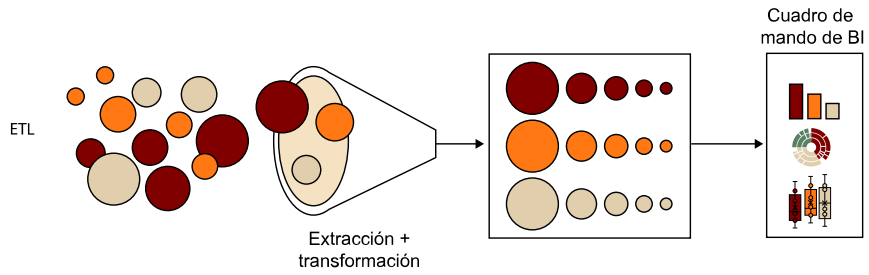
Fuente: Wikicommons (imagen libre de derechos).

Durante los últimos años, han proliferado en Internet una gran cantidad de herramientas gratuitas o de semipago (*freemium*) con las cuales crear una infografía es una tarea de lo más sencilla. Entre los recursos más populares actualmente, destacan productos como *canva.co*, *visme.co*, *venngage.com* o Adobe Spark. A pesar de que la mayoría de recursos están orientados a la elaboración de infografías en formatos de imagen o de PDF, cada vez hay más servicios que prevén visualizaciones interactivas para soportes virtuales.

El segundo grupo de «nuevas» herramientas, que han cobrado importancia durante la década de 2010 a 2020, son los paneles de control (*dashboards*) de BI. Suelen constituir la última fase automatizada en un sistema de ETL: los datos se extraen de sus fuentes, se procesan, se transforman, se almacenan y se

visualizan en un panel de control BI para que el analista pueda concentrarse en interpretarlos (figura 32). Entre las herramientas de BI más populares actualmente están Tableau, Google Data Studio, Microsoft Power BI o SAP Analytics Cloud, y todas ofrecen una gran cantidad de conectores a diferentes bases de datos estructuradas y semiestructuradas y capacidades gráficas. La capacidad de interacción de datos diversos es uno de los puntos más importantes de este tipo de software, orientado a la inmersión en el dato mediante notificaciones emergentes configurables y filtros que permiten dirigir preguntas *ad hoc* por parte del analista.

Figura 32. De la extracción de datos a la visualización

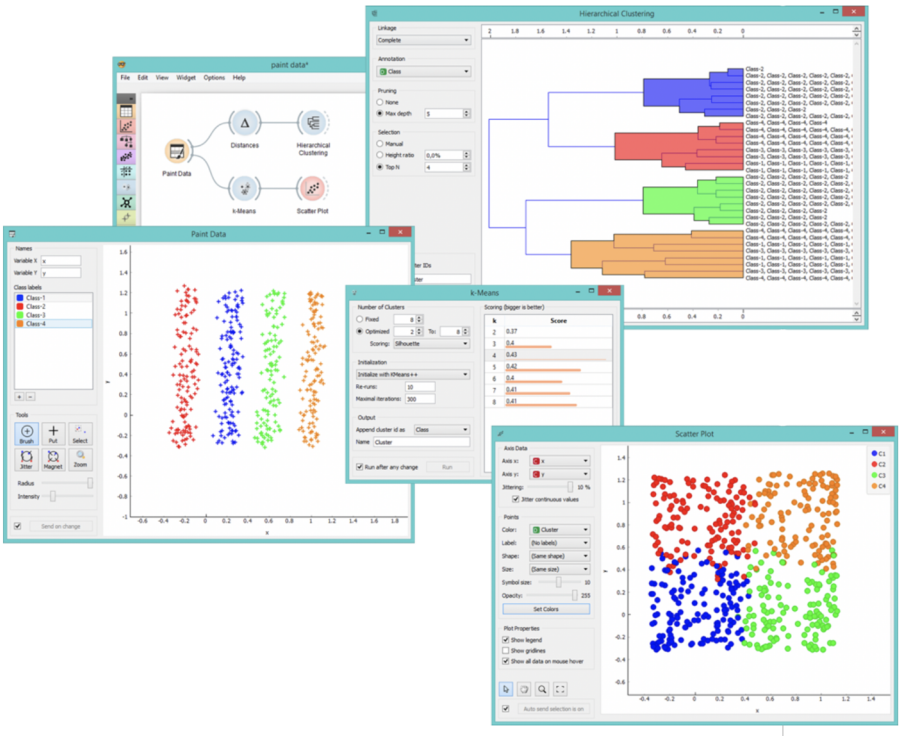


Fuente: elaboración propia.

Los **paneles de control de BI** son herramientas especialmente interesantes en cuanto a los escenarios de ETL, cuando disponemos de datos ya procesados, sobre los cuales se han aplicado los algoritmos necesarios para su análisis.

En este caso, el *software* se limita a leer y a «pintar» los datos según los parámetros que determine el analista. Los sistemas de BI también son capaces de crear nuevas variables mediante operaciones sencillas de procesamiento y transformación de datos, pero no suelen estar preparados para llevar a cabo operaciones complejas ni algoritmos de PLN o ML. Esto hace que no sean el mejor tipo de software en entornos de ELT, en los cuales querremos trabajar con los datos antes de visualizarlos. Actualmente, hay algunas propuestas en el mercado que combinan elementos de BI e IA, como IBM Cognos o Salesforce Einstein, y que permiten cierta integración de procesos, pero para la mayoría de usos, el analista tendrá que disponer de software de minería de datos como SPSS Modeler, RapidMiner, KNIME u Orange Data Mining si se encuentra en un escenario de ELT. Afortunadamente, muchos de estos programas también presentan opciones de visualización de datos interactivas muy interesantes (figura 33).

Figura 33. Visualización interactiva de datos con Orange Data Mining



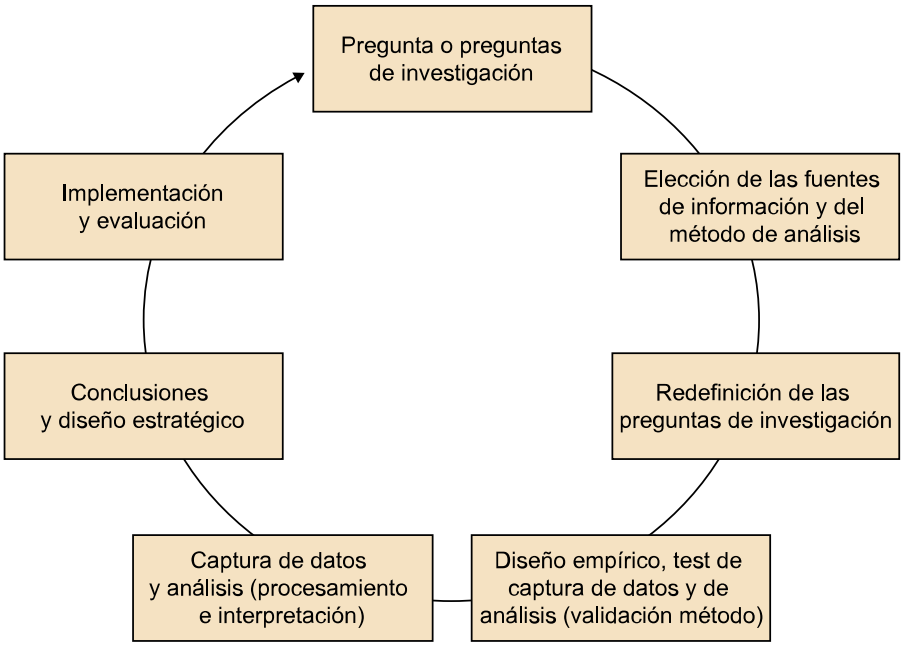
Fuente: Wikicommons (imagen libre de derechos).

4. Planificar un proyecto de investigación en medios sociales

A lo largo de estos tres módulos, hemos podido ver qué es el paradigma de macrodatos, en qué consiste, cuáles son las innovaciones principales que aporta respecto a paradigmas anteriores y cómo podemos acercarnos como analistas, con la intención de generar valor a partir de los macrodatos. Hemos conocido una serie de técnicas de análisis de datos que nos permiten articular estrategias de generación de conocimiento en contextos de macrodatos y en los cuales, a menudo, querremos generar nuevas variables que nos ayuden a responder a nuestras preguntas de investigación. Finalmente, hemos visto una serie de técnicas y herramientas clave en las tareas de visualización de datos, que ayudan al analista y al lector a interpretar y comprender los fenómenos representados.

El universo de los macrodatos y las metodologías de investigación asociadas –en las cuales el elemento inductivo gana fuerza ante el método hipotético-deductivo tradicional– trastocan muchos de los elementos clásicos que deben estar presentes en una investigación. Pero hay un elemento que no cambia, porque es consustancial con la investigación misma: todo parte, siempre, de una pregunta de investigación. La propuesta siguiente (figura 34) sintetiza los siete pasos que hay que llevar a cabo en una investigación en entornos de macrodatos.

Figura 34. Los siete pasos de la investigación en macrodatos



Fuente: elaboración propia.

La propuesta de siete pasos ha sido elaborada teniendo en cuenta que cualquier investigación social es en realidad una actividad creativa (Hans Joas, 1992) en que medios y finalidades se reelaboran mutuamente en función de las necesidades que emergen en el mismo proceso. También se han tenido en cuenta las particularidades del paradigma de los macrodatos y su orientación hacia la creación de valor.

1) Pregunta o preguntas de investigación. Partir de una pregunta o de un grupo acotado de preguntas es la mejor manera de optimizar los recursos en un contexto de investigación. Estas nacen de la curiosidad y las inquietudes del investigador, pero también nacen de la familiaridad con un campo de estudio y del conocimiento de sus carencias en un momento particular. En un contexto de investigación aplicada (por ejemplo, en una empresa, una institución o un proyecto de cualquier tipo), las preguntas también nacen de las necesidades específicas de creación de valor.

2) Elección de las fuentes de información y del método de análisis. Una vez se tiene claro qué se quiere investigar y por qué razón, es el momento de identificar qué fuente o fuentes de información pueden contribuir a responder a la pregunta y qué procesos y técnicas hay que aplicar para hacerlo. En este punto, es importante ver las fuentes de datos no solamente como lo que son (los datos sucios), sino como lo que pueden llegar a ser después de aplicar las transformaciones y los algoritmos necesarios.

3) Redefinición de las preguntas de investigación. Investigar es una actividad creativa y, como tal, requiere retroalimentación entre medios y finalidades. Después de considerar la viabilidad metodológica de un proyecto desde el punto de vista de la captura, el tratamiento y el análisis de datos, siempre es conveniente revisar cómo el universo de posibilidades disponibles modifica las preguntas iniciales. Ahora es el momento de reelaborar las metas de una investigación para adaptarlas al principio de realidad y disponibilidad, y de constatar que todavía vale la pena continuar el proyecto.

4) Diseño empírico, test de captura de datos y de análisis. En esta fase, es importante diseñar el modelo de captura, tratamiento y análisis de datos y asegurarse de que todo funciona según lo previsto. Según el tipo de proyecto, será suficiente conseguir datos retroactivos o de un periodo corto en tiempo real, especialmente si lo que se quiere es llevar a cabo un proyecto de ETL en ejecución permanente. Si la cosa no funciona como es debido, toca recular y volver a definir las preguntas. Si todo va bien, pasamos a la fase siguiente.

5) Captura de datos, visualización y análisis. Esta fase suele ser la más larga en cualquier proyecto. De hecho, muchos proyectos lo ejecutan de manera permanente, puesto que la captura de datos de los medios sociales puede llevarse a cabo a largo plazo. Al margen de que esta fase sea más o menos costosa en tiempo y recursos, la captura de datos y su análisis no tendrían que ser muy

problemáticos si las fases anteriores se han implementado adecuadamente. Si aparecen errores importantes, hay que corregirlos antes de pasar a la próxima fase.

6) Conclusiones y diseño estratégico. Tanto en el caso de un estudio temporalmente acotado como en el caso de un proyecto estable, es importante documentar los resultados y discutirlos a partir de las preguntas de investigación. Como el paradigma de los macrodatos está fuertemente orientado a la acción y a la creación de valor, ahora también es el momento de transformar el conocimiento generado en acciones estratégicas realizables y evaluables.

7) Implementación y evaluación. Finalmente, es importante poner en marcha las acciones planificadas a partir de los resultados del análisis y evaluarlos desde el punto de vista de su rendimiento y para la obtención de resultados. Las necesidades de evaluación de las acciones estratégicas probablemente derivarán en nuevas necesidades analíticas, que habrá que convertir en preguntas de investigación y articular en el plan metodológico. En un proyecto estable, es importante medir también los efectos de las acciones implementadas en el propio instrumento de captura.

Los medios sociales son un entorno privilegiado para investigar, implementar los resultados de investigación y evaluar su rendimiento. Gran parte de la actividad que se desarrolla y registra es enormemente trazable, y puede ser utilizada para la generación de valor por una empresa o proyecto. Las métricas de los medios sociales, el análisis de redes y los algoritmos de PLN proporcionan una gran variedad de opciones de análisis, y muchos se pueden implementar con un coste de captura de datos muy bajo. A la hora de generar valor, es muy importante no perder la perspectiva de lo que se está haciendo y de por qué razón (qué es aquello que se quiere obtener) y no distanciarse demasiado del principio de realidad y disponibilidad. Por eso, es imprescindible que el analista elabore una lista corta de prioridades por cada proyecto, y que dimensione correctamente cada una de las fases y partes antes de empezar a sumergirse en el apasionante mundo del análisis de datos.

Bibliografía

Croxton, Frederick E.; Stein, Harold (1932). «Graphic Comparisons by Bars, Squares, Circles, and Cubes». *Journal of the American Statistical Association* (vol. 27, n.º 177, págs. 54-60).

Gastner, Michael T.; Seguy, Vivien; More, Pratyush (2018). «Fast Flow-Based Algorithm for Creating Density-Equalizing Map Projections». *Proc. Natl. Acad. Sci. USA* (vol. 115, n.º 10, págs. E2156–E2164).

Joas, Hans (1992). *Die Kreativität des Handelns*. Fráncfort am Main: Suhrkamp.

Lima, Manuel (2017). *Material Design. Data Visualization* [en línea]. *Medium.com*. [Fecha de consulta: 25 de junio de 2020]. <https://material.io/design/communication/data-visualization.html#principles>.

Ponjuán Dante, Gloria (1998). *Gestión de información en las organizaciones. Principios, conceptos y aplicaciones*. Santiago de Chile: Universidad de Chile. Centro de Capacitación en Información.

Robinson, Imogen (2016). *Data Visualisation: Contributions to Evidence-Based Decision-Making*. Wallingford: SciDev.Net.

Smiciklas, Mark (2012). *The Power of Infographics. Using Pictures to Communicate and Connect with Your Audiences*. Indianápolis: Que Publishing.

Telea, Alexandru C. (2014). *Data Visualization. Principles and Practice*. CRC Press.

Wilke, Claus O. (2019). *Fundamentals of Data Visualization. A Primer on Making Informative and Compelling Figures*. Sebastopol, CA: O'Reilly Media.