

Guía de visualización de datos



Generalitat de Catalunya

Aviso legal:



Esta obra está sujeta a una licencia Reconocimiento 4.0 de Creative Commons. Está permitida su reproducción, distribución, comunicación pública y transformación para generar una obra derivada, sin ninguna restricción, siempre que se cite al titular de los derechos (Generalitat de Catalunya). La licencia completa se puede consultar en:

<https://creativecommons.org/licenses/by/4.0>

Primera edición: abril de 2019

ISBN: 978-84-393-9734-2

Guía de visualización de datos

1. Introducción a la guía	5
2. Qué es la visualización de datos	10
2.1. Visualizaciones para explorar	12
2.2. Visualizaciones para analizar	13
2.2.1. Cuadro de mando	14
2.2.2. Aplicación de analítica visual	14
2.3. Visualizaciones para explicar	15
2.3.1. Infografía	16
2.3.2. Narrativa por desplazamiento	17
2.3.3. Presentación	17
2.3.4. Vídeo	20
3. Metodología	21
3.1. Fase 1 - Estrategia	21
3.1.1. Investigación y análisis	21
3.1.2. Objetivos	23
3.1.3. Indicadores	25
3.2. Fase 2 - Datos	27
3.2.1. Obtención	27
3.2.2. Formateo y limpieza	27
3.2.3. Procesamiento	28
3.3. Fase 3 - Diseño	28
3.3.1. Esbozar	28
3.3.2. Prototipar	29
3.3.3. Finalizar	30
4. Gráficos	31
4.1. Comparaciones	32
4.1.1. Gráfico de barras	33
4.1.2. Gráfico de barras agrupadas	37
4.1.3. Gráfico de barras apiladas	38
4.1.4. Gráfico de radar	41
4.1.5. Gráfico de intensidad de colores	43
4.1.6. Gráfico de marcas	45
4.1.7. Gráfico múltiple	47

4.2. Tendencias	48
4.2.1. Gráfico de líneas	49
4.2.2. Gráfico de pendientes	50
4.2.3. Minigráfico	51
4.3. Mapas	52
4.3.1. Mapa de coropletas	53
4.3.2. Mapa de símbolos proporcionales	54
4.4. Partes de un total	55
4.4.1. Gráfico de sectores	56
4.4.2. Pictograma	60
4.4.3. Mapa de árbol	61
4.5. Distribuciones	62
4.5.1. Histograma	63
4.5.2. Diagrama de caja	64
4.6. Correlaciones	67
4.6.1. Gráfico de dispersión	68
4.6.2. Gráfico de burbujas	72
4.6.3. Coordenadas paralelas	74
4.7. Conexiones, relaciones y redes	76
4.7.1. Diagrama de nodos y aristas	77
4.7.2. Diagrama de Sankey	78
5. Principios del diseño	80
5.1. Pensamiento visual	80
5.2. Color	82
5.3. Forma	84
5.4. Interacción	85
5.5. Recomendaciones generales	87
5.5.1. Menos es más	87
5.5.2. Equilibrio entre funcionalidad y estética	87
5.5.3. La forma sigue a la necesidad	88
5.5.4. Texto	89
5.5.5. Maquetación	90
6. Bibliografía	91
7. Anexos	92
7.1. Glosario	92
7.2. Herramientas	93

1. Introducción a la guía

Actualmente se generan datos sobre prácticamente cualquier fenómeno humano o natural. Ámbitos tan diversos como la salud, el deporte, la política, la educación o el *marketing*, por citar solo algunos, utilizan los datos para entender mejor la realidad. Por lo tanto, cada vez más personas necesitan desarrollar competencias para trabajar con datos y entender su significado e impacto. En cambio, un estudio reciente estima que solo el 17% de los ciudadanos europeos están preparados para utilizar datos.

Esta guía quiere ayudar a todas las personas que trabajan con datos, en particular a explorarlos, analizarlos y explicarlos gráficamente. En ella se utiliza el marco de trabajo de la visualización de datos, una disciplina que explica cómo tratar la información de un modo visual y que ofrece dos grandes ventajas. En primer lugar, el lenguaje visual es el más adecuado para hacer accesibles los datos a públicos no especialistas. En segundo lugar, la visualización facilita destacar en un contexto de sobrecarga informativa y, por lo tanto, ayuda a hacer llegar un mensaje a un público determinado.

Ahora bien, un gráfico mal diseñado puede dar lugar a una interpretación incorrecta de los datos y, por lo tanto, provocar una mala decisión. Asimismo, un gráfico correcto que responde a una pregunta irrelevante no tendrá ningún impacto. Es por eso que en esta guía se tratan tanto los aspectos de diseño como los metodológicos, con el fin de asegurar que se formulan las preguntas adecuadas y que se escogen los indicadores pertinentes.

Además, se hace referencia a las pautas para la elaboración de gráficos, tablas e infografías que deben armonizarse con el diseño del resto de soportes comunicativos de la Generalitat de Cataluña. La identidad corporativa es un elemento básico de la imagen que la institución proyecta a la ciudadanía y que identifica la oficialidad y veracidad de los datos que se presentan.

La guía se divide en cuatro grandes bloques: qué es la visualización de datos, metodología, gráficos y principios del diseño. Aunque la guía se puede leer linealmente, habrá lectores que la podrán utilizar como referencia en momentos puntuales. Por este motivo, a continuación se presentan cuatro infografías que resumen los cuatro grandes bloques de contenido. Estas infografías funcionan como un índice interactivo para acceder a aquellos contenidos de la guía que se pueden consultar puntualmente.

¹ Qlik. Data Equality Campaign. <<http://dataequality.org/>> [recurso electrónico]. [Consultado: noviembre 2017]

Qué es la visualización de datos

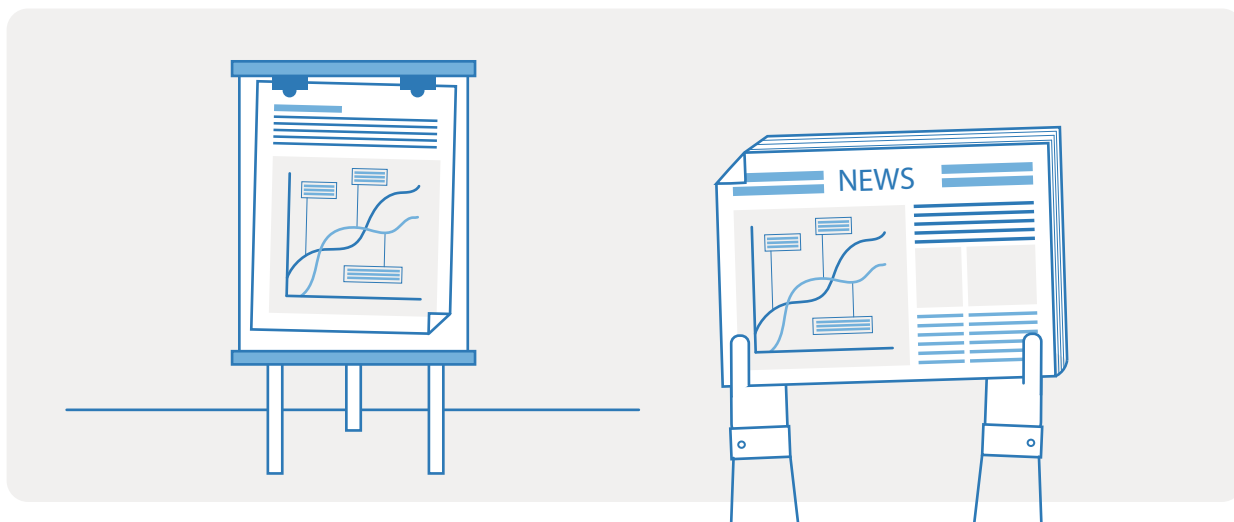
ES EXPLORAR unos datos para responder preguntas o plantear nuevas cuestiones.



ES ANALIZAR patrones, relaciones y valores atípicos entre los datos.



ES EXPLICAR una historia con datos.



Cómo se elabora una visualización

DEFINIENDO una estrategia



① Investigar



② Definir objetivos



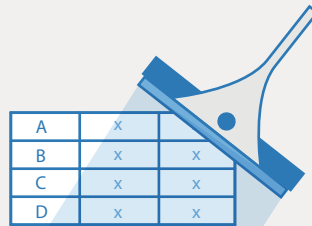
③ Definir indicadores

PREPARANDO unos datos



① Obtener los datos

BD C
A 1983 1
0 67% 230€ ff
33,57 12°

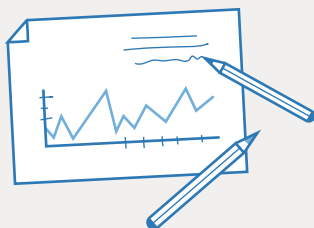


② Darles forma y limpiar



③ Procesar

DISEÑANDO unos gráficos



① Esbozar



② Prototipar



③ Finalizar

Gráficos

Comparaciones

Para comparar diferentes variables o categorías entre sí.



Tendencias

Para entender la evolución temporal de variables cuantitativas.



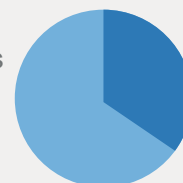
Mapas

Para encontrar patrones geográficos.



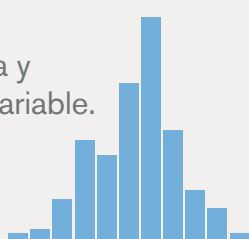
Partes de un total

Para entender la contribución de diferentes categorías en un total.



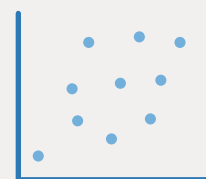
Distribuciones

Para entender la forma y propiedades de una variable.



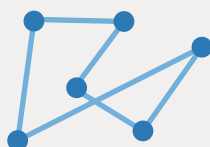
Correlaciones

Para entender la relación entre diferentes variables.



Conexiones, relaciones y redes

Para entender la relación entre los elementos de un conjunto de datos.



Principios de diseño

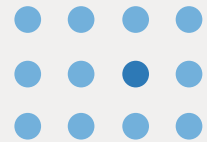
Pensamiento visual

Cómo aprovechar las capacidades cognitivas del ser humano.



Color

Cómo utilizar el color para comunicar datos adecuadamente.



Forma

Qué atributos existen para codificar los datos mediante la forma.



Interacción

Cómo utilizar la interacción para facilitar la exploración y la comprensión.



Recomendaciones generales

Otras consideraciones para desarrollar buenas visualizaciones.



2. Qué es la visualización de datos

La visualización de datos es una disciplina muy reciente y que todavía está en proceso de consolidación. Colin Ware fue uno de los primeros investigadores en ofrecer una definición de visualización de datos lo bastante acotada y, a la vez, lo bastante amplia para no quedar desfasada frente a los constantes avances tecnológicos en el tratamiento y representación de datos. Según Ware, la visualización es:

“la representación gráfica de datos o conceptos, que tiene como resultado una imagen mental o un artefacto externo que ayuda en la toma de decisiones”²

Así pues, la visualización de datos tiene dos componentes principales:

- Es una representación gráfica de datos, pero también de conceptos;
- Ayuda a tomar decisiones.

Aunque para Ware la visualización es la representación gráfica de datos o conceptos, en esta guía se utiliza la palabra visualización para hacer referencia solo a la visualización de datos.

A nivel formal, y de un modo muy general, se pueden distinguir dos tipos de visualizaciones: las estáticas y las interactivas.

Las visualizaciones estáticas se acostumbran a utilizar para comunicar unos datos que se han analizado previamente. Este tipo de visualizaciones ayudan a explicar patrones y tendencias generales.

En la siguiente imagen se puede ver una visualización en la que se muestra la evolución de la tasa de paro en los Estados Unidos en forma de gráfico de líneas.

² Ware, Colin. *Information Visualization, Third Edition: Perception for Design*. Morgan Kaufmann, 2012



Figura 2.1. Evolución de la tasa de paro en los Estados Unidos.

Las visualizaciones interactivas se utilizan para que el usuario pueda interactuar con los datos. Por ejemplo, la visualización interactiva de *The New York Times* que referenciamos a continuación permite que el usuario dibuje su estimación para una serie de indicadores económicos y sociales de los Estados Unidos:

<https://www.nytimes.com/interactive/2017/01/15/us/politics/you-draw-obama-legacy.html>.

Gracias a la interacción, conseguimos centrar la atención del usuario en los datos. En este caso, le hacemos pensar sobre la evolución de ciertos indicadores bajo la presidencia de George Bush y de Barack Obama. Un gráfico estático podría comunicar lo mismo, pero seguramente no conseguiría el mismo grado de implicación del lector.

Por otra parte, la visualización tiene tres objetivos principales:

- Explorar
- Analizar
- Explicar

A continuación, explicamos estos objetivos detalladamente, así como los formatos más adecuados para cada uno de ellos.

2.1. Visualizaciones para explorar

La naturaleza bidireccional de los medios digitales facilita que el usuario pueda interactuar con los datos. Así pues, en vez de explicar una historia con datos, el lector se puede convertir en protagonista. Facilita que el lector interroge los datos, de modo que le ayuden a responder inquietudes particulares y preguntas adaptadas a su contexto. En estos casos, la visualización ayuda a explorar un conjunto de datos.

Así pues, una visualización de datos no solo ayuda a resolver dudas, sino que invita a plantearse preguntas que ni siquiera se habían imaginado. Las visualizaciones explorativas son un buen ejemplo de ello: facilitan la interacción del lector, contribuyen a abrir un debate sobre un tema y responden a muchas preguntas, a la vez que generan otras nuevas.

Las visualizaciones de datos explorativas suelen requerir de un equipo de desarrollo multidisciplinar, capaz de pensar una historia, recoger y analizar los datos, y diseñar interfaces interactivas. Herramientas como Tableau facilitan el desarrollo de aplicaciones interactivas de datos, pero el formato web (mediante librerías como D3) es el más flexible y eficiente, y el que se adapta mejor a diferentes dispositivos.

En el siguiente ejemplo, en vez de ofrecer unas estadísticas medias del rendimiento del jugador de baloncesto Stephen Curry, se optó por mostrar todos los datos de los lanzamientos. El resultado es una aplicación que permite ver la evolución temporal de los puntos que marcó, en qué partidos marcó más, y la localización de estos dentro de la pista. Podemos filtrar por diferentes partidos, momentos del juego o tipo de tiros. Con esta visualización no se pretende responder a ninguna pregunta concreta; simplemente se trata de poder explorar un conjunto de datos.

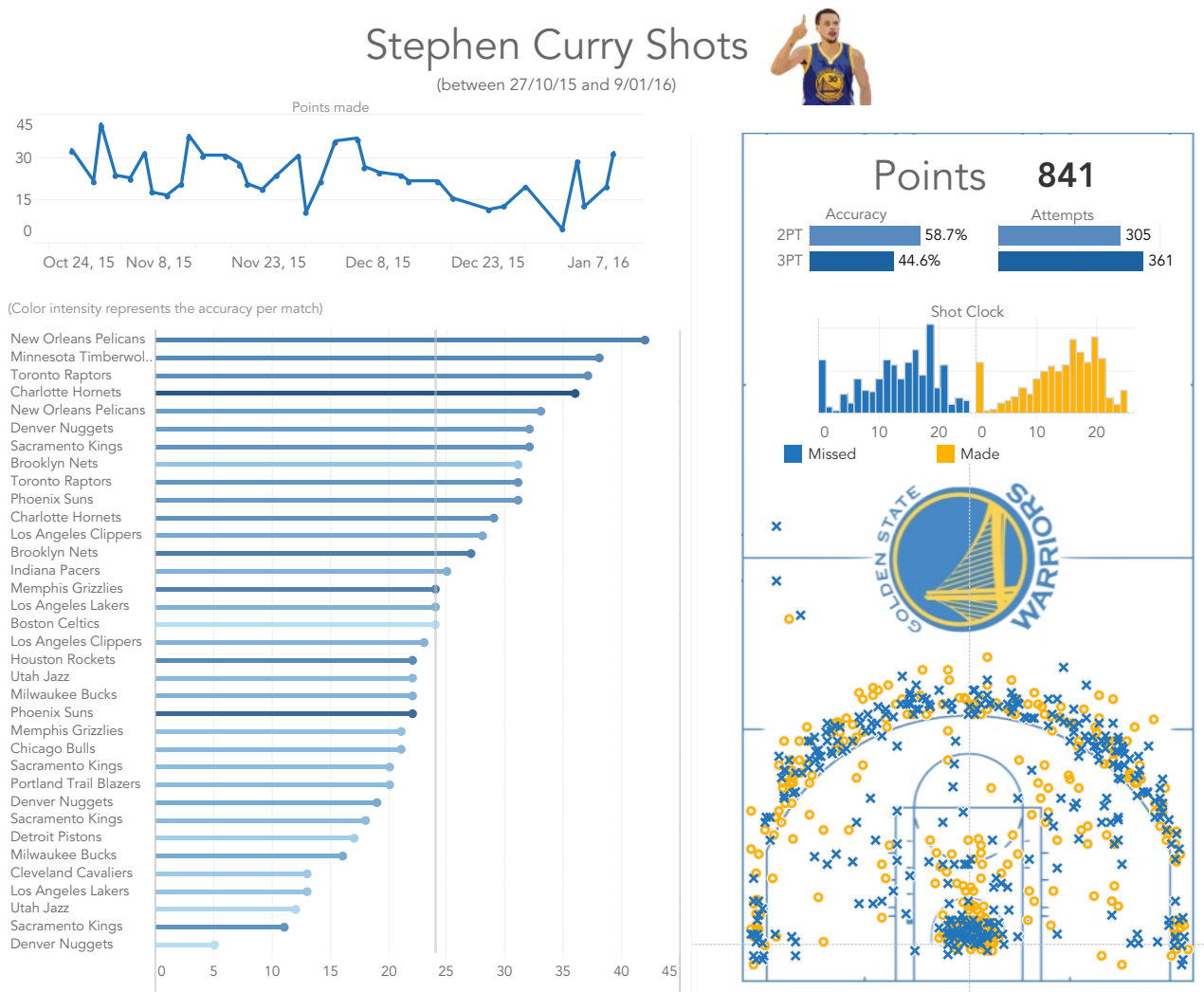


Figura 2.2. Visualización de los lanzamientos efectuados por el jugador Stephen Curry. Imagen cedida por OneTandem.
Fuente: <https://public.tableau.com/profile/onetandem#!/vizhome/curry/StephenCurryShots>

2.2. Visualizaciones para analizar

Además de explicar una historia o de permitir la exploración de un conjunto de datos, las visualizaciones también son útiles al analizar estos datos. Por lo tanto, primero se puede utilizar la visualización para analizar un conjunto de datos y extraer una serie de conclusiones, y después para explicar una historia o permitir su exploración según una serie de parámetros.

Aunque las visualizaciones interactivas han llegado al gran público gracias al periodismo de datos, su uso principal sigue siendo el tratamiento y análisis de problemas complejos, en campos tan diversos como el ejército, la medicina, la biología, las políticas públicas o las redes sociales. Son especialmente útiles cuando la complejidad del problema que debe tratarse o la cantidad de datos excede las capacidades de las visualizaciones y técnicas de análisis tradicionales.

2.2.1. Cuadro de mando

El cuadro de mando es una herramienta con la que se puede controlar el estado de un sistema. Por ejemplo, un cuadro de mando financiero permite controlar los ingresos y los gastos de una organización por diferentes departamentos y segmentos de productos y clientes.

Como característica diferencial, un cuadro de mando va dirigido a usuarios expertos en la temática que cubre, los cuales, al mismo tiempo, son personas con capacidad de alterar el funcionamiento del sistema sobre el cual informa. Por lo tanto, es muy importante entender los objetivos y los conocimientos del usuario que lo utilizará.

2.2.2. Aplicación de analítica visual

Una aplicación de analítica visual requiere de unas capacidades superiores de análisis y exploración respecto de un cuadro de mando.

Hablaremos de aplicaciones de analítica visual siempre que se puedan utilizar técnicas estadísticas para analizar los datos. Por ejemplo:

- Facilitar medidas como la media, la mediana o la correlación;
- Agrupar elementos mediante técnicas de análisis de clústers³;
- Construir modelos predictivos que faciliten responder a qué pasará, y no solo a qué ha pasado, según diferentes escenarios.

El siguiente ejemplo es una aplicación de analítica visual que permite contextualizar los rankings de diferentes universidades, analizar qué factores contribuyen a su posición y qué correlaciones existen entre los diferentes indicadores. No es una visualización adecuada para comunicar un ranking de universidades (para eso habría bastante con un **gráfico de barras**). En cambio, es la adecuada para que una universidad pueda entender por qué ocupa una posición determinada y qué puede hacer para mejorarla.

³ El análisis de clústers es una técnica de análisis de datos con la que se puede segmentar, automáticamente, un conjunto de datos en subgrupos (denominados clústers) que tengan valores similares a sus variables.



Figura 2.3. Aplicación analítica para analizar *rankings* de universidades. Imagen cedida por la empresa SIRIOS Academic.
Fuente: <http://university-analytics.com/rankings/#/explorer/ARWU/2017>.

2.3. Visualizaciones para explicar

Internet, las redes sociales y los datos abiertos han democratizado el acceso a los datos. En este contexto, han surgido iniciativas como el periodismo de datos, enfocado a generar noticias utilizando precisamente los datos como fuente principal de información.

La visualización es vital para comunicar y ayudar a entender adecuadamente los datos en el contexto de una historia o de un propósito determinado. No sirve de nada un buen análisis si no se sabe explicar. Pero lo contrario también es cierto. Se puede crear una visualización de datos formalmente perfecta, pero si el análisis que se hace o la historia que se explica no tienen importancia en un contexto determinado, la visualización tampoco servirá de nada. Asimismo, debe tenerse presente que la visualización puede aclarar el significado de los datos, pero también puede servir para engañar al lector. Es por eso que tener conocimientos de visualización de datos es indispensable si se aspira a entender mejor el mundo y la sociedad que nos rodean.

Por otra parte, la visualización de datos permite que destacar como comunicador en un contexto de saturación informativa. Hay más datos que nunca, pero no se dedican bastantes esfuerzos para refinarlos y ofrecer una información relevante y de calidad. La metodología de la visualización de datos nos obliga a pensar qué se quiere comunicar con ellos, qué es importante y cómo se tiene que hacer. Al mismo tiempo, se consigue comunicar los contenidos de un modo visualmente más atractivo y en formatos más adecuados para propagarlos por las redes sociales.

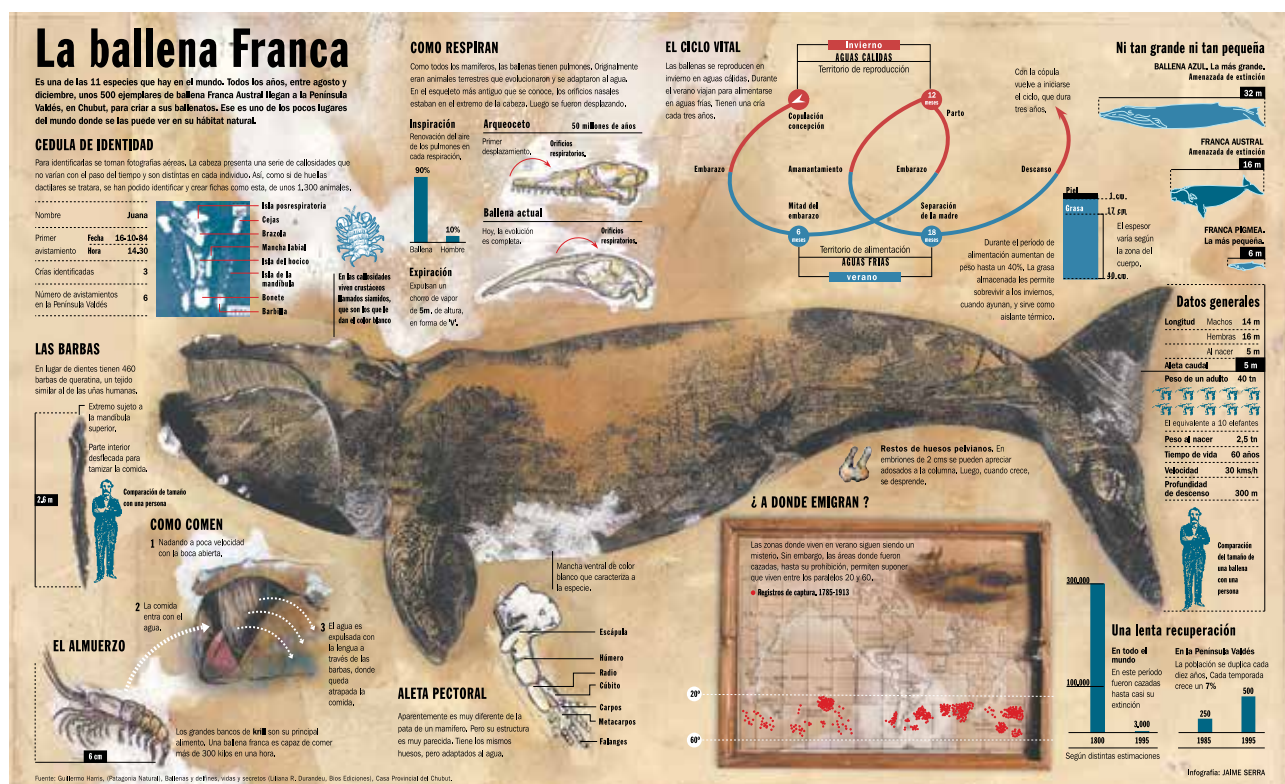
A continuación, se presenta una introducción a los principales formatos que utiliza la visualización para explicar historias basadas en datos.

2.3.1. Infografía

Las infografías se popularizaron en revistas y diarios como una forma de comunicar temas complejos al público en general. Con las redes sociales, las infografías han incrementado su audiencia de manera exponencial por tres motivos:

- El componente visual ayuda a que destaquen y eso hace que sean más compartidas y que se propaguen rápidamente por las redes;
- Están diseñadas para ser consumidas en poco tiempo y, por lo tanto, se adaptan a las pautas de consumo rápido de información de las redes sociales;
- No suelen requerir conocimientos previos del tema tratado y, por lo tanto, se dirigen a un público amplio.

Un muy buen ejemplo de infografía es el siguiente trabajo sobre la ballena, de Jaime Serra:



2.3.2. Narrativa por desplazamiento

La narrativa por desplazamiento (*scrollytelling*) es una forma de narración interactiva donde el usuario debe desplazarse (utilizando la barra de desplazamiento, con el ratón o deslizando el dedo en un dispositivo táctil) para obtener más información. Se utiliza para explicar historias desde múltiples puntos de vista y/o con múltiples elementos, como gráficos, vídeos, animaciones, fotografía y texto. La interacción facilita que la historia aparezca en el orden requerido y que se disfrute de la atención del usuario.

Un buen ejemplo de este formato es la narrativa por desplazamiento hecha por Bloomberg, que explica los motivos del calentamiento global. Esta narrativa se puede consultar en: <https://www.bloomberg.com/graphics/2015-whats-warming-the-world/>.

La narrativa por desplazamiento es un formato exclusivamente pensado para páginas web, generalmente asociado a grandes medios de comunicación digitales. Se puede enmarcar dentro de la tradición de la narrativa interactiva⁴. Su elaboración requiere muchos recursos porque necesita la colaboración de perfiles muy variados: guionistas, diseñadores, programadores, fotógrafos, realizadores, etc.

2.3.3. Presentación

Las presentaciones con herramientas como PowerPoint o similares se han convertido en un formato habitual de comunicación en las organizaciones. Como suelen presentar la información de un modo conciso, las visualizaciones de datos juegan un papel importante.

En general, se recomienda incorporar solo una idea en cada diapositiva. Esta idea se puede transmitir con una frase, una imagen o una visualización de datos, y debe acompañar el discurso de un modo natural. La regla de limitarnos a una idea por diapositiva nos invita a introducir la visualización de datos en diferentes pasos.

A continuación se reproduce el caso presentado por Cole Nussbaumer, especialista en *storytelling* con datos, en su libro y blog *Storytelling with data*⁵ (<http://www.storytellingwithdata.com/>).

⁴ Wikipedia, Interactive Storytelling. <https://ca.wikipedia.org/wiki/Projecci%C3%B3_cartogr%C3%A0fica> [Consulta: no viembre 2017]

⁵ Nussbaumer, Cole. *A Data Visualization Guide for Business Professionals*. Wiley, 2015

En el caso que se analiza, se ve la evolución del número de usuarios activos de una aplicación de móvil denominada Moonville. La forma habitual de presentar los datos es la siguiente:

Moonville: active users over time



Figura 2.5. Visualización utilizada para mostrar la evolución de los usuarios activos de una aplicación móvil. Imagen cedida por Cole Nussbaumer.

Para comunicar los datos con una presentación en PowerPoint o similar, Cole Nussbaumer propone descomponer el gráfico en una serie de diapositivas. En primer lugar, empezaría con esta imagen para situar el problema:

Moonville: active users over time



Figura 2.6. Imagen inicial recomendada para explicar el problema. Imagen cedida por Cole Nussbaumer

Después, se iría presentando la evolución con la siguiente serie de diapositivas, cada una acompañada del discurso pertinente:

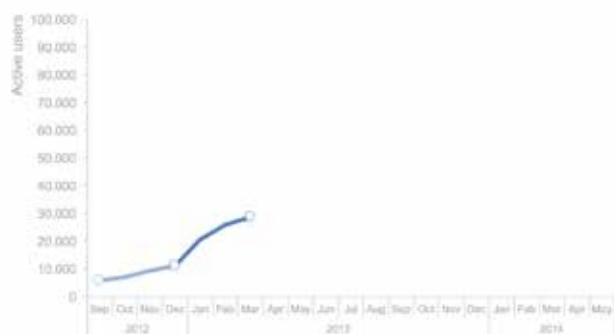
Moonville: active users over time



Moonville: active users over time



Moonville: active users over time



Moonville: active users over time



Moonville: active users over time



Figura 2.7. Serie de imágenes utilizadas para explicar y presentar la evolución de los usuarios activos en una aplicación móvil. Imágenes cedidas por Cole Nussbaumer.

Finalmente, se puede presentar una transparencia que resuma todo el discurso. Esta será la diapositiva que se puede incluir en un fichero PowerPoint (o similar) que se envíe a la audiencia por correo electrónico o que se entregue impresa.

Moonville: active users over time

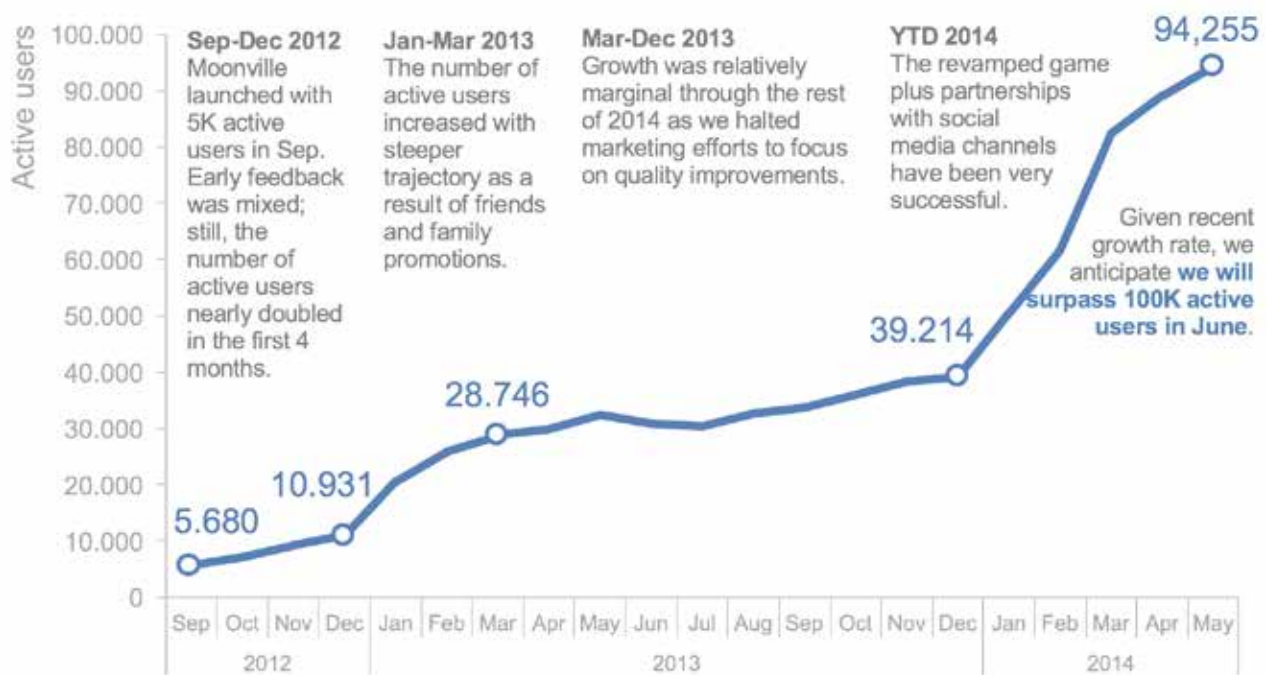


Figura 2.8. Visualización final que resume todo el discurso pronunciado previamente. Imagen cedida por Cole Nussbaumer.

Con este ejemplo, se muestra que el reto de una presentación es doble: conocer el tema a tratar y saber explicarlo. Por eso, además de ser buenos diseñadores y oradores, es vital conocer a la audiencia. Por ejemplo, si se prepara una presentación para una clase, hace falta averiguar qué quieren aprender los alumnos y cuál es el nivel de conocimientos sobre el tema que se tratará. Se puede diseñar una presentación muy completa, pero si no responde al nivel y objetivos de la audiencia, no funcionará. Un error habitual es utilizar terminología que la audiencia no conoce o bien visualizaciones de datos demasiado complejas.

2.3.4. Vídeo

El vídeo va por el camino de convertirse en el formato por excelencia de la red. La visualización de datos, tan vinculada a Internet, no es ajena a este hecho. Por ejemplo, los gráficos en movimiento (*motion charts*) se han convertido en un mecanismo habitual para transmitir datos de forma animada. Los GIF también ocupan un espacio importante en redes de consumo rápido de información como Twitter. Asimismo, numerosos canales de Youtube y medios digitales apuestan por formatos de vídeo que incorporan datos e infografías de forma innovadora.

Un medio que apuesta por este formato es Vox (<https://www.vox.com/>). En su canal de Youtube (<https://www.youtube.com/user/voxdotcom>) se puede ver una buena muestra de vídeos que hacen uso de la visualización de datos, como por ejemplo el titulado “The real reason American health care is so expensive” https://www.youtube.com/watch?v=k1vE_LVBx4s.

3. Metodología

En este capítulo se describe el proceso para desarrollar una visualización de datos, que consta de tres fases:

- Fase 1: estrategia
- Fase 2: datos
- Fase 3: diseño

3.1. Fase 1 - Estrategia

La estrategia hace referencia a todo aquello que nos conduce a definir los objetivos de la visualización de datos. En particular, se tratan los siguientes puntos:

- Investigación y análisis
- Objetivos
- Indicadores

3.1.1. Investigación y análisis

Para elaborar una visualización sobre un tema en particular debe conocerse en profundidad. Así se puede tratar desde diferentes ángulos y asegurar que se dispone de los datos adecuados para mostrar un buen análisis. Por eso, a menudo debe hacerse investigación del tema sobre el que se quiere tratar. Para hacerlo adecuadamente, conviene investigar los siguientes aspectos:

- Temática
- Audiencia
- Análisis

Temática

Todo proyecto de visualización de datos gira en torno a una temática en particular. Por ejemplo, si se quieren visualizar datos sobre la contaminación en Cataluña, se puede definir la temática como el medio ambiente.

El primer paso es plantear la temática desde diferentes puntos de vista. El esquema 5W1H⁶ puede ayudar. Aplicado a la contaminación, este esquema plantearía las siguientes preguntas:

- Qué – ¿Qué quiere decir tener un problema de contaminación?
- Quién – ¿Quién es responsable / Quién la sufre?
- Dónde – ¿Dónde se sufre más la contaminación? ¿Dónde han arreglado el problema?
- Cuándo – ¿Cuándo se sufre más la contaminación? ¿Qué evolución ha habido?
- Por qué – ¿Por qué varían los niveles de contaminación? ¿Por qué pueden suponer un problema?
- Cómo – ¿Cómo se puede reducir / aumentar la contaminación? ¿Cómo se regula?

Una vez definidos uno o más ángulos para tratar la temática, es recomendable consultar publicaciones y expertos en esta área, así como otras visualizaciones de datos que ya se hayan hecho, para asegurarse de que no se repite el mismo trabajo.

Finalmente, es útil encontrar testimonios. Por ejemplo, en el caso de la contaminación sería muy enriquecedor incorporar la opinión de personas afectadas por el problema. Los datos ayudan a cuantificar un problema, pero los testimonios personales ayudan a hacerlo más próximo y, por lo tanto, aumenta la probabilidad de causar un mayor impacto en el consumidor de la visualización.

Audiencia

Por audiencia se entiende quién utilizará la visualización de datos. Diseñar una buena visualización de datos es similar al diseño de una aplicación y, por lo tanto, saber cuáles son los conocimientos y las motivaciones del usuario es vital para asegurar el éxito de la visualización.

Para conocer a la audiencia, se recomienda responder a estas preguntas:

Relación

- ¿La conocemos personalmente?
- ¿Nos conoce?
- ¿Es homogénea o heterogénea?

Conocimientos sobre el tema

- ¿Nos reconoce como autoridad en el tema?
- ¿Qué conocimientos tiene del tema?
- ¿Qué experiencia tiene en el uso de los datos?

⁶ Wikipedia, Five Ws. <https://en.wikipedia.org/wiki/Five_Ws> [Consulta: 20 noviembre 2017]

Contexto de uso

- ¿Qué dispositivo o dispositivos utilizará para consultar la visualización?
- ¿Qué objetivos tiene?
- ¿Qué acciones esperamos que lleve a cabo?
- ¿Cuánto tiempo tiene para consultar la visualización?
- ¿Por qué canales llegaremos a la audiencia?

En definitiva, se trata de entender el modelo mental de los usuarios para poder diseñar productos adaptados a sus objetivos, capacidades y contexto de uso.

Análisis

Otro punto a tener en cuenta a la hora de definir la estrategia es el análisis de los datos. Y es que en el proceso de exploración y análisis de un conjunto de datos, a menudo se encuentra la idea de una visualización. También puede pasar que los datos demuestren que la idea inicial no es válida o que es irrelevante.

Para este primer análisis, se recomienda hacer las siguientes exploraciones preliminares:

- Distribución de los principales indicadores, con histogramas y diagramas de caja: eso ayudará a encontrar valores atípicos, que a menudo esconden historias interesantes.
- Tendencias de los principales indicadores: se detectará si hay grandes cambios en determinados momentos.
- Correlación entre diferentes indicadores: la mente tiende a establecer relaciones entre fenómenos, por lo tanto, las historias que expliquen relaciones funcionan muy bien.

3.1.2. Objetivos

Ahora que ya se conocen la temática y la audiencia, es el momento de definir los objetivos de la visualización de datos. En general, se dice están definidos si se pueden responder claramente las siguientes preguntas:

- ¿Cómo queremos ayudar al usuario?
- ¿Por qué?
- ¿Cómo lo evaluaremos?

Para hacerlo más comprensible, conviene imaginar qué se plantea, por ejemplo, una visualización de datos sobre la contaminación en Cataluña.

¿Cómo queremos ayudar al usuario?

Tal y como se ha explicado en el capítulo de introducción, las visualizaciones de datos pueden ayudar a explorar, analizar y/o explicar un tema. Por lo tanto, lo primero que conviene decidir es cómo se ayuda al usuario, es decir:

- ¿Se quiere ayudar a explorar unos datos sobre un tema?
- ¿Se quiere ayudar a analizar los datos sobre un tema?
- ¿Se quiere explicar un tema?

En el ejemplo de la visualización sobre la contaminación en Cataluña, se podría responder de tres maneras bien distintas a la pregunta “¿Cómo se quiere ayudar al usuario?”:

- Se quieren facilitar al usuario los datos de la evolución de la contaminación por comarca, para que sea capaz de explorar cómo han variado en la zona donde vive y trabaja.
- Se quieren facilitar datos sobre contaminación, densidad de población, industria y tráfico, así como mecanismos para establecer correlaciones entre diferentes indicadores, para poder analizar qué factores contribuyen más a la presencia de contaminación en diferentes zonas del país.
- Se quiere explicar cuánto ha disminuido la contaminación en Cataluña en los últimos diez años y cuáles han sido los principales motivos.

Las tres respuestas pueden dar lugar a buenas visualizaciones de datos, pero seguramente no se podrán resolver con una misma visualización de datos. Es decir, se puede enfocar un mismo tema desde múltiples puntos de vista, y para cada punto de vista habrá diferentes maneras de visualizar los datos.

¿Por qué?

Esta pregunta pretende hacernos reflexionar sobre la importancia y el alcance de lo que se está haciendo y, por lo tanto, los recursos que se tendrían que dedicar. También provoca que se empiece a pensar en el formato y los canales de comunicación.

Volviendo al caso de la contaminación en Cataluña, se podrían definir los siguientes *porqués*:

- Sensibilizar a la población sobre los efectos nocivos de la contaminación para la salud;
- Ayudar a los técnicos medioambientales a planificar mejor las políticas de actuación.

En el primer caso, el destinatario es la población en general, que no tiene conocimientos avanzados de análisis e interpretación de datos. Por lo tanto, una serie de visualizaciones en formato GIF acompañadas de una campaña en redes sociales, seguramente serán más efectivas que una sofisticada visualización interactiva. En cambio, en el segundo caso, sería mejor desarrollar una visualización interactiva con múltiples indicadores y herramientas de análisis que faciliten la interpretación de los datos de la mejor manera posible, de acuerdo con las competencias de un técnico.

¿Cómo se evaluará?

Se trata de definir unas métricas para evaluar si se ha conseguido ayudar al usuario adecuadamente, de modo que se alcancen los objetivos planteados a la pregunta “¿por qué?”.

Siguiendo con el ejemplo de antes, si se quiere sensibilizar a la población sobre los efectos de la contaminación, se puede medir el porcentaje de ciudadanos capaces de responder a una serie de preguntas sobre la contaminación antes y después de consultar la visualización de datos.

3.1.3. Indicadores

A menudo, las visualizaciones de datos no alcanzan los objetivos planteados porque tratan unos datos que no son relevantes o que no captan todos los ángulos del problema. Por eso es vital definir unos indicadores adecuados. En general, se pueden definir tres tipos de indicadores:

- Volumen
- Calidad
- Contexto

Por ejemplo, el número de accidentes de tráfico (indicador de volumen) no es un buen indicador para compararlo entre provincias, aunque se haga con un gráfico adecuado como el siguiente:



Figura 3.1. Número de accidentes por provincia.

En cambio, un indicador como la tasa de accidentes por habitante (indicador de calidad) es más adecuado, ya que es independiente de la población de la comunidad y, por lo tanto, se pueden comparar. En el gráfico siguiente se puede ver que Ceuta y Melilla pasan a estar en la primera posición de la clasificación y que, por ejemplo, las islas Baleares se sitúan por encima de Madrid.

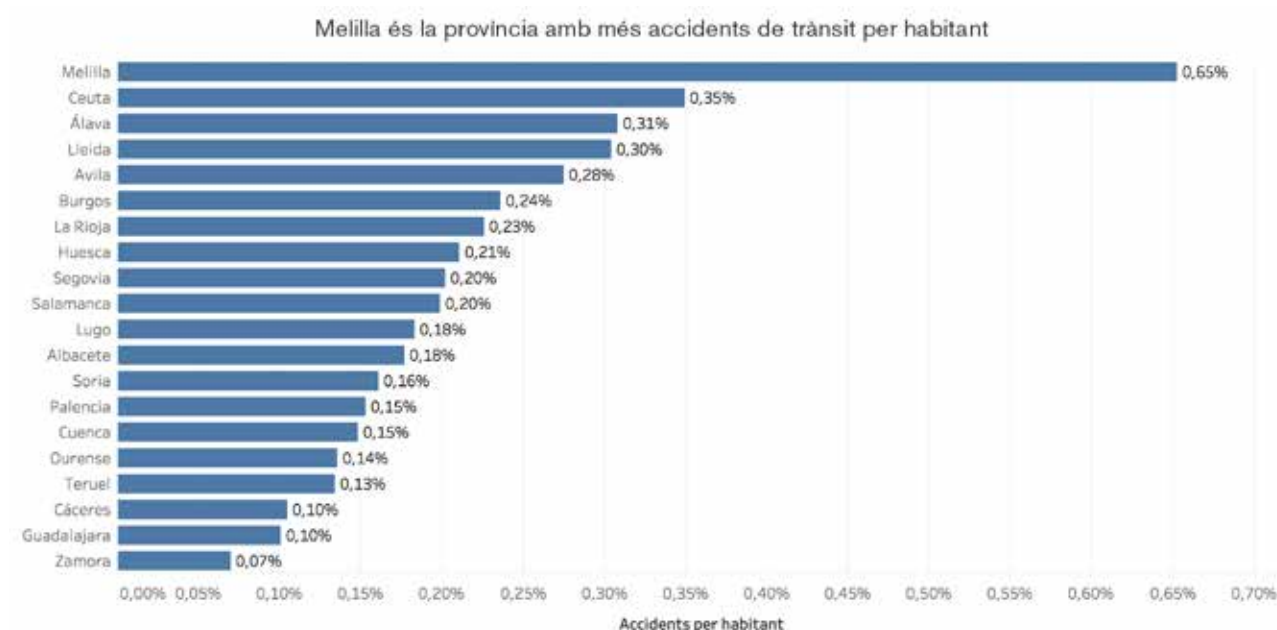


Figura 3.2. Número de accidentes por habitante y provincia.

Finalmente, disponer de la tendencia de la tasa de accidentes en los últimos años (indicador contextual) ayudará a interpretar los indicadores. Por ejemplo, una tasa del 0,2% sería positiva si hace cinco años era del 0,5%, pero negativa si era del 0,1%.

Además de utilizar los indicadores adecuados, conviene tener presente el nivel de agregación de los datos. Por ejemplo, si se quiere analizar la tendencia anual de los accidentes de tráfico, conviene decidir si se quieren mostrar los datos por horas, por días o por meses. Cada nivel de agregación nos permitirá llegar a diferentes conclusiones y requerirá un tipo de gráfico diferente.

En resumen, la fase de estrategia finaliza cuando:

- Se conoce el tema a tratar y la audiencia a la que se dirige.
- Se define cómo se quiere ayudar al usuario y por qué.
- Se definen unos indicadores adecuados (volumen, calidad, contexto).

Se recomienda tener siempre presentes estos puntos durante el posterior proceso de diseño de la visualización de datos, para asegurar que se alcanzan los objetivos marcados inicialmente.

3.2. Fase 2 - Datos

Aunque no es el objetivo de esta guía hacer una introducción al tratamiento de los datos, conviene ser consciente de que en muchos casos se tienen que trabajar para poder disponerlos en un formato y estructura adecuados.

A menudo se dice que la información son datos puestos en contexto o datos organizados y estructurados.

A continuación se mencionan los pasos que debe tenerse en cuenta.

3.2.1. Obtención

El primer paso para poder trabajar una visualización será obtener los datos necesarios, definidos durante el proceso de estrategia. A menudo éstos ya vienen dados, pero otras veces hace falta buscarlos y extraerlos de bases de datos, o implementar sistemas que los generen.

3.2.2. Formateo y limpieza

Una vez que se dispone de los datos, debe validarse el contenido, asegurarse de que no contienen errores y que el formato es adecuado y consistente para su consumo. A menudo contienen errores o valores poco coherentes que conviene revisar. Los valores decimales expresados con comas y al mismo tiempo con puntos son un ejemplo típico.

Será necesario, pues, en la medida de lo posible, repasar cada uno de los registros de nuestros datos para validar que todos los valores están bien escritos y tienen sentido.

Se puede hablar de dos tipos de formatos:

- **Formato general de los datos:** hay una infinidad de formatos que expresan datos de manera estructurada. Las hojas de cálculo, que contienen filas y columnas, son un ejemplo de organización de datos que se pueden expresar en diferentes formatos (XLS, CSV, etc.). Otro formato bastante utilizado, pero al mismo tiempo más avanzado, es el formato JSON. Se escoge el formato de los datos en función de los que se dispone y en función de las opciones de importación de las que disponga el software que se utilice para trabajar
- **Formato de cada uno de los indicadores:** es importante comprobar que cada uno de los indicadores sigue un formato consistente y, preferiblemente, estándar. Por ejemplo, si hay una columna que expresa una fecha, es importante que todos los valores de esta sigan el mismo formato. Por ejemplo, día/mes/año (15/10/1981). En este caso, lo más importante es que haya consistencia, de modo que todos los valores del indicador estén expresados de la misma forma.

3.2.3. Procesamiento

El procesamiento de los datos es, por si mismo, una disciplina propia. En función del volumen y de lo que se quiera analizar, se podrán ejecutar algoritmos muy complejos que podrán dar mucho más valor a los datos. No obstante, normalmente conviene disponer de grandes conocimientos de matemáticas, estadística y programación para poder implementarlos y utilizar.

Finalmente, una vez acabada la visualización (una vez efectuado el proceso de diseño descrito en la siguiente sección), conviene ser conscientes de que es importante ser metódicos a la hora de almacenar nuestros datos. En grandes instituciones es deseable disponer de un repositorio de datos único y estándar donde se puedan guardar los archivos y documentarlos.

En cualquier caso, es recomendable guardar los ficheros finales donde se encuentren los datos en un lugar fácilmente identificable. Además, también es una buena idea acompañarlos con un pequeño documento de texto que describa los datos, su procedencia, cómo se han procesado y la fecha de obtención. Estos ficheros serán muy útiles si se quieren reutilizar los datos en el futuro o si eventualmente se quieren hacer públicos.

3.3. Fase 3 - Diseño

El diseño hace referencia a todo aquello que nos permite conseguir los objetivos definidos en la estrategia. El resultado es el producto final, que debe evaluarse en función de cómo ayuda a cumplir estos objetivos.

El proceso de diseño que se presenta es iterativo y basado en la metodología de diseño centrado en el usuario. En particular, se siguen estas fases:

- Esbozar
- Prototipar
- Finalizar

3.3.1. Esbozar

Esta fase consiste en elaborar borradores con el objetivo de descubrir maneras de representar los datos de acuerdo con la estrategia definida. También es habitual referirse con el término inglés *sketching*.

En particular, la fase de esbozo es muy útil para definir el aspecto de la visualización de datos, que debe ser coherente con los objetivos definidos en la estrategia. Cuando se hacen los primeros borradores, se decide el tipo de gráficos que se utilizarán, qué textos deben acompañarlos y cómo se ha de maquetar.

Generalmente, en esta fase se trabaja con papel y bolígrafo para tener el máximo de libertad a la hora de pensar la visualización de datos. Asimismo, se recomienda trabajar con libretas de diferentes medidas, que ayuden a pensar diseños para pantalla de ordenador, para tableta y para móvil.

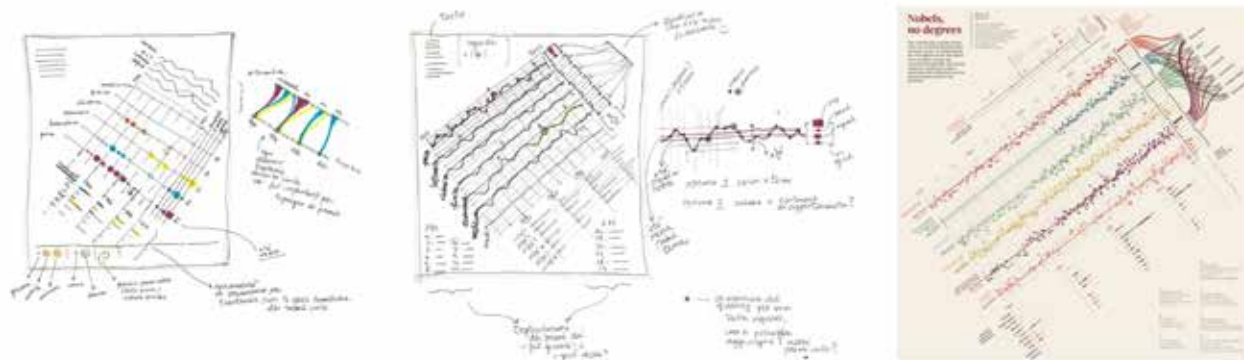


Figura 3.3. Esbozos presentados en el artículo "Sketching with Data Opens the Mind's Eye", de Giorgia Lupi. Imagen cedida por la autora. <https://medium.com/accurat-studio/sketching-with-data-opens-the-mind-s-eye-92d78554565>

3.3.2. Prototipar

En esta fase se utilizan los datos reales para ver si los esbozos siguen siendo válidos. A veces, algunas visualizaciones escogidas no son las adecuadas. En este caso, conviene volver a trabajar con papel y bolígrafo para obtener nuevas ideas. Es decir, durante el prototipaje, se alterna el uso de papel y bolígrafo con el uso de herramientas digitales de prototipaje para trabajar con datos reales.

Una vez que se tiene una primera versión del prototipo, debe presentarse al usuario final con el objetivo de validar si responde a sus necesidades. Si no se conoce el usuario final, conviene escoger a un grupo de usuarios representativos y hacer un test de usuario parecido a los que se hacen para evaluar la usabilidad de una aplicación. Se considera que un test con 5 usuarios ayuda a detectar gran parte de los problemas de un diseño.

Es importante centrar el test con usuarios en los objetivos del diseño. Por ejemplo, si se pregunta "¿qué te parece este cuadro de mando?", lo más habitual será que el usuario hable de si le gusta o no, basándose en su experiencia particular. La estética es lo más fácil de juzgar, pero no es el elemento más importante de un prototipo. En cambio, se recomienda recordar al usuario los objetivos de la visualización de datos que se le presenta y hacerle preguntas para averiguar si es capaz de alcanzar los objetivos marcados.

Otra técnica consiste en presentar la visualización de datos al usuario y dejarlo interaccionar durante un tiempo, sin ninguna interferencia por nuestra parte. Finalmente, se le pregunta qué ha aprendido con la visualización y cómo la ha utilizado para entender qué funciona y qué no funciona.

Una vez que se conocen los aspectos a mejorar, se elabora un nuevo prototipo y se repite el test con otros usuarios. Es recomendable hacer un par de iteraciones, antes de pasar a la fase final.

3.3.3. Finalizar

En esta fase se transforma el prototipo en el producto final. Es el momento de tratar aspectos como:

- La redacción de textos finales
- La adaptación a la guía de estilo requerida (colores, tipografías, espacios, logotipos, etc.)
- La elaboración de mecanismos de ayuda

Además, una vez que el producto está acabado, conviene hacer una revisión de calidad, es decir:

- Asegurar que el diseño funciona correctamente en todos los dispositivos donde habrá que utilizarlo, con diferentes resoluciones y medidas de pantalla.
- Comprobar cada función y asegurarse de que no da ningún error (particularmente, los filtros y otros elementos de interacción).
- Comprobar que los datos son correctos y fijarse especialmente en que no aparezcan resultados inesperados para determinadas combinaciones de filtros o elementos interactivos.

Aunque ya se haya finalizado el diseño, todavía no se habrá utilizado en entornos reales. Por lo tanto, es probable que surjan nuevos problemas y que hagan falta otras mejoras. También es probable que los requerimientos iniciales puedan cambiar al cabo de un tiempo. Es por eso que se recomienda tener presente que un diseño nunca será definitivo y que deben reservarse recursos para adaptarlo a diferentes circunstancias.

Para asegurarse de que el diseño evoluciona, se recomienda lo siguiente:

- Realizar encuentros periódicos con el usuario o usuarios de la visualización de datos.
- Analizar periódicamente las métricas de rendimiento y el uso de la visualización de datos.

4. Gráficos

Para visualizar datos, ya sea para explorarlos, analizarlos o explicarlos, hay infinitas posibilidades. Sin embargo, hay un conjunto de gráficos considerados estándar, como por ejemplo los gráficos de barras o los gráficos de líneas. En este capítulo, se clasifican estos gráficos en 7 categorías diferentes, en función de lo que se quiera conseguir:

- comparaciones
- tendencias
- mapas
- partes de un total
- distribuciones
- correlaciones
- conexiones, relaciones y redes

A continuación, se presentan las diferentes visualizaciones de cada categoría. Sin embargo, conviene tener en cuenta que siempre puede haber varias opciones válidas para representar un conjunto de datos. Se tendrá que valorar qué se quiere representar y cómo son los datos para acabar de escoger cuál es la mejor visualización a utilizar.

4.1. Comparaciones

Debe escogerse un gráfico con el que se puedan hacer comparaciones cuando se quieren comparar diferentes variables entre sí, cuando queramos comparar el valor de una o más categorías entre ellas, o cuando se quieren comparar valores con el fin de establecer un *ranking* (es decir, cuando se quiere saber qué elemento de un conjunto de datos es el primero, cuál es el segundo y qué diferencia de valor hay entre los unos y los otros).

Gráfico de barras

Para comparar un conjunto de valores y establecer un ranking.

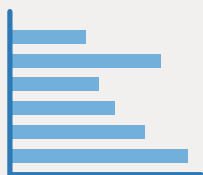


Gráfico de barras agrupadas

Para comparar el valor de los diferentes segmentos que forman parte de nuestras categorías.

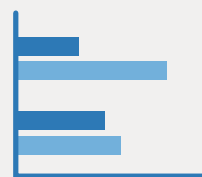


Gráfico de barras apiladas

Para comparar el valor total de la suma de los segmentos que forman cada una de las barras.

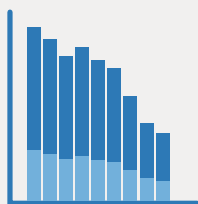


Gráfico de radar

Para entender los valores de un elemento en diferentes variables a la vez.



Gráfico de intensidades de colores

Para descubrir la variabilidad de un conjunto de variables.

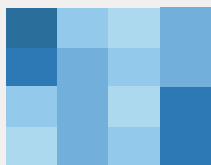


Gráfico de marcas

Para evaluar un valor concreto respecto al que se esperaba y respecto a unos rangos de calidad establecidos.

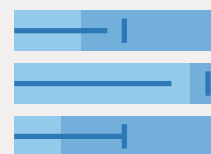
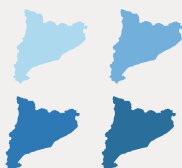


Gráfico múltiple

Para representar datos de diferentes categorías que no se pueden ver correctamente en un único gráfico.



4.1.1. Gráfico de barras

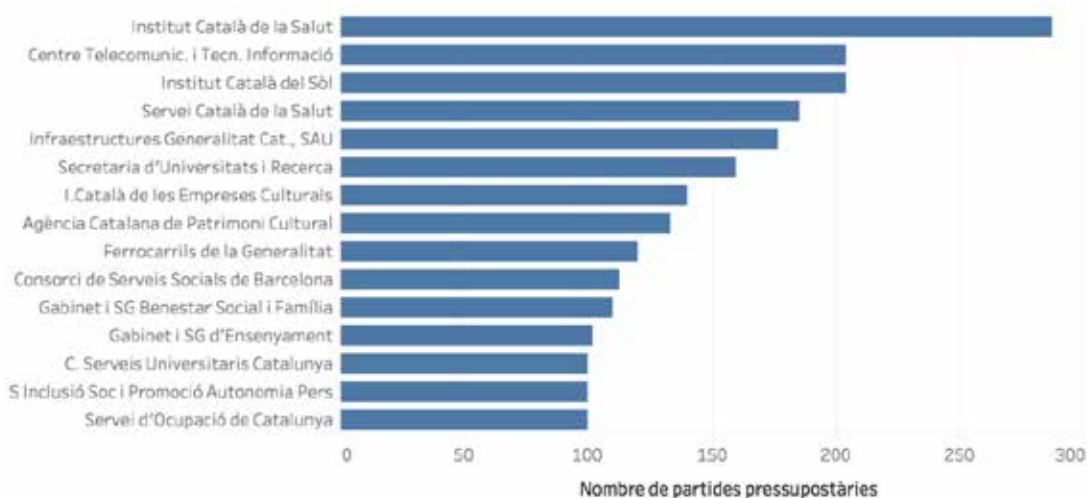


Figura 4.1. Ejemplo de los 15 servicios con mayor número de partidas presupuestarias durante el 2015.

Fuente: Presupuestos de la Generalitat de Catalunya 2015

(http://economia.gencat.cat/ca/70_ambits_actuacio/pressupostos/els-pressupostos-de-la-generalitat-de-catalunya-per-al-2015/).

Este gráfico está formado por dos ejes: uno cuantitativo, que muestra la escala de los valores que se representan; y uno textual, que representa la categoría a la cual pertenecen los datos representados. En el eje textual se sitúan un conjunto de barras, cuya longitud codifica el valor de cada categoría.

¿Cuándo se debe utilizar?

Cuando se quiere comparar un conjunto de valores y establecer un *ranking*. Como el ojo humano es muy bueno comparando distancias de elementos que están situados sobre un mismo eje, el gráfico de barras no solo descubre los valores más altos y los más bajos, sino que también ayuda a tener una intuición bastante acertada de la diferencia existente entre unos valores y otros.

Recomendaciones

Se recomienda ordenar las categorías de mayor a menor, o al revés. Eso facilita la comunicación del *ranking* de categorías.

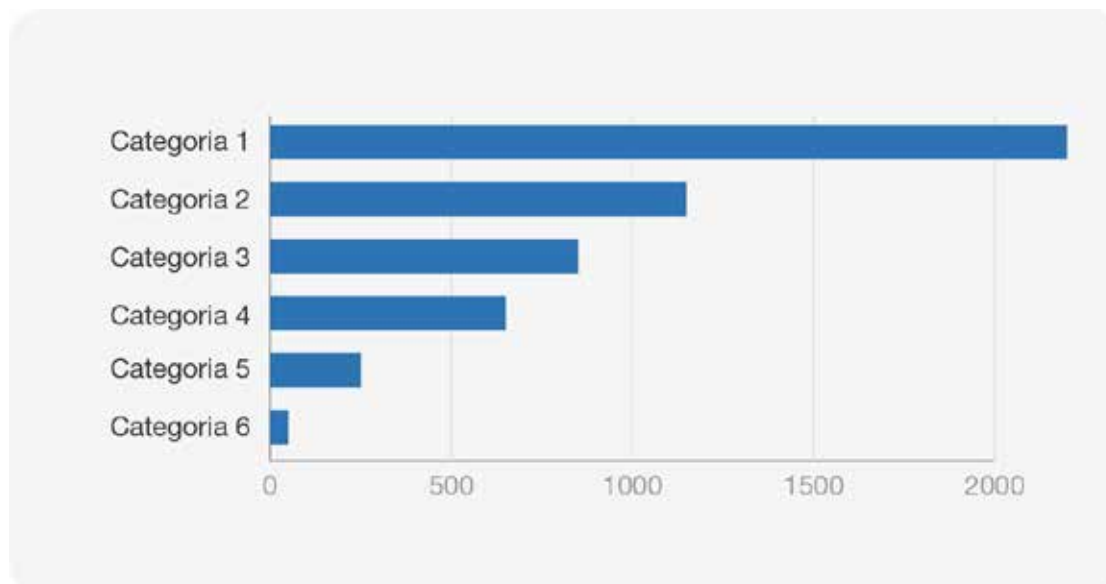


Figura 4.2. Ejemplo de un gráfico de barras ordenado para facilitar la comunicación del ranking de las categorías.

Cuando se representan datos temporales (y las categorías son horas, días, meses, etc.) conviene ordenar el eje de forma temporal. En este caso, se recomienda la disposición vertical de las barras.

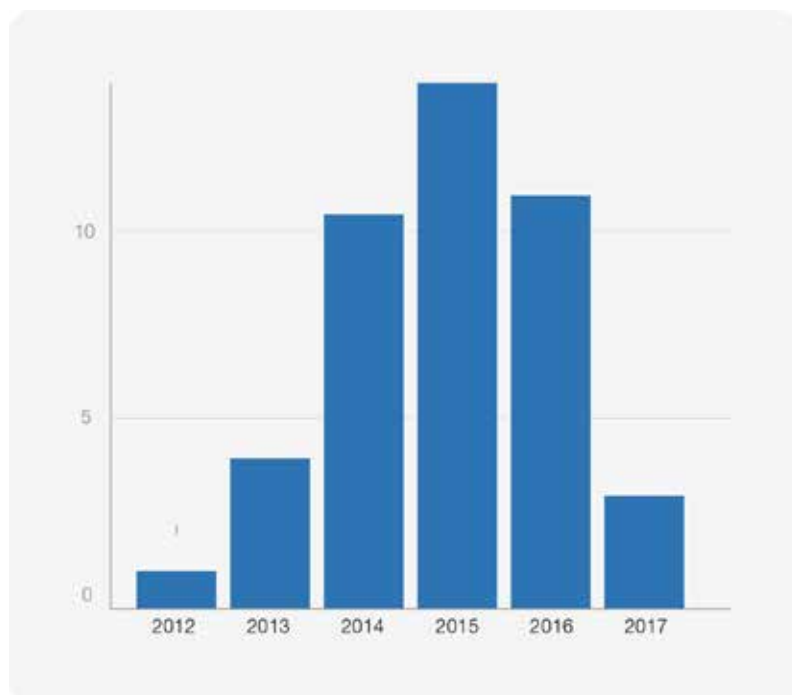


Figura 4.3. En caso de disponer de categorías que tengan un orden temporal, será mejor utilizar el gráfico de barras en vertical.

El eje cuantitativo debe empezar siempre en el cero, porque, si no, las diferencias de valores se magnifican y confunden al lector.

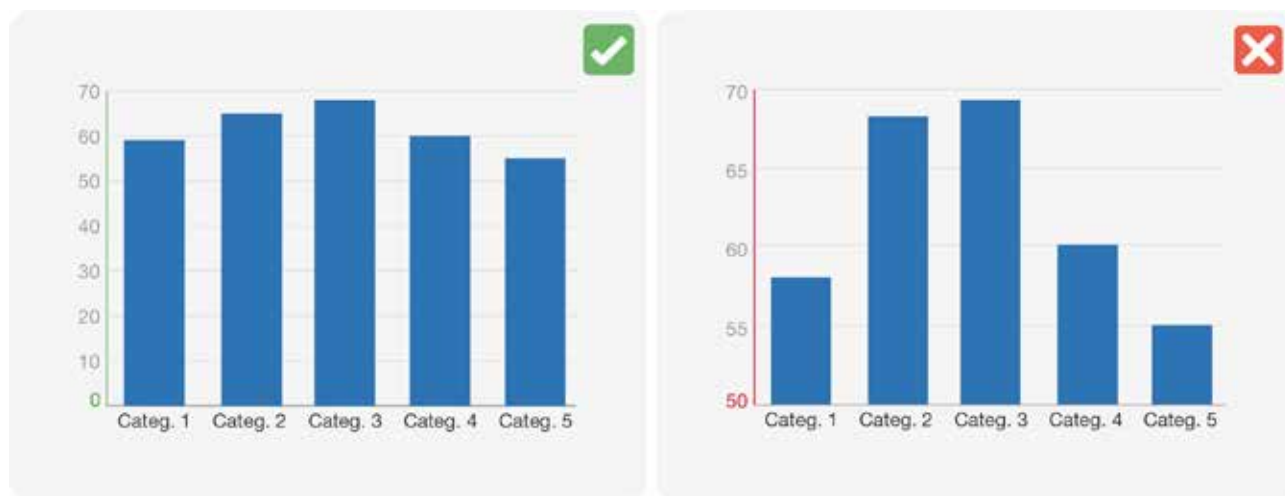


Figura 4.4. El eje cuantitativo de los gráficos de barras siempre debe empezar en el 0.

Cuando los nombres de las categorías son largos, es conveniente orientarlo horizontalmente para facilitar la lectura.

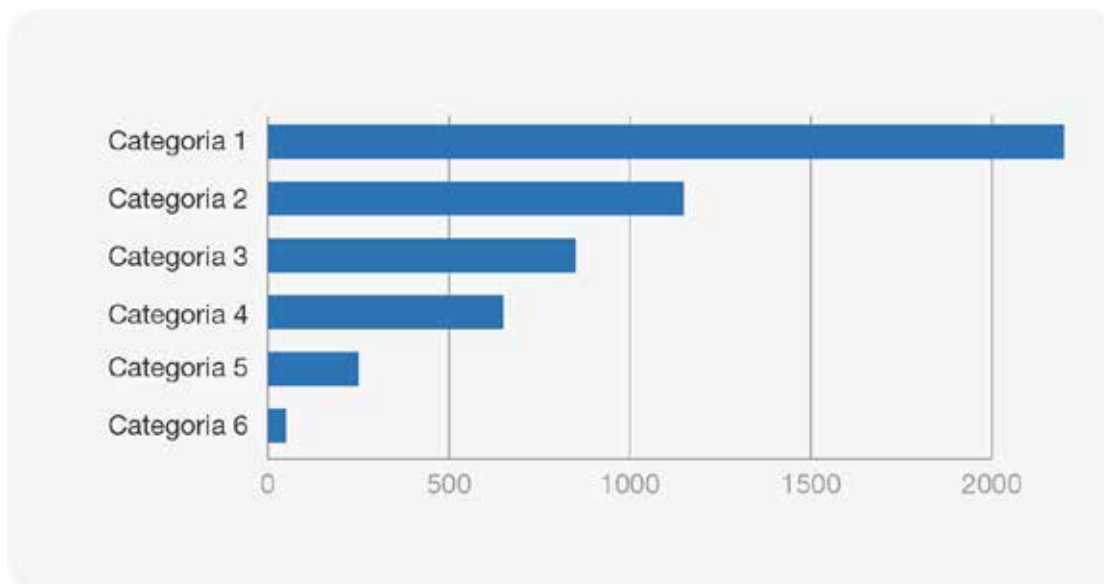


Figura 4.5. Ejemplo de gráfico de barras horizontales para facilitar la lectura del nombre de las categorías.

Otra práctica interesante es añadir una línea de referencia al gráfico de barras. Esta línea puede representar la media, la mediana, los objetivos o cualquier otro valor que ayude a descubrir en cuánto se han superado o cuánto se está por debajo de estos.

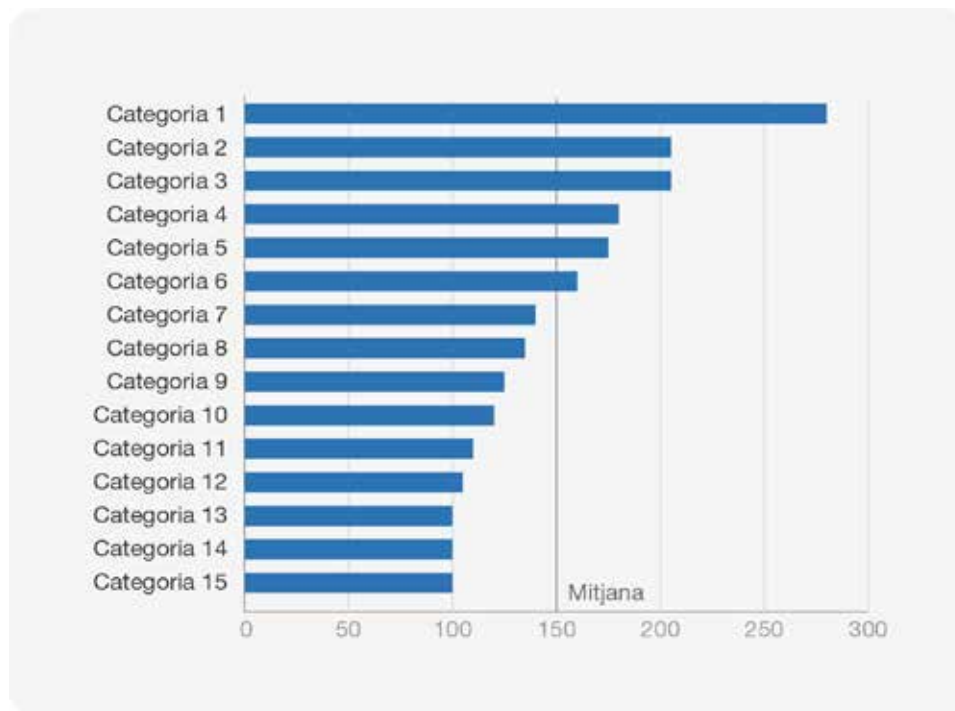


Figura 4.6. Ejemplo de gráfico de barras que muestra la media.

A menudo se quiere comparar uno de los elementos con el resto. En estos casos, es interesante resaltar la barra que le corresponde con un color diferente, de modo que sea más fácil situarlo y, por lo tanto, compararlo con el resto de elementos.

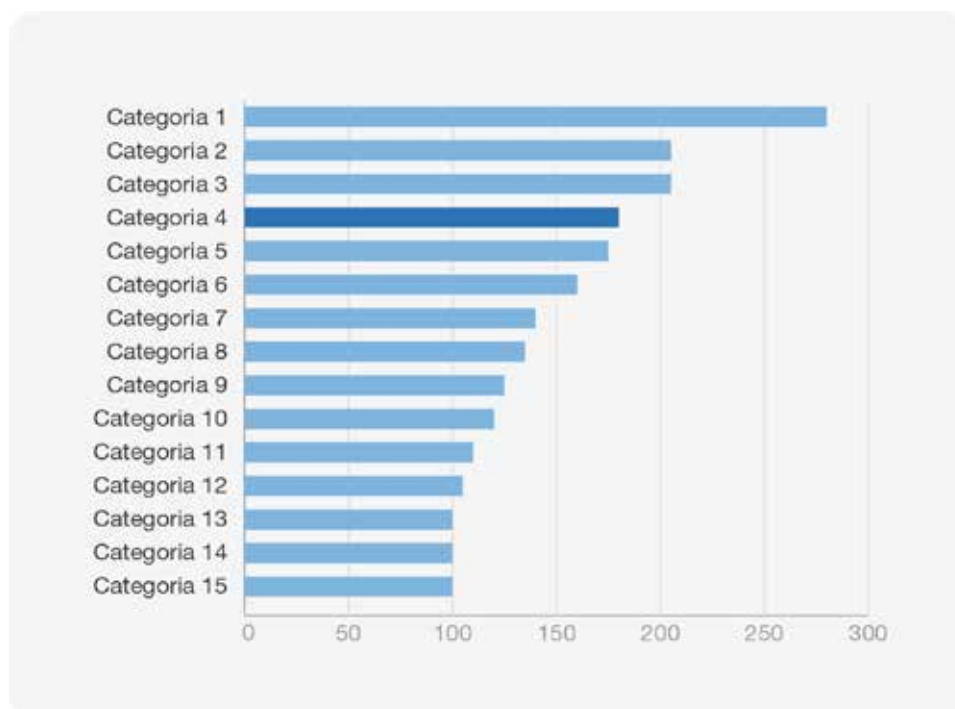


Figura 4.7. Ejemplo de gráfico de barras donde se resalta la categoría que se quiere comparar con el resto.

4.1.2. Gráfico de barras agrupadas

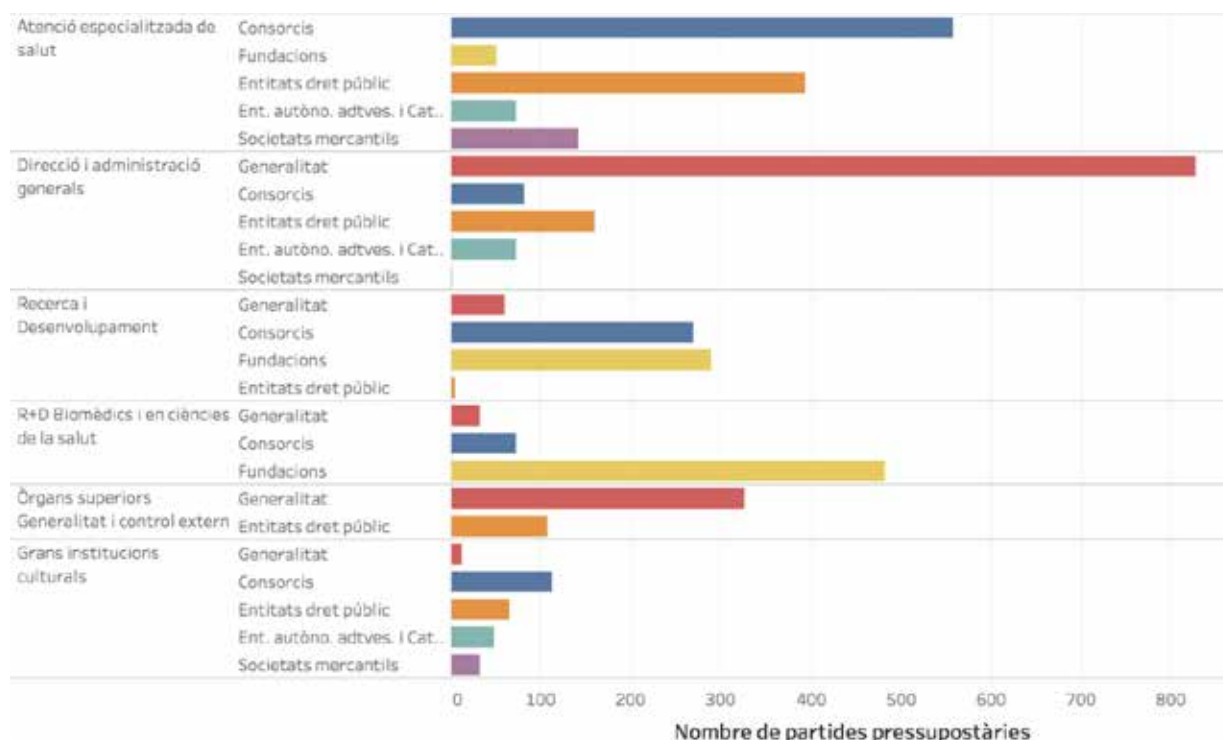


Figura 4.8. Número de partidas presupuestarias por categoría y subcategoría.

Este gráfico es una versión del gráfico de barras estándar. En este caso, consiste en un conjunto de gráficos de barras dispuestos sobre un mismo eje.

¿Cuándo se debe utilizar?

Cuando se quiere comparar el valor de los diferentes segmentos que forman parte de las categorías seleccionadas. Este gráfico permite entender la descomposición de una categoría en los diferentes segmentos que la componen (por ejemplo, cómo se reparten las partidas presupuestarias de Investigación y Desarrollo entre diferentes subsectores, como se puede ver en el ejemplo) y, al mismo tiempo, se pueden comparar las subcategorías por diferentes categorías (por ejemplo, para qué programa recibe más partidas el subsector de Fundaciones).

Recomendaciones

Para este gráfico debe tenerse en cuenta las mismas recomendaciones que las del gráfico de barras estándar. Pero es importante no sobrecargarlo con demasiadas categorías, ya que puede acabar generando problemas de comprensión. En casos así, es recomendable convertir el gráfico en diferentes gráficos de barras estándar separados.

4.1.3. Gráfico de barras apiladas

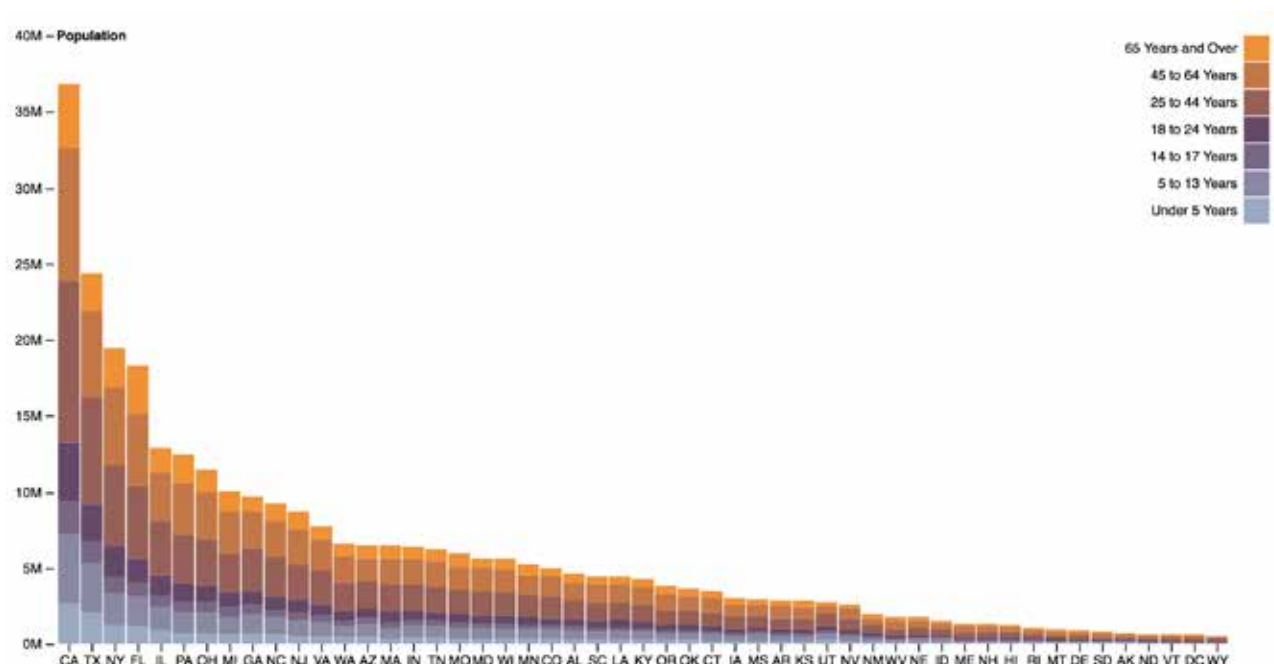


Figura 4.9. Población de los estados americanos segmentados por rangos de edad. Fuente: <https://bl.ocks.org/mbostock/3886208>.

Este tipo de gráfico representa otra extensión del gráfico de barras estándar. En este caso, se divide cada una de las barras en función de los segmentos que la componen.

¿Cuándo se debe utilizar?

El gráfico de barras apiladas se utiliza cuando se quiere comparar el valor total de la suma de los segmentos que forman cada una de las barras. Al mismo tiempo, ofrece información sobre cómo son de grandes estos segmentos.

Cuando las barras apiladas suman un 100%, o sea que cada barra segmentada ocupa toda la altura de la representación, el gráfico se puede considerar un gráfico que permite representar partes de un total.

Recomendaciones

Si se quiere comparar la composición de las barras entre sí, este gráfico no es una buena opción ya que, como se puede apreciar en la siguiente imagen, es difícil comparar el número de partidas presupuestarias destinadas a cada subsector.

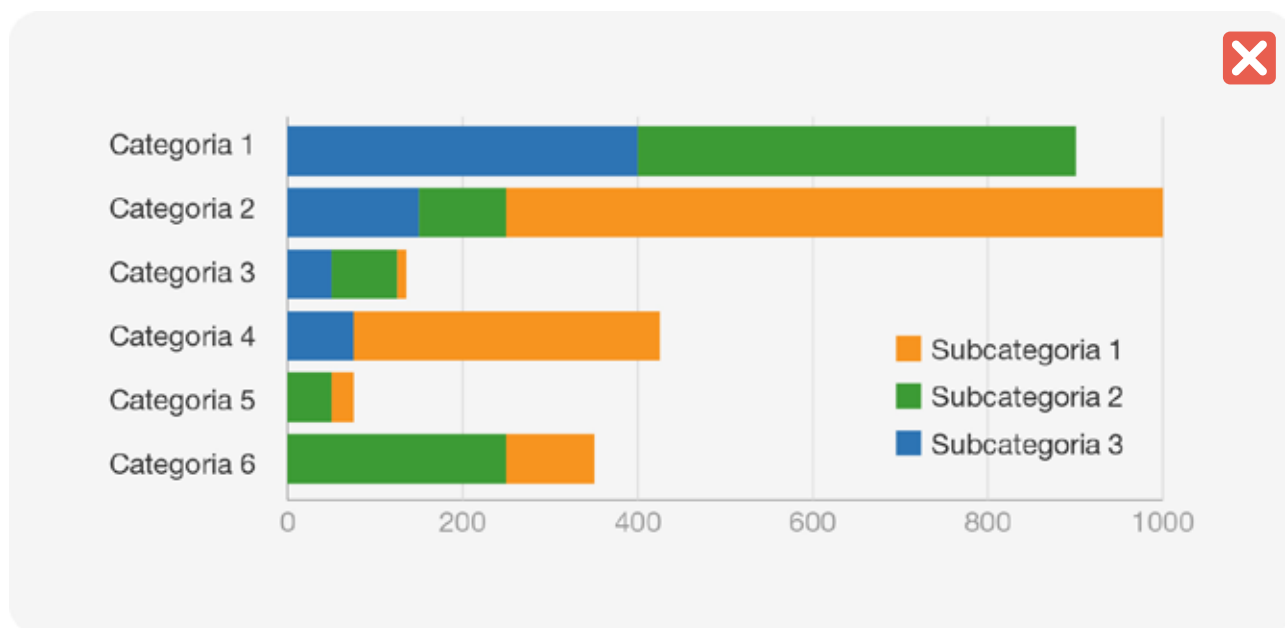


Figura 4.10. Ejemplo de barras apiladas que representan una mala práctica, ya que cuesta comparar los valores de las subcategorías 1 y 2.

En este caso, se recomienda descomponer el gráfico en diferentes gráficos de barras, como se puede ver en el siguiente ejemplo.

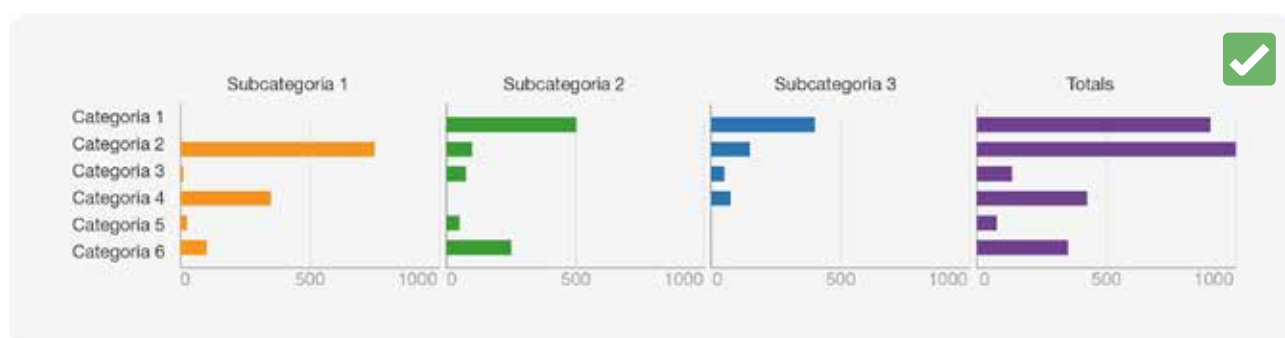


Figura 4.11. Una alternativa al gráfico de barras apiladas es separar cada subcategoría en diferentes gráficos de barras.

El número de segmentos que se utilicen para apilar las barras es muy importante. En el caso de utilizar solo dos, y que encima formen parte de un total, del 100%, se podrán obtener dos gráficos complementarios que pueden ser muy esclarecedores.

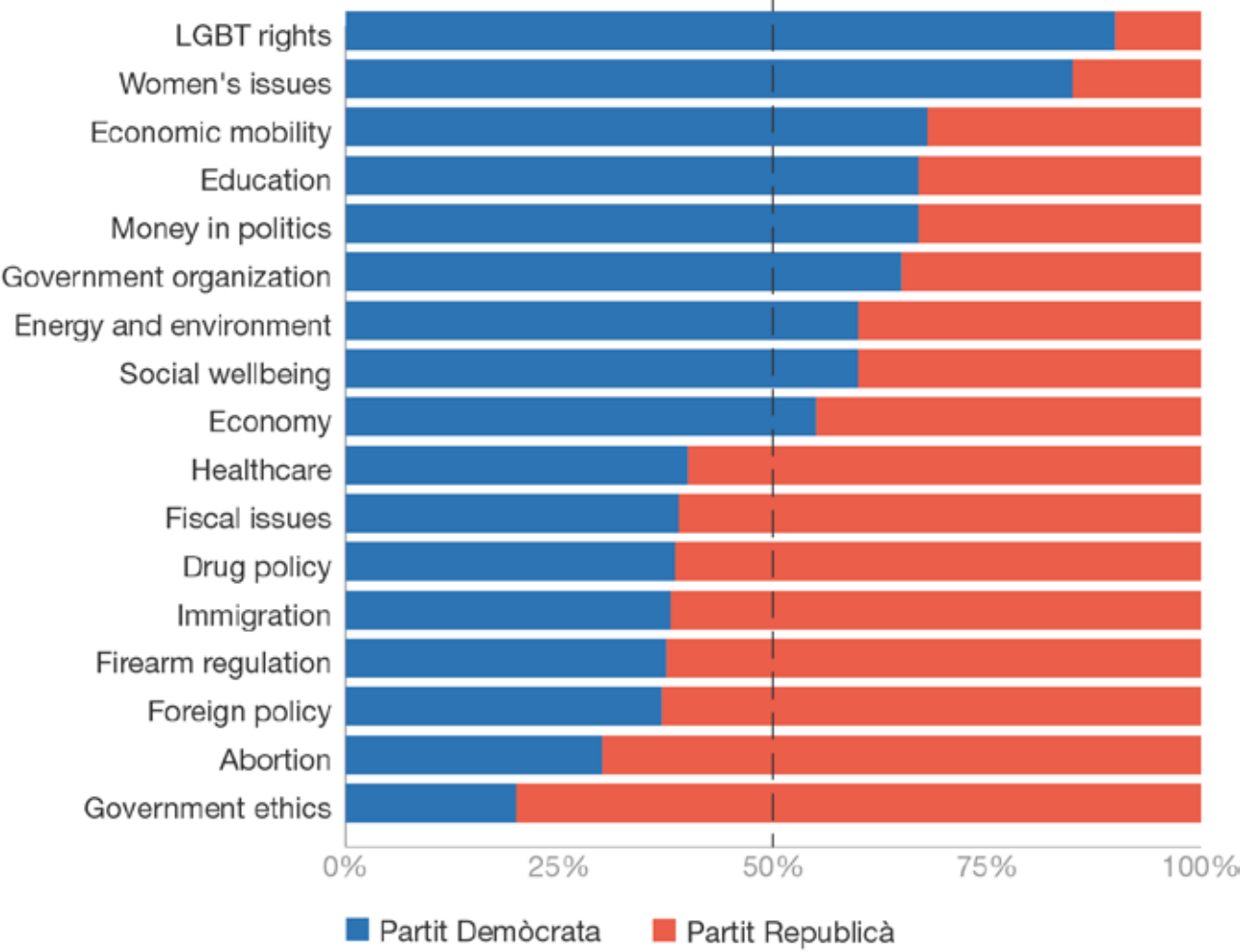


Figura 4.12. Adaptación del gráfico que muestra el porcentaje de publicaciones de Facebook dedicadas a diferentes temas de actualidad para el Partido Demócrata y el Republicano de los Estados Unidos.
Fuente: <https://www.facebook.com/notes/10152581594083859/>.

4.1.4. Gráfico de radar



Figura 4.13. Ejemplo de gráfico de radar que representa las características más valoradas en un teléfono.

Fuente: <http://bl.ocks.org/nbremer/21746a9668ffdf6d8242>

Los gráficos de radar muestran de forma circular un conjunto de ejes, cada uno relativo a una variable de nuestro conjunto de datos. El valor que toma cada variable se sitúa en un punto de su eje correspondiente y, después, se traza una línea que une estos puntos. A menudo, se colorea el área que crea el polígono resultante.

El siguiente diagrama detalla la estructura de un gráfico de radar:

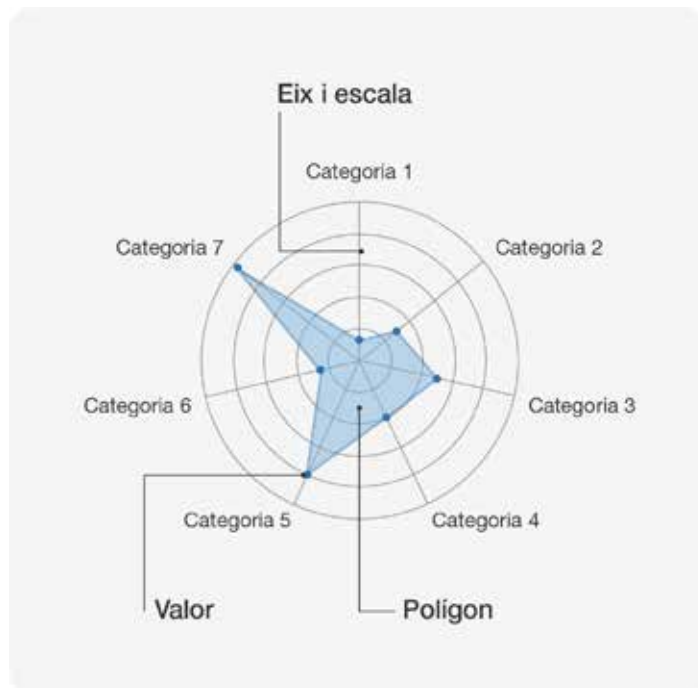


Figura 4.14. Partes que forman un gráfico de radar.

En esencia, este gráfico convierte un conjunto de valores en una forma intuitiva y fácil de interpretar.

¿Cuándo se debe utilizar?

Este tipo de gráfico se utiliza cuando se quieren entender los valores de un elemento en diferentes variables al mismo tiempo. Por ejemplo, se podría representar el rendimiento de un jugador de fútbol a partir de diferentes variables. En caso de que quisiéramos comparar diferentes jugadores, podríamos utilizar un gráfico de radar para cada jugador.

Recomendaciones

Se debe evitar mostrar demasiadas variables al mismo tiempo, de modo que no se generen formas demasiado anguladas o extrañas, dificultando la lectura y la comparación entre diferentes gráficos.

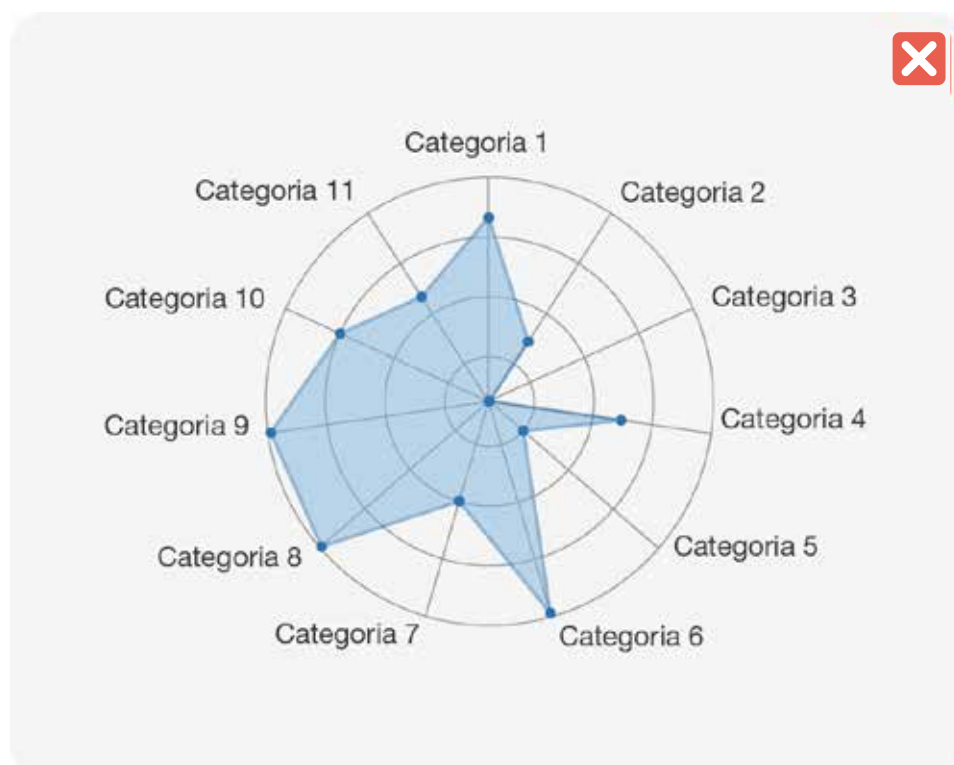


Figura 4.15. Ejemplo de gráfico de radar con demasiadas variables donde se generan polígonos demasiado angulados poco reconocibles.

Este tipo de gráfico es especialmente útil cuando el orden en que se sitúan las diferentes variables tiene algún significado. Por ejemplo, en análisis futbolísticos, las variables relacionadas con el ataque de un equipo (número de goles, chutes a portería, etc.) se sitúan en los ejes que quedan en la parte superior del gráfico, mientras que aquellas variables de aspecto defensivo (número de paradas, faltas, intercepciones de pelota, etc.) se sitúan en la parte inferior. De esta manera, un gráfico de radar que muestre un círculo casi perfecto, indicará un equipo que es muy bueno tanto en ataque como en defensa. Si, por el contrario, el gráfico es más completo en la parte superior, indicará que el equipo es mejor en ataque que en defensa.

Este gráfico no permite representar variables que puedan tener valores negativos y positivos.

4.1.5. Gráfico de intensidad de colores



Figura 4.16. Gráfico de intensidades por colores que representa el número de infectados por malaria en diferentes países. Imagen cedida por OneTandem. Fuente: https://public.tableau.com/views/malaria_5/Dashboard1.

Los gráficos de intensidad por colores (*heatmap*) son una evolución de las tablas. En lugar de representar los valores utilizando números, estos se representan mediante la intensidad del color de la celda que ocupan.

¿Cuándo se debe utilizar?

Cuando se quiere descubrir la variabilidad de un conjunto de variables. El gráfico de intensidades por colores también ayuda a revelar patrones cuando muestra si diferentes variables se comportan igual e, incluso, puede ayudar a encontrar correlaciones entre ellas.

Recomendaciones

Ordenar las filas y las columnas del gráfico de intensidades por colores según un criterio establecido puede ser muy interesante para facilitar el descubrimiento de elementos que tengan datos similares. Siguiendo el ejemplo anterior, habría que ordenar los países por número de infectados y los años, siguiendo un orden temporal.

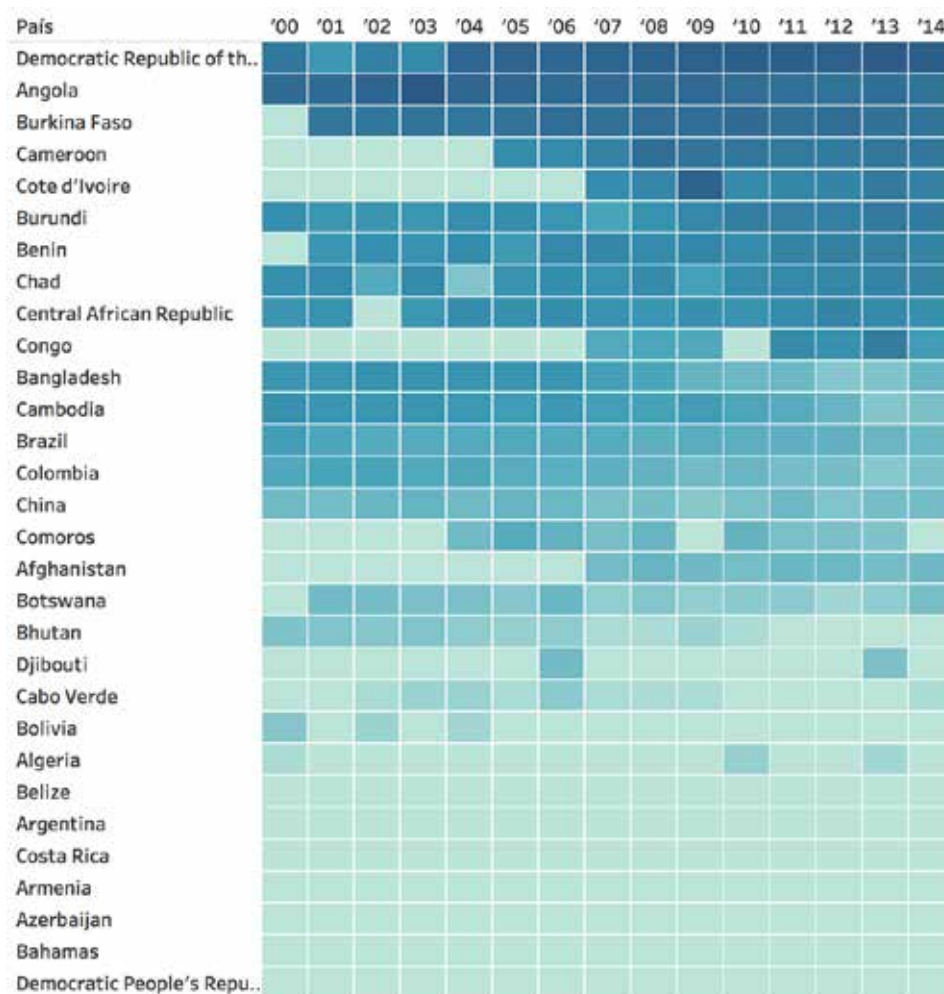


Figura 4.17. Gráfico de intensidades por colores con la evolución de infectados por malaria en algunos países del mundo, ordenados según la media anual de infectados por país. Adaptación de la imagen cedida por OneTandem. Fuente: https://public.tableau.com/views/malaria_5/Dashboard1.

Si el gráfico de intensidades por colores se utiliza para un conjunto de datos donde sus variables tienen escalas diferentes, es aconsejable normalizarlas, de modo que todas pasen a tener el mismo rango de valores. En este artículo se representan diferentes indicadores de jugadores de la NBA: <http://flowingdata.com/2010/01/21/how-to-make-a-heatmap-a-quick-and-easy-solution/>.

4.1.6. Gráfico de marcas

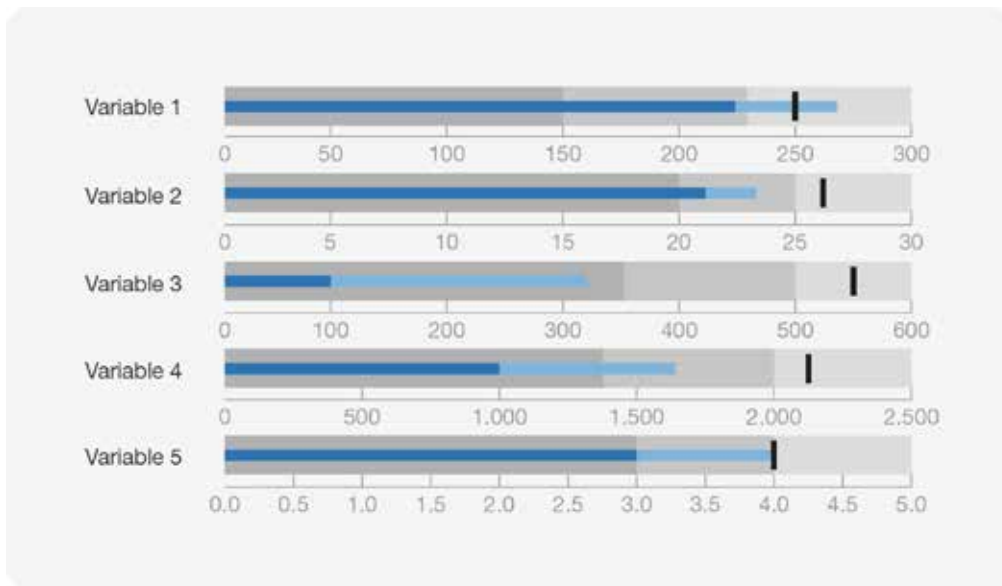


Figura 4.18. Ejemplo de un gráfico de marcas mostrando cinco categorías diferentes.

Los gráficos de bala (*bullet chart*) son una modificación del gráfico de barras, pero con dos elementos que ayudan a proporcionar contexto:

- Una marca vertical que representa un valor que permite comparar la variable representada (por ejemplo, si la barra indica ventas, la marca vertical podría indicar el objetivo de ventas).
- Un fondo con áreas de diferentes tonalidades, que indiquen los rangos de valores “aceptables”, “buenos” o “malos” que puede tomar la variable en cuestión.

El siguiente diagrama detalla la estructura de un gráfico de marcas:

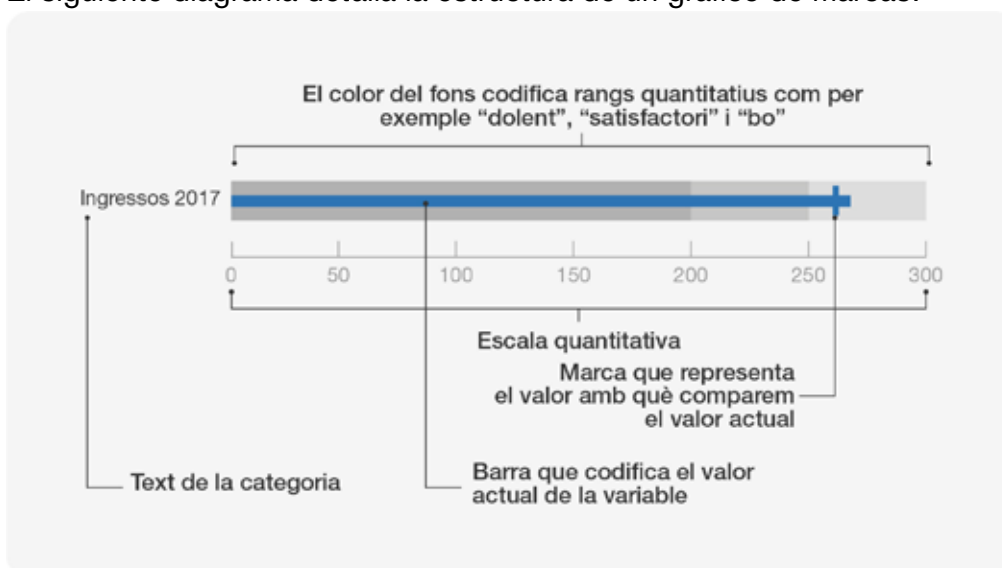


Figura 4.19. Partes que forman un diagrama de bala.

¿Cuándo se debe utilizar?

Los gráficos de bala se utilizan cuando se quiere evaluar un valor concreto respecto de lo que se esperaba y respecto de unos rangos de calidad establecidos.

La particularidad del gráfico de marcas es que permite no solo evaluar el valor de la variable re-presentada, sino compararla con un valor esperado, de manera que contextualiza el valor representado y ayuda a su interpretación. De la misma manera, las bandas pintadas del fondo también ayudan a saber si el valor se encuentra dentro de un rango identificado como aceptable, bueno o malo.

Recomendaciones

A menudo el objetivo que se marca para un indicador es anual (por ejemplo, objetivo de ventas anuales). Por lo tanto, si se hace este gráfico, por ejemplo, a mitad de año, lo más frecuente es no llegar al objetivo. Por eso, hay una versión del gráfico que muestra el valor actual (azul oscuro al ejemplo inicial) y el valor esperado (azul claro).

4.1.7. Gráfico múltiple

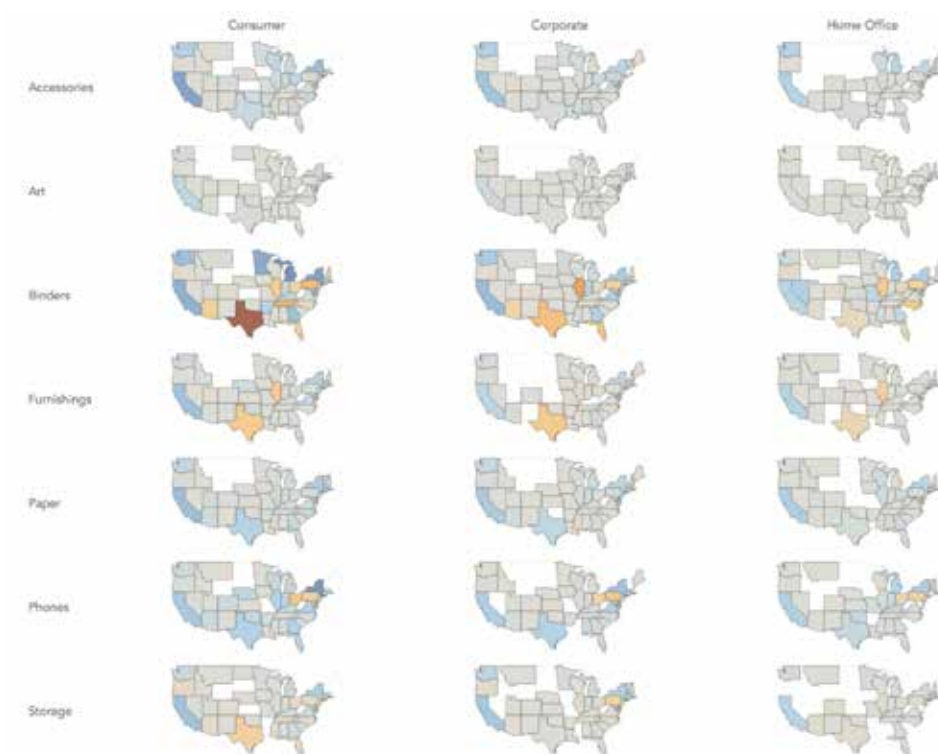


Figura 4.20. Ejemplo del uso del gráfico múltiple: una cadena de tiendas de material de oficina utiliza esta técnica para comparar los beneficios de cada estado de los Estados Unidos por diferentes categorías de producto y por sus tres tipos de clientes.

Más que un tipo de gráfico, el concepto del gráfico múltiple (*small multiples*) hace referencia a una técnica de composición de gráficos. Concretamente, esta técnica propone utilizar el mismo gráfico repetidamente, para mostrar el comportamiento de una misma variable por diferentes categorías. De esta manera, gracias a la repetición y al hecho de que los gráficos están muy juntos, se facilita la comparación entre ellos.

¿Cuándo se debe utilizar?

Cuando representar todos los datos de las diferentes categorías en un único gráfico genera representaciones complejas. Eso acostumbra a pasar cuando hay demasiadas categorías que dificultan la comprensión.

También se puede utilizar el gráfico múltiple para representar datos en el tiempo y mostrar diferentes gráficos para cada momento determinado.

Recomendaciones

Es importante que el tipo de gráfico utilizado sea lo bastante claro para que se entienda su significado a pesar de su pequeña medida.

También conviene ser coherente en el uso de los rangos en los ejes y en los colores, con el objeto de que sea fácil comparar todos los gráficos entre sí.

4.2. Tendencias

Los gráficos de tendencia permiten ver la evolución temporal de una o más variables cuantitativas. Es uno de los tipos de análisis más comunes ya que, a menudo, no solo interesa ver el valor que tiene un elemento en una variable, sino que queremos entender qué comportamiento ha tenido a lo largo del tiempo.

Gráfico de líneas

Para ver la evolución temporal de una o más variables.

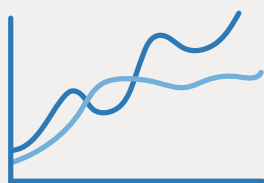
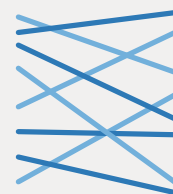


Gráfico de pendientes

Para comparar el antes y el después de una variable en varias categorías.



Minigráfico de línea

Para contextualizar un indicador.



4.2.1. Gráfico de líneas



Figura 4.21. Ejemplo de gráfico de líneas que representa la evolución del total de ventas de dos productos de una empresa.

El gráfico de líneas permite representar el cambio de una variable a lo largo del tiempo. La variable temporal o continua se sitúa en el eje x , y se construye disponiendo una serie de puntos conectados por una línea según la altura que marca la variable colocada en el eje y .

Un gráfico de líneas puede contener unas o más líneas, de modo que permite comparar la evolución de datos de diferentes categorías, a menudo representadas con colores diferentes.

En caso de que solo haya una línea, el área que queda entre esta y el eje x se puede pintar de un color concreto; así obtendremos lo que se denomina un gráfico de área.

¿Cuándo se debe utilizar?

Cuando queremos ver la evolución temporal de unas o más variables.

Recomendaciones

Conviene ser muy cuidadoso con el rango de valores utilizado en el eje y , ya que se corre el riesgo de enfatizar demasiado los cambios de la variable en el tiempo. Este efecto se puede observar en los gráficos siguientes, que muestran la evolución del número de infectados por malaria a lo largo del tiempo. El gráfico de la izquierda tiene como valor mínimo del eje y el 0, mientras que el de la derecha empieza en 60.000 infectados. Como se puede observar, en el gráfico de la derecha las pendientes de las líneas son mucho más pronunciadas y, por lo tanto, dan una sensación de grandes cambios.

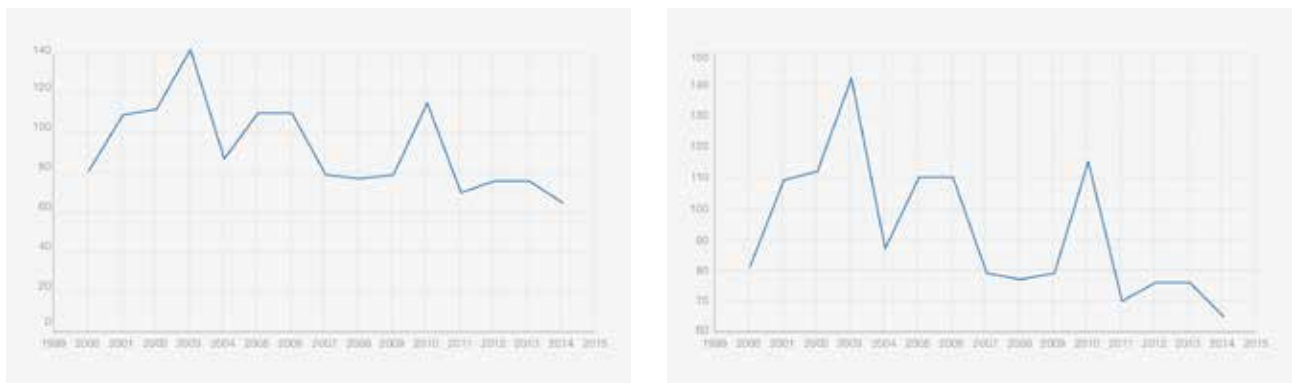


Figura 4.22. Ejemplo de dos gráficos de barras con diferentes valores en el eje y.

En general, es una buena idea añadir líneas de referencia en forma de retícula que ayuden a estimar los valores de las variables de cada punto.

El gráfico de líneas también hace que se pueda mostrar más de una línea al mismo tiempo, como se ha visto en el primer ejemplo. No obstante, la representación de demasiadas líneas puede reducir considerablemente la legibilidad del gráfico.

4.2.2. Gráfico de pendientes

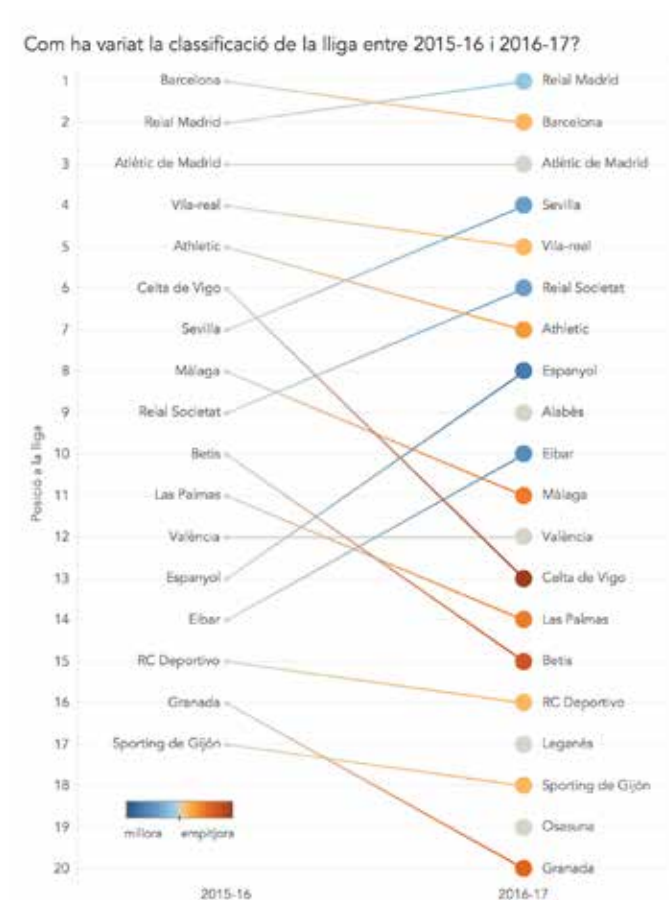


Figura 4.23. Ejemplo de gráfico de pendientes que muestra la clasificación de los diferentes equipos de la primera división de fútbol en las temporadas 2015-16 y 2016-17.

El gráfico de pendientes (*slope chart*) es un gráfico de líneas con solo dos puntos y está pensado para mostrar, a simple vista, el antes y el después de una variable concreta. Generalmente, un gráfico de pendientes acostumbra a contener más de una línea para facilitar la comparación entre los valores de diferentes categorías.

¿Cuándo se debe utilizar?

Cuando se quiere comparar el antes y el después de una variable en varias categorías. En el ejemplo anterior se puede ver la diferencia de clasificación de los diferentes equipos de la liga española en las temporadas 2015-2016 y 2016-17.

Recomendaciones

Para que el gráfico de pendientes sea coherente, los valores máximos y mínimos de cada uno de los ejes deben de ser los mismos.

4.2.3. Minigráfico



Figura 4.24. Minigráficos de línea utilizados en la cabecera de un cuadro de mando que muestran la evolución de los indicadores principales de una empresa.

Un minigráfico (*sparkline*) es un gráfico diseñado para dar contexto a una cifra. Generalmente se utilizan en cuadros de mandos y se ponen al lado del valor de un indicador clave del sistema, de manera que deben ser gráficos más bien pequeños.

Aunque inicialmente los minigráficos de líneas eran básicamente pequeñas líneas de tendencia que ayudaban a entender la evolución de una determinada métrica, el concepto se ha extendido en los últimos años y ha permitido la utilización de otros tipos de gráficos, como por ejemplo los gráficos de barras.

¿Cuándo se debe utilizar?

Cuando se quiere contextualizar un indicador. Generalmente se utilizan para representar patrones o tendencias en el tiempo, utilizando un espacio pequeño al lado del indicador.

Recomendaciones

Como norma general, la escala de los ejes de un minigráfico no se muestra, ya que el objetivo es poder ver el comportamiento del valor mostrado a grandes rasgos.

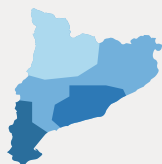
4.3. Mapas

Los mapas son uno de los tipos de gráficos más utilizados. Su principal función es ayudar a encontrar patrones geográficos en los datos.

Además, los mapas son proyecciones que intentan representar de la manera más fidedigna posible la esfera de la Tierra en dos dimensiones. Por eso hay muchas proyecciones matemáticas, cada una con sus problemáticas particulares, que posibilitan la generación de los mapas que utilizamos hoy día. En general, los programas que nos permiten mostrar datos sobre mapas ya llevan incorporados mapas que utilizan unas proyecciones que deforman al mínimo la realidad⁷.

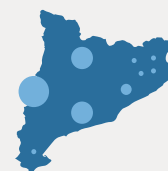
Mapa de coropletas

Para encontrar patrones geográficos en datos categorizados por zonas o regiones.



Mapa de símbolos proporcionales

Para encontrar patrones geográficos en datos que contienen información de latitud y longitud.



⁷ Más información sobre las proyecciones de mapas: Wikipedia, Proyección cartográfica.
<https://ca.wikipedia.org/wiki/Projecci%C3%B3_cartogr%C3%A0fica> [Consulta: 21 noviembre 2017]

4.3.1. Mapa de coropletas

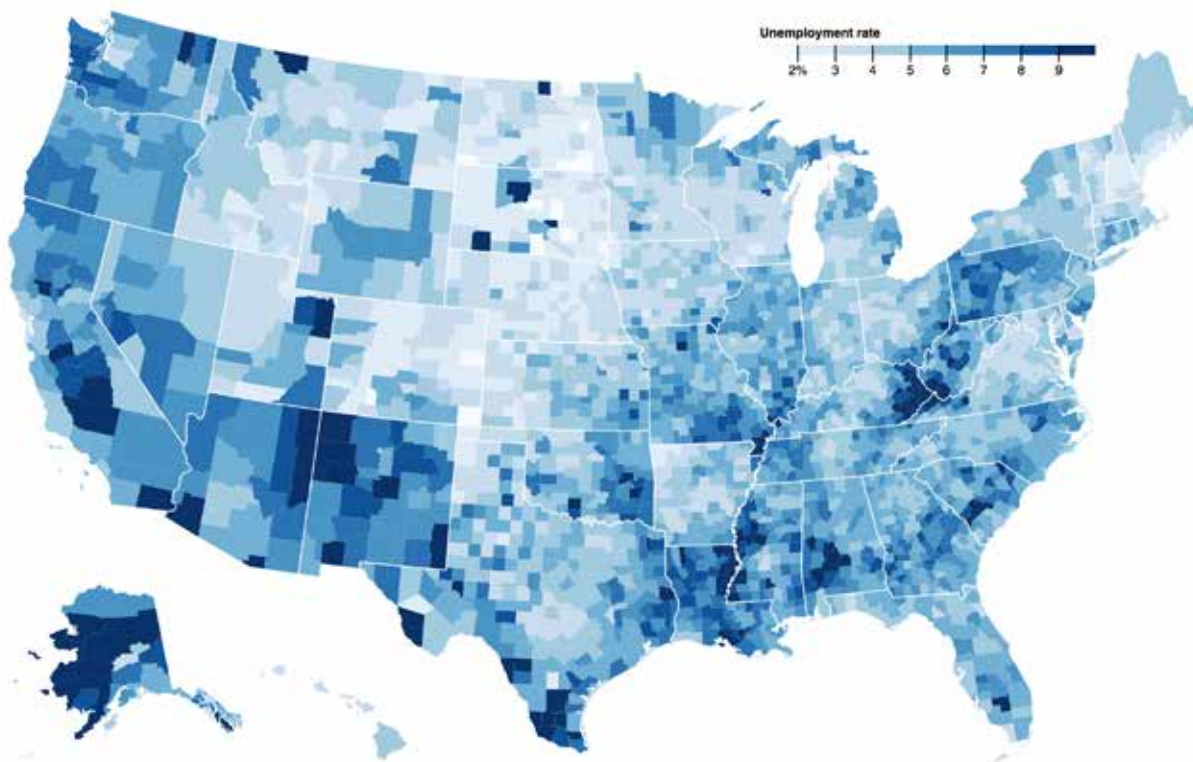


Figura 4.25. Mapa de coropletas que representa la tasa de paro en los Estados Unidos en agosto del 2016.

Fuente: <https://bl.ocks.org/mbostock/4060606>.

Los mapas de coropletas muestran los valores de una variable sobre un mapa pintando las áreas de cada región afectada de un color determinado. Los colores se utilizan para representar una variable numérica o bien para representar la pertenencia de una región a una categoría concreta. En función del objetivo que se busque se utilizará una escala de colores u otra.

¿Cuándo se debe utilizar?

Cuando se quiere encontrar patrones geográficos en los datos que están categorizadas por zonas o regiones.

Recomendaciones

En general, un mapa se debe utilizar para resolver dudas de índole geográfica. Por lo tanto, un mapa de coropletas no será el adecuado si lo que se desea es ver un ranking de las diferentes regiones.

4.3.2. Mapa de símbolos proporcionales

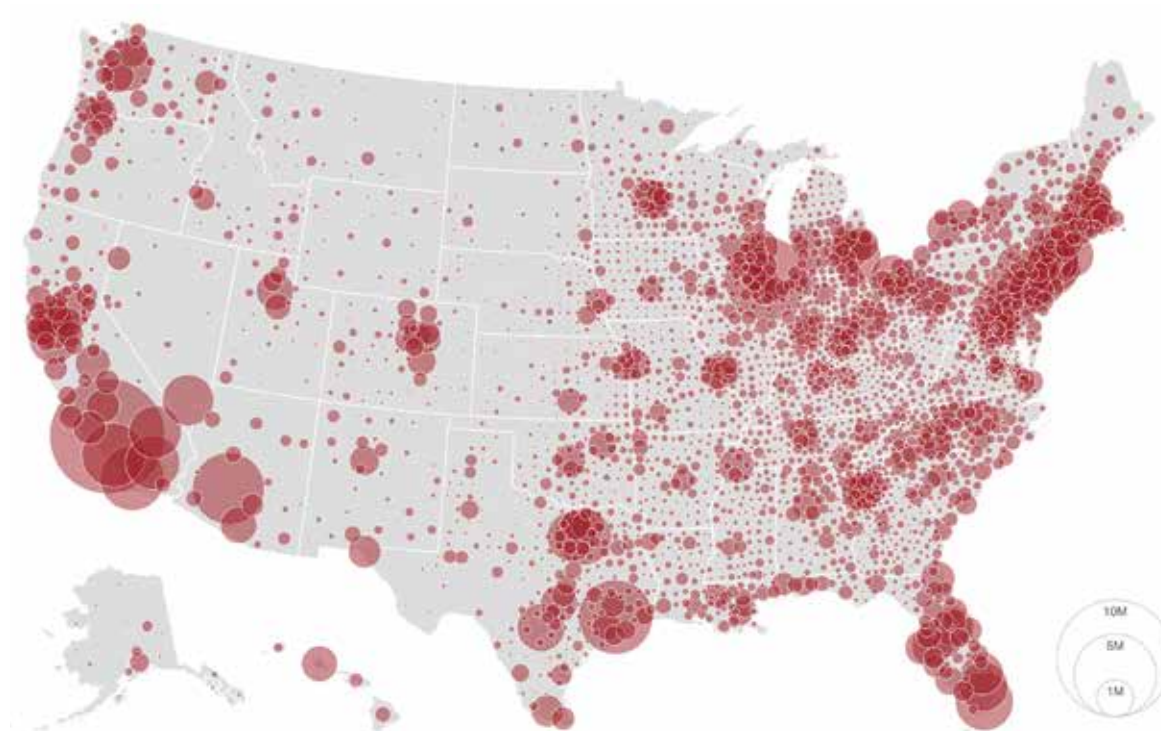


Figura 4.26. Mapa de símbolos proporcionales que muestra el número de personas censadas en los Estados Unidos.
Fuente: <https://bost.ocks.org/mike/bubble-map/>.

El mapa de símbolos proporcionales consiste en un mapa donde se sitúa un icono o símbolo, generalmente un círculo, de medida proporcional a la variable que se representa sobre el centro de la región a la que corresponde.

A diferencia del mapa de coropletas, el mapa de símbolos proporcionales permite representar datos geográficos sin necesidad de que representen un área política. Al mismo tiempo, se puede utilizar el color para añadir una segunda variable en el gráfico, o para enfatizar la variable utilizada en la medida.

¿Cuándo se debe utilizar?

Cuando se buscan patrones geográficos en unos datos que contienen puntos geográficos específicos con latitud y longitud.

Recomendaciones

Como con los mapas de coropletas, un mapa de símbolos proporcionales se debe utilizar únicamente cuando se persigue encontrar patrones geográficos, y no cuando se quiere crear un *ranking* de lugares geográficos a partir de una variable.

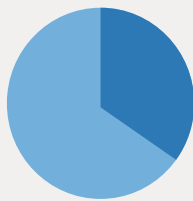
Debe tenerse en cuenta que, tal y como se puede ver en el ejemplo, a menudo este tipo de mapas generan un encabalgamiento entre los símbolos utilizados para representar los valores de la variable seleccionada. En estos casos, hace falta utilizar la transparencia para facilitar la lectura del gráfico e identificar los símbolos que quedan unos debajo de los otros.

4.4. Partes de un total

Analizar datos que forman parte de un total es una tarea que hacemos muy a menudo. Por ejemplo, cuándo queremos entender la contribución de cada departamento al total de gastos de una organización, o cuándo queremos analizar las respuestas de una encuesta. A continuación, describimos una serie de gráficos que son muy útiles para comunicar este tipo de información.

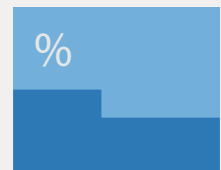
Gráfico de sectores

Para mostrar como se distribuyen proporcionalmente nuestros datos.



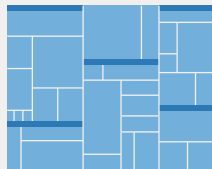
Pictograma

Para expresar un valor y a la vez ofrecer un contexto que nos permita identificar si es grande o pequeño.



Mapa de árbol

Para representar datos jerárquicos y comparar una o dos variables entre los diferentes elementos.



4.4.1. Gráfico de sectores

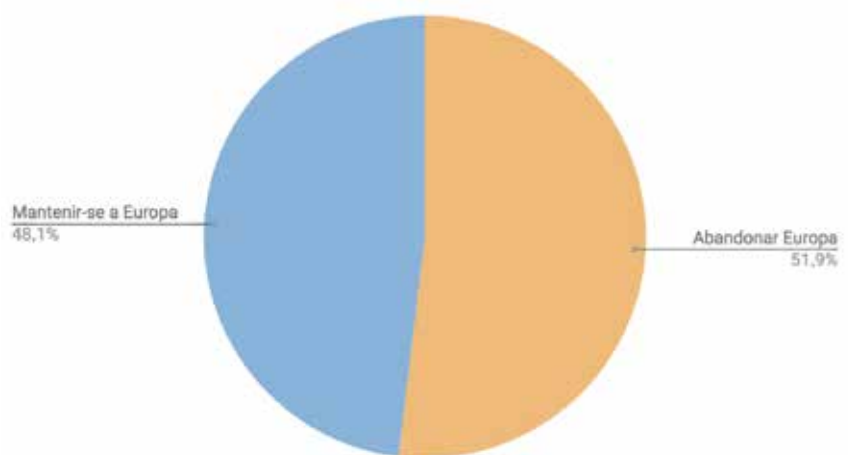


Figura 4.27. Resultado de la votación del Brexit en el Reino Unido.

También denominado *gráfico de pastel*, el gráfico de sectores utiliza una representación circular que se divide en sectores proporcionales a los valores de cada categoría representada.

¿Cuándo se debe utilizar?

El gráfico de sectores se utiliza cuando se quiere mostrar cómo se distribuyen proporcionalmente los datos. Es un gráfico ideal cuando hay dos valores, como por ejemplo el número de personas que contestan sí o no a una pregunta.

También es especialmente útil cuando se quiere comparar un valor con la suma de los otros, como se puede apreciar en la siguiente imagen, donde queda claro que el sector naranja es mayor que la suma del resto de valores.

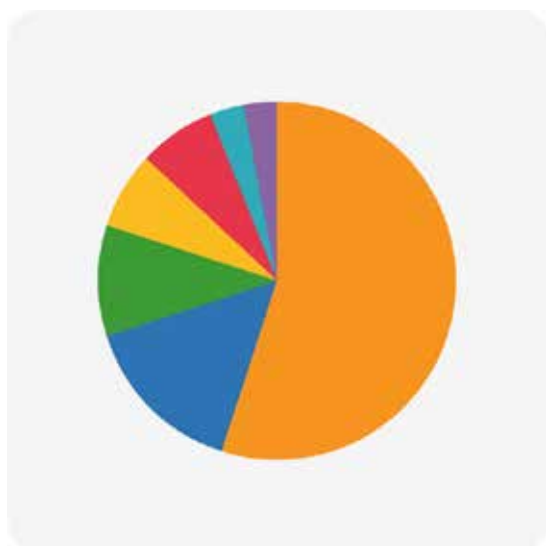
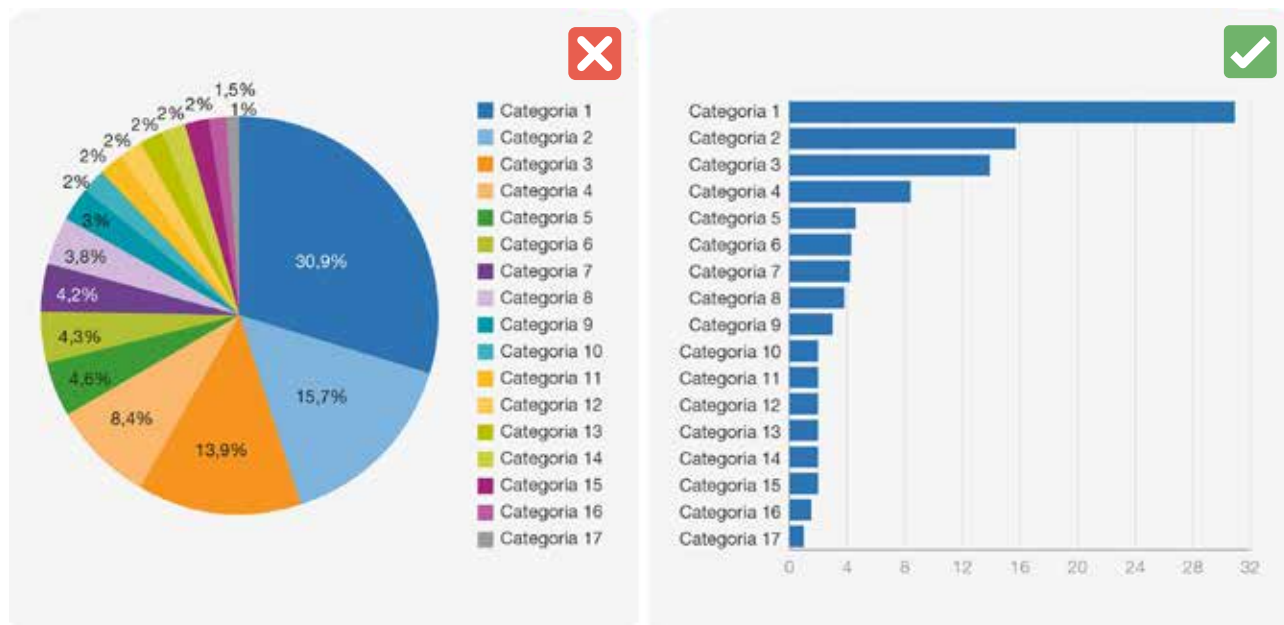


Figura 4.28. En este gráfico de sectores podemos ver cómo la categoría naranja ocupa más de la mitad del total, una cifra a la cual no llegan el resto de sectores juntos.

Recomendaciones

Excepto en los casos mencionados antes, no es recomendable utilizar el gráfico de sectores para representar más de dos o tres partes de un total. En este caso, es mucho mejor utilizar un gráfico de barras.



Figuras 4.29. Los gráficos de sectores con muchas divisiones no son buenas representaciones ya que confunden y no ayudan a entender la diferencia entre los valores. Fuente: <http://www.conceptdraw.com/solution-park/resource/images/solutions/pie-charts/Graphs-and-Charts-Pie-Chart-Business-Report.png>.

El uso del 3D es especialmente popular para este tipo de gráfico; sin embargo, esta técnica todavía dificulta más su comprensión por culpa de la perspectiva.

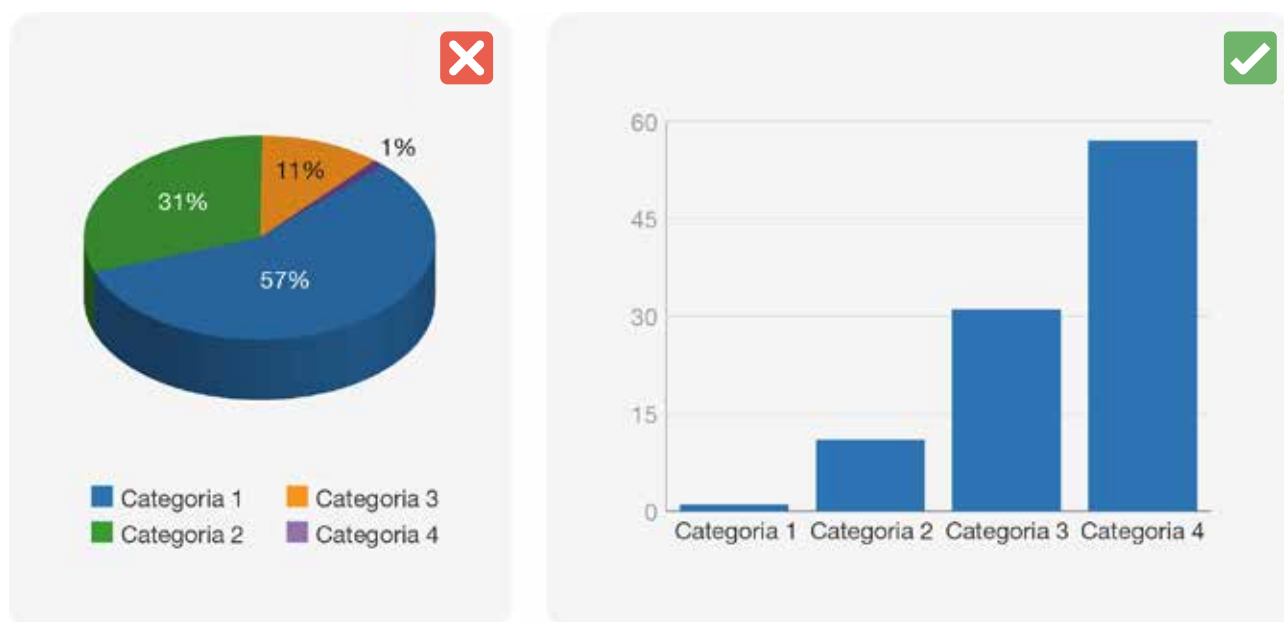


Figura 4.30. Gráfico de sectores en 3D representado con un gráfico de barras. Adaptación de la imagen cedida por Ann K. Emery. Fuente: <http://annkemery.com/pie-chart-guidelines/>.

Si se tienen datos ordinales (que siguen un orden secuencial), es preferible utilizar un gráfico de barras. Así, se puede jugar con los extremos para mostrar la gradación (en el ejemplo, desde el extremo de gran acuerdo al extremo de gran desacuerdo).



Figura 4.31. Datos ordinales representados con un gráfico de barras apiladas. Adaptación de la imagen cedida por Ann K. Emery.
Fuente: <http://annkemery.com/pie-chart-guidelines/>.

Si se tienen datos que siguen un orden temporal, es preferible utilizar un gráfico de líneas o de barras.

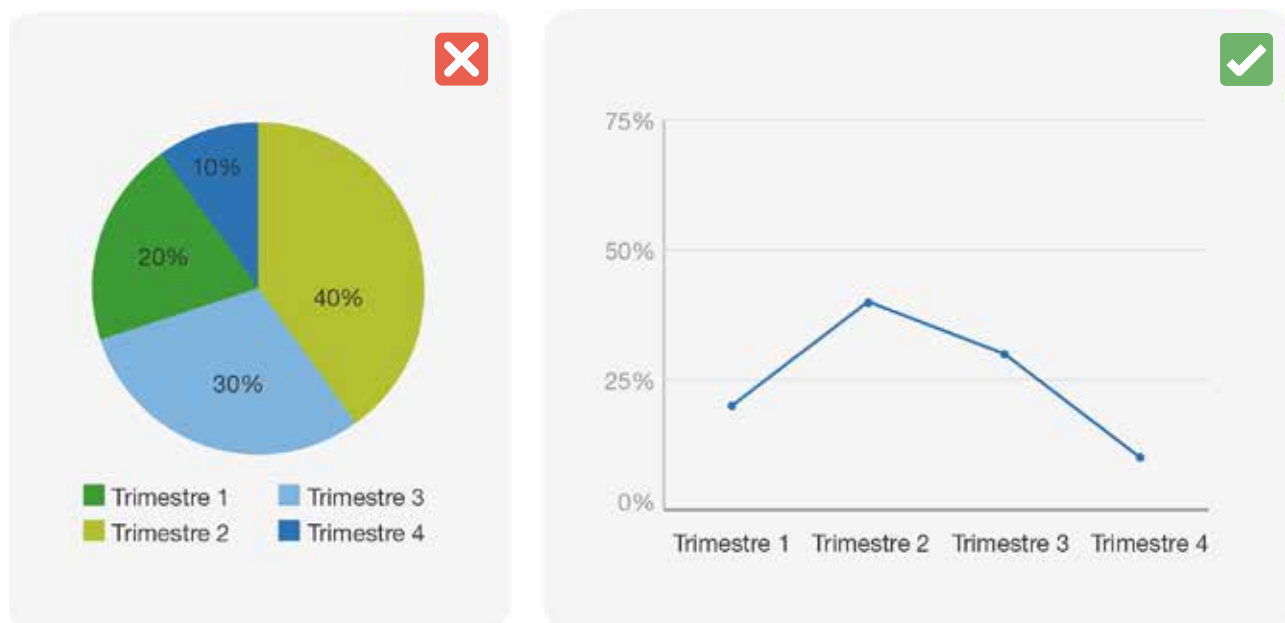


Figura 4.32. Es mejor representar los datos temporales con un gráfico de línea. Adaptación de la imagen cedida por Ann K. Emery.
Fuente: <http://annkemery.com/pie-chart-guidelines/>.

Si se tienen que comparar diferentes categorías, nuevamente es preferible utilizar barras, que facilitan la comparación de su composición, ya que se pueden ordenar visualmente.

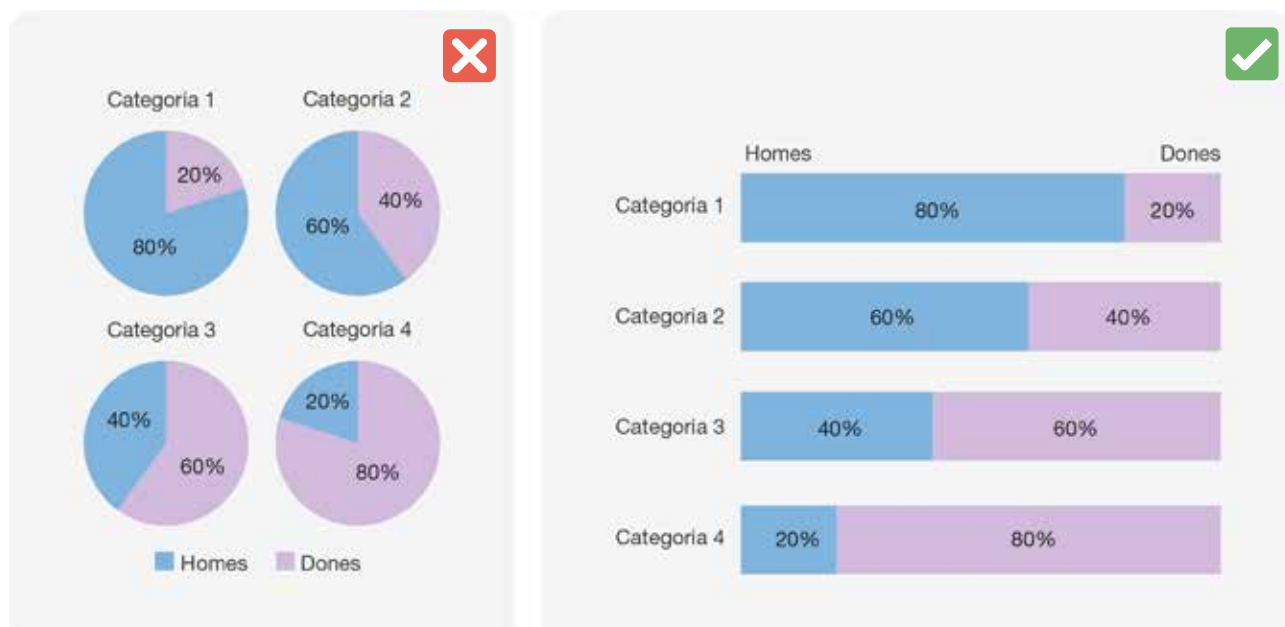


Figura 4.33. Es preferible utilizar barras para comparar valores a través de varias categorías. Adaptación de la imagen cedida por Ann K. Emery. Fuente: <http://annkemery.com/pie-chart-guidelines/>.

Una alternativa al gráfico de sectores es el llamado gráfico de corona. Este se construye exactamente igual que el de sectores, pero añadiendo un agujero en medio donde se puede mostrar un valor. Esta variación se utiliza a menudo para mostrar el progreso de un valor.

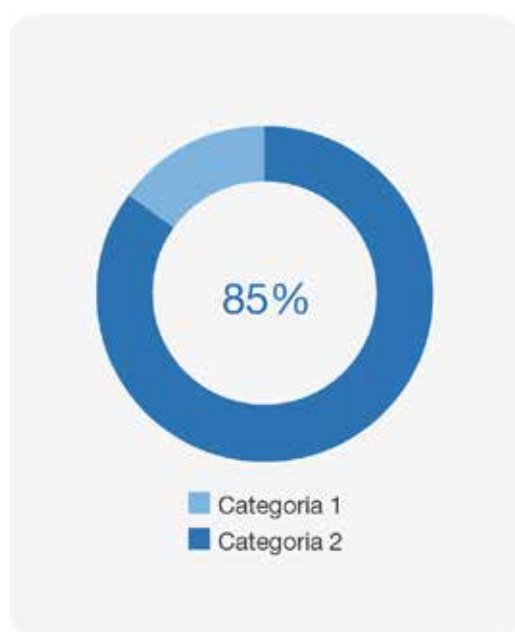


Figura 4.34. Ejemplo de gráfico de corona.

4.4.2. Pictograma

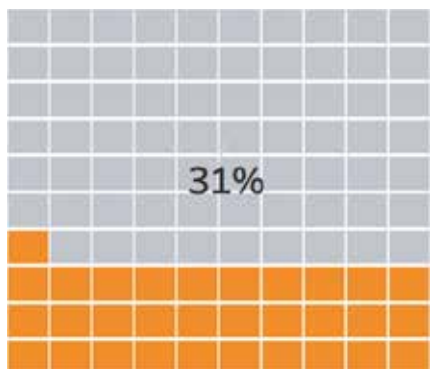


Figura 4.35. Ejemplo de pictograma.

Este gráfico se utiliza para expresar un valor concreto, generalmente un porcentaje. Se utiliza en forma de cuadrado o de rectángulo. El cuadrado exterior representa el valor máximo, y el número de cuadrados pintados de otro color representa el valor actual.

¿Cuándo se debe utilizar?

Cuando se quiere expresar un valor y al mismo tiempo ofrecer un contexto que permita identificar si es grande o pequeño. Por ejemplo, en el siguiente gráfico se puede ver el porcentaje de exportaciones de cava por cada continente. Es fácil ver que las exportaciones se concentran en Europa y que los otros continentes pueden representar una gran oportunidad comercial.

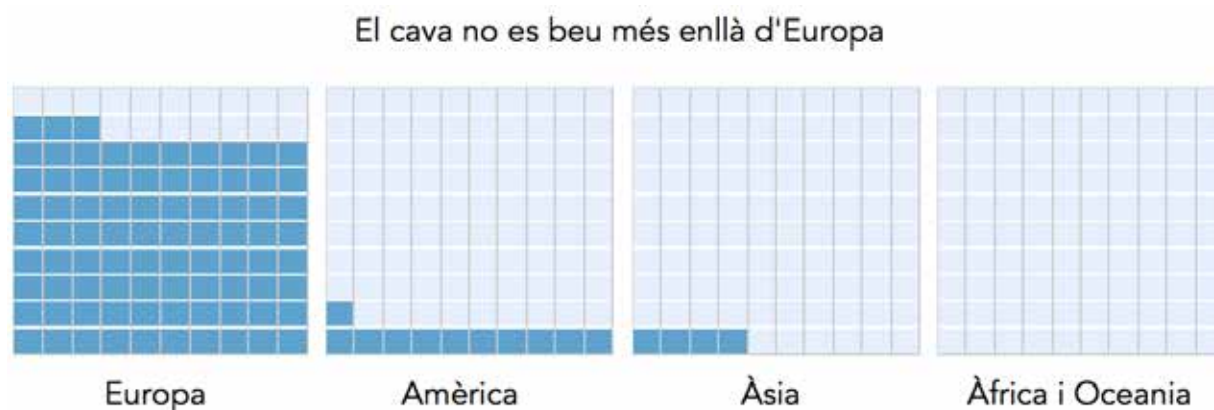


Figura 4.36. Percentatge de població que beu cava als diferents continents. Visualització feta per OneTandem.

Font: <https://public.tableau.com/profile/onetandem#!/vizhome/cava/cavaexports>.

Los pictogramas son una buena alternativa gráfica para mostrar un único número, especialmente cuando se trata de un indicador principal que va del 0% al 100%.

Recomendaciones

Son óptimos para representar indicadores con un valor máximo de 100, por ejemplo, los porcentajes que nunca superarán el 100%.

Si es importante detallar los decimales del indicador, no son una buena alternativa porque cada cuadrado de color representa una unidad.

4.4.3. Mapa de árbol

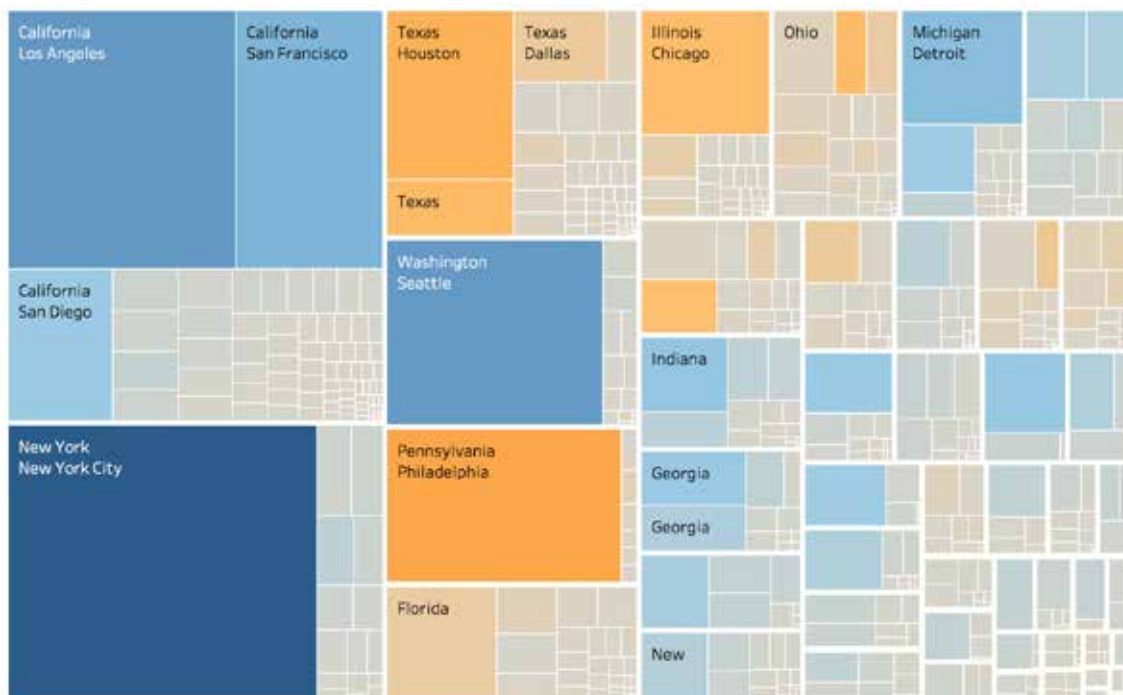


Figura 4.37. Mapa de árbol que representa el volumen de ventas (representado con la medida) y el beneficio (representado con el color) de una empresa ficticia en diferentes ciudades americanas agrupadas por estado.

El mapa de árbol (*treemap*) permite ver una agrupación jerárquica de valores. Se construye dividiendo un rectángulo en rectángulos más pequeños, donde la medida de estos y su color dan una información determinada.

¿Cuándo se debe utilizar?

Cuando los datos tienen una jerarquía y se quiere comparar una o dos variables entre los diferentes elementos de los datos. En el ejemplo, podemos ver la jerarquía que se establece entre los estados y las ciudades donde una empresa vende productos. Con este mapa de árbol se compara el volumen de ventas en cada una de las ciudades, representado mediante la medida de los rectángulos. El color muestra el beneficio que se obtiene en cada ciudad (naranja para beneficios negativos y azul para beneficios positivos).

Recomendaciones

La medida de los rectángulos hace que no se vean fácilmente las diferencias entre valores muy parecidos. Por lo tanto, este gráfico no es adecuado para identificar cuáles son los valores mayores (o sea, para representar un *ranking*).

El mapa de árbol es especialmente útil en versión interactiva. Eso hace, por ejemplo, que se muestre el nombre de cada rectángulo cuando se pasa por encima con el cursor. De esta manera se evita uno de los problemas principales de este gráfico: que si el cuadrado es demasiado pequeño, no se ve el nombre de la categoría que representa.

4.5. Distribuciones

Antes de abordar el análisis y la comunicación de un conjunto de datos, conviene entender la distribución de las diferentes variables. La distribución ayudará a comprender la forma y las propiedades principales de una variable. Asimismo, se detecta si hay una distribución normal, de larga cola o más bien aleatoria de los valores. Eso nos indicará, por ejemplo, si es adecuado utilizar medidas estadísticas como la media o la mediana.

Histograma

Para entender como están distribuidos los valores de una variable.

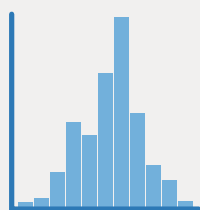
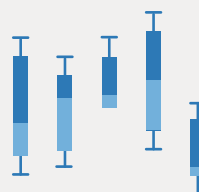


Diagrama de caja

Para comparar la distribución de diferentes variables o la distribución de una de ellas en diferentes categorías.



4.5.1. Histograma

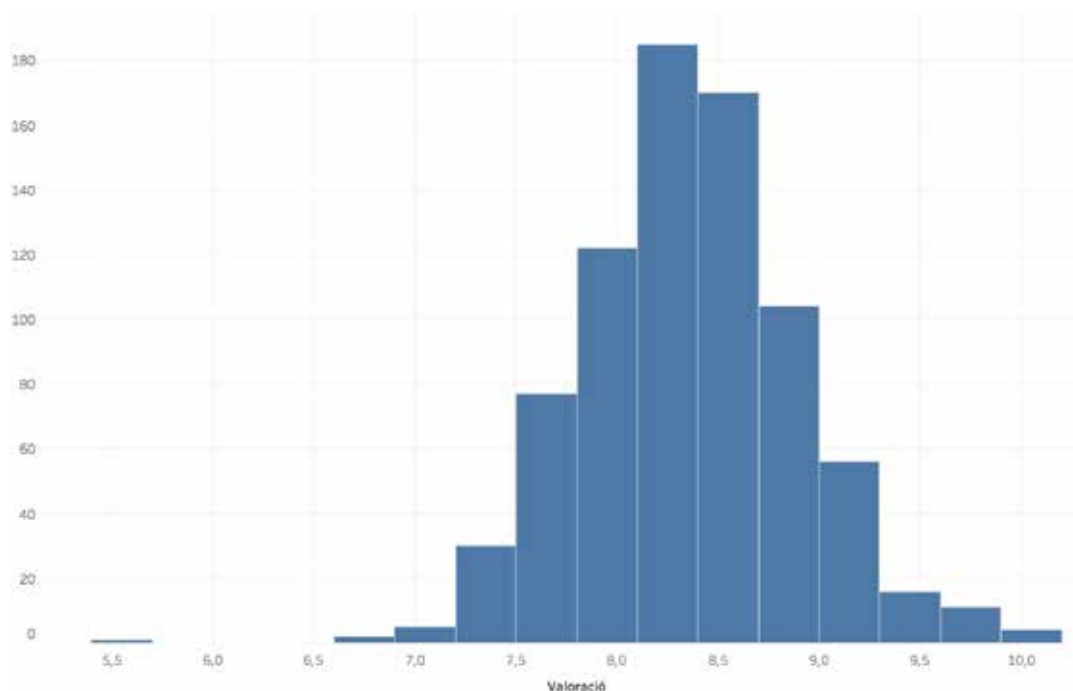


Figura 4.38. Histograma de una distribución normal que muestra las valoraciones de un conjunto de series de televisión. Visualización hecha por OneTandem. Fuente: <https://public.tableau.com/views/series2017/Dashboard1>.

Un histograma es un gráfico de barras que muestra la distribución de valores de una variable. La altura de cada barra representa la frecuencia de aparición de valores dentro del rango que define cada barra.

¿Cuándo se debe utilizar?

Cuando se quiere entender cómo están distribuidos los valores de una variable. En particular, cuando los valores se distribuyen de forma similar en torno a la media, como en el ejemplo, se habla de una distribución normal. Si, en cambio, la forma dibujada por el histograma no tiene forma de campana, y el valor máximo se sitúa en uno de los extremos, se habla de una distribución asimétrica.

Recomendaciones

Por convención, las barras de los histogramas no se deben separar entre sí.

La forma de un histograma dependerá del rango de valores que representa cada barra. Este rango es importante ya que es lo que nos permitirá agrupar los valores y, por lo tanto, acabará definiendo la forma del histograma. Generalmente, los programas deciden automáticamente la medida de este rango en función del conjunto de barras, aunque se puede editar. A continuación se muestran dos distribuciones de los mismos datos, utilizando dos rangos diferentes para agrupar los valores del conjunto de datos. El de la izquierda utiliza un rango de 0,5, mientras que el de la derecha utiliza uno de 0,25.

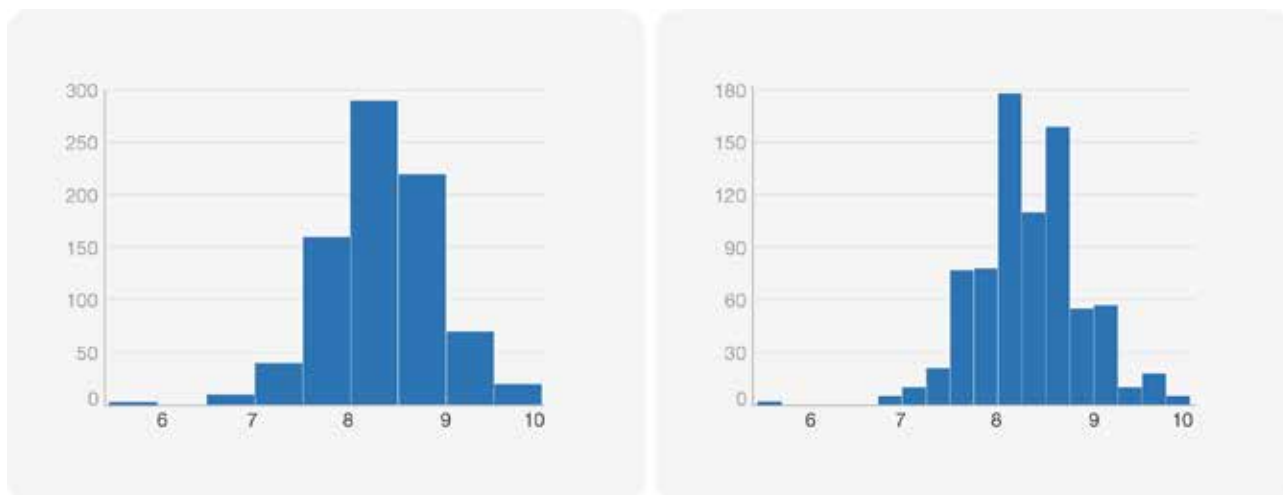


Figura 4.39. Dos histogramas sobre los mismos datos que utilizan rangos diferentes para agrupar los datos.

No hay una fórmula exacta e ideal para calcular este rango, por tanto, será trabajo del analista de datos probar diferentes fórmulas para hacerse una idea de cómo se distribuyen los datos.

4.5.2. Diagrama de caja

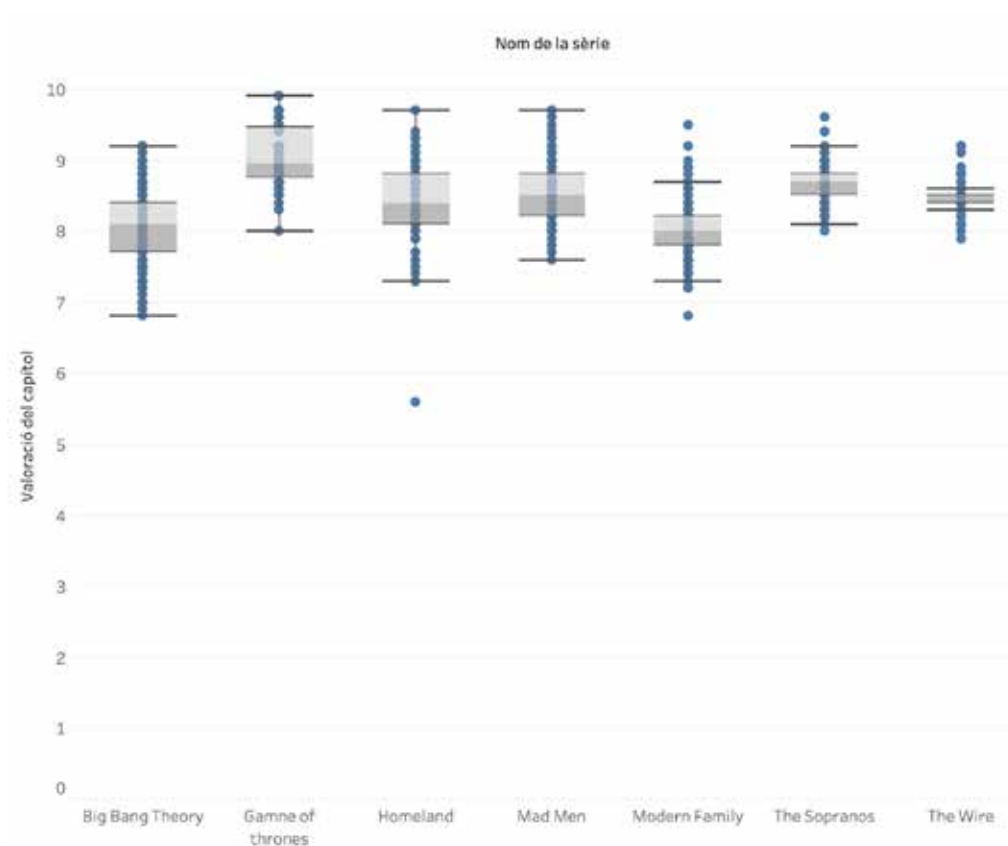


Figura 4.40. Diagrama de caja de las valoraciones de los capítulos de series de televisión (datos extraídos de www.imdb.com). Cada punto es un capítulo diferente. Gráfico elaborado por OneTandem.

Un diagrama de caja (*box plot*) permite representar la distribución de los valores de una variable y, opcionalmente, compararlos por diferentes categorías.

En el siguiente diagrama se explica el significado de los elementos de un diagrama de caja. La caja está dividida por una línea que representa la mediana. Los extremos de la caja corresponden al primer y al tercer cuartil. De la caja, emergen unas líneas denominadas “bigotes”, que pueden calcularse de diferentes maneras. Todos los puntos que se encuentren más allá de los bigotes se consideraran valores atípicos.

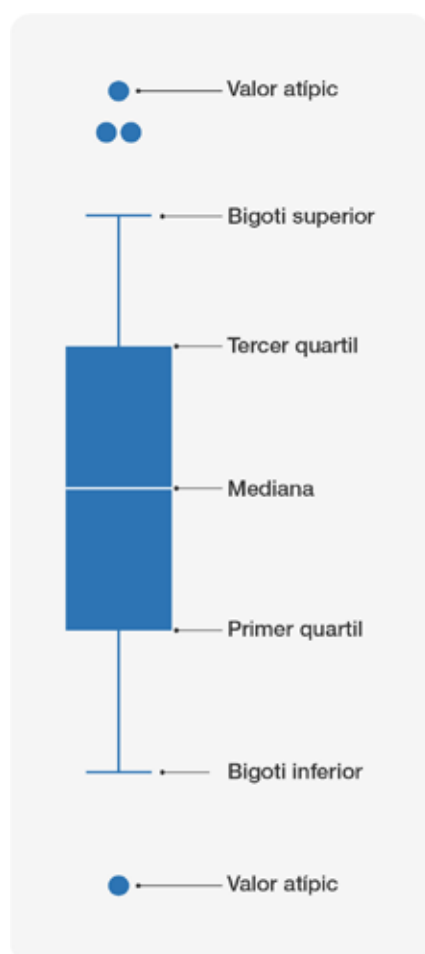


Figura 4.41. Esquema de un diagrama de caja.

¿Cuándo se debe utilizar?

Cuando queremos comparar la distribución de diferentes variables o la distribución de una variable en diferentes categorías.

El diagrama de caja ayuda a entender la forma de los datos, un paso previo en todo análisis de datos. Si la línea de la mediana está situada en el centro de la caja (como la serie *MadMen* en la imagen), querrá decir que hay una distribución simétrica (a menudo, una distribución “normal”). En caso contrario, querrá decir que hay una distribución asimétrica (ver [Histogramas](#)).

Por otra parte, si la caja es muy pequeña, quiere decir que los valores de la variable están poco dispersos (como en el caso de *The Wire* en el ejemplo). Dicho de otra manera, es fácil predecir un valor aproximado de la variable, ya que siempre presenta valores muy similares.

El diagrama de caja también es idóneo para detectar valores atípicos, que son aquellos puntos que se encuentran más allá de las líneas horizontales situadas en los extremos de cada caja o bigotes (se puede apreciar uno muy importante en la serie *Homeland*). Estos valores pueden ser fruto de errores de medida y, por lo tanto, es importante descartarlos para evitar llegar a conclusiones erróneas (por eso el uso del diagrama de caja es común en ámbitos de investigación científica).

A menudo, sin embargo, los valores atípicos indican que alguna cosa está pasando. Hará falta analizar los datos más a fondo, o encontrar otros nuevos, porque seguramente estamos a las puertas de un hallazgo interesante.

Recomendaciones

Para hacer un buen uso de este gráfico conviene conocer conceptos fundamentales de la estadística descriptiva, en particular, de la mediana y los cuartiles. Por lo tanto, no es un gráfico adecuado para todos los públicos.

De la misma manera que los gráficos de barras, los diagramas de caja pueden ser horizontales o verticales. Generalmente se hacen verticales, pero si los nombres de las categorías son muy largos, habrá que hacerlos horizontales para que quepan cómodamente.

Se denomina bigotes a las líneas horizontales que se sitúan en los extremos del diagrama de caja. Estas líneas ayudan a identificar valores atípicos. Hay diferentes convenciones para situar estas líneas:

- Situarlas a 1,5 veces la distancia intercuartil, a partir del primer y tercer cuartil respectivamente
- Situarlas en los percentiles 9 y 91 de los datos
- Situarlas en los percentiles 2 y 98 de los datos

4.6. Correlaciones

Los gráficos de correlaciones ayudan a entender la relación entre diferentes variables. Por ejemplo, permiten ver si el precio de un producto está relacionado con su demanda. Sin embargo, conviene tener presente que la correlación de dos variables no implica necesariamente que la una sea causa de la otra.

Gráfico de dispersión

Para ver la relación que existe entre dos variables.

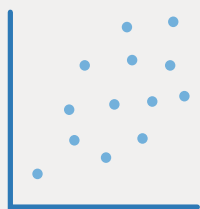


Gráfico de burbujas

Cuando se quiere emplear un gráfico de dispersión y, al mismo tiempo, tener más información de los elementos representados



Coordenadas paralelas

Para explorar conjuntos de datos multidimensionales.



4.6.1. Gráfico de dispersión

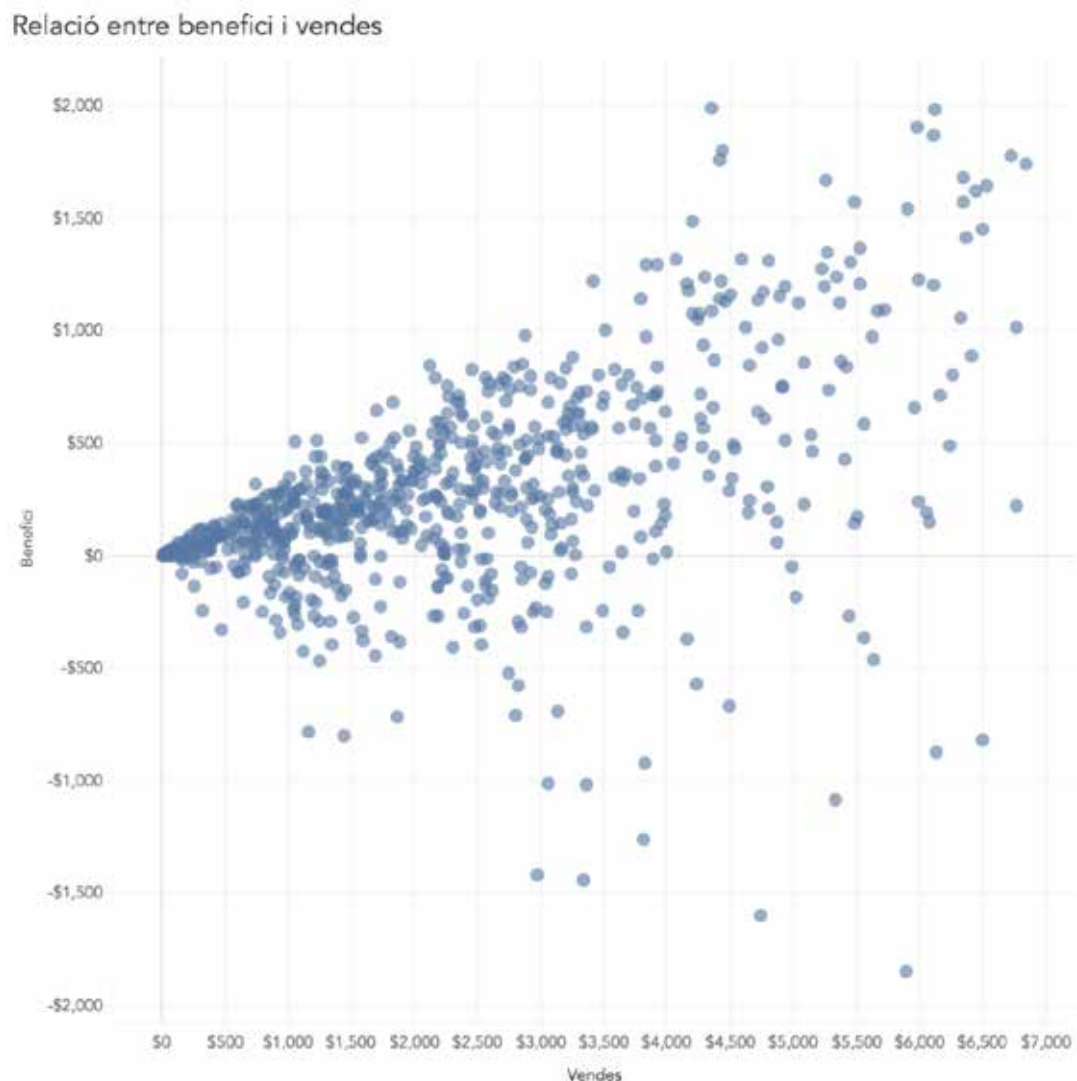


Figura 4.42. Gráfico de dispersión que muestra las ventas y el beneficio por producto. Cada punto es un producto.

Es un tipo de gráfico que permite representar los valores de dos variables sobre dos ejes de coordenadas x y y. Cada elemento se representa como un punto situado en el espacio en función de sus valores en cada uno de los ejes.

¿Cuándo se debe utilizar?

Cuando se quiere ver la relación que hay entre dos variables.

El gráfico de dispersión también es útil para detectar elementos atípicos, que no siguen la misma relación de correlación que el resto de puntos.

Recomendaciones

Para analizar un gráfico de dispersión, debe tenerse en cuenta lo siguiente:

- **La forma:** si los puntos se sitúan en torno a una recta, quiere decir que las variables están correlacionadas linealmente. Si los puntos se distribuyen en torno a una recta de pendiente positiva, quiere decir que están positivamente correlacionados (es decir, cuando aumenta el valor de x , también lo hace el de y , como en el ejemplo principal). En caso contrario, están negativamente correlacionados. Por otra parte, si no se distribuyen en torno a ninguna línea, no están correlacionados. Además de la correlación lineal, expresada como una recta, también podría existir correlación exponencial, logarítmica o polinomial.

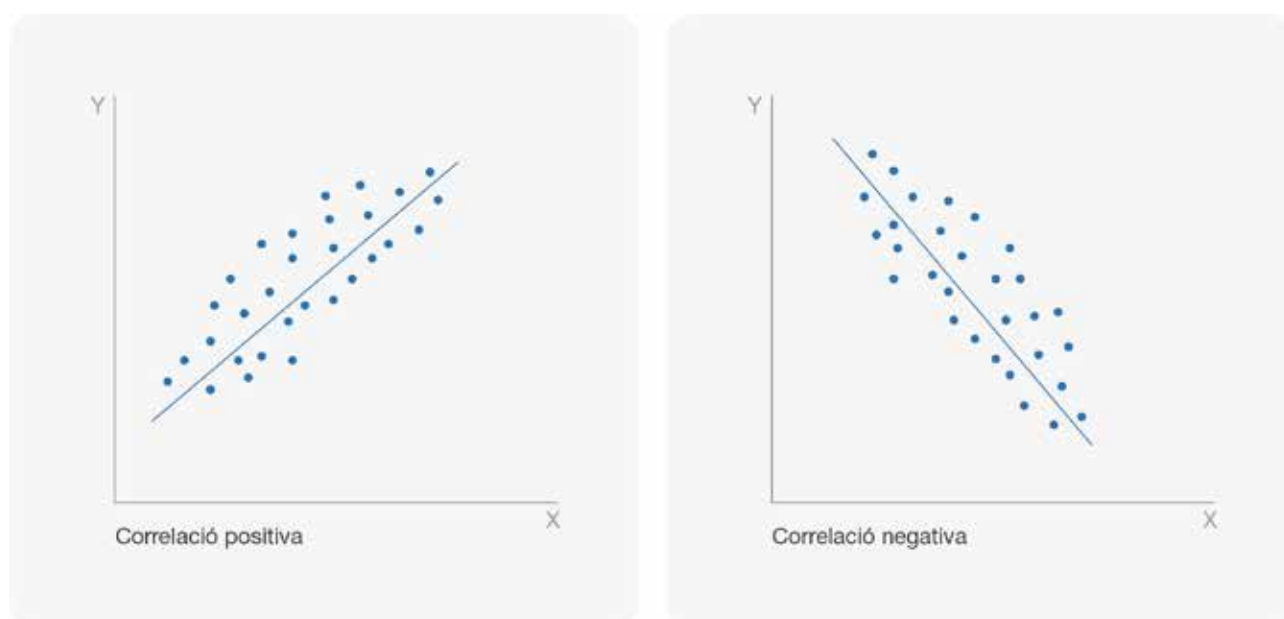


Figura 4.43. Ejemplos de correlación positiva y negativa.

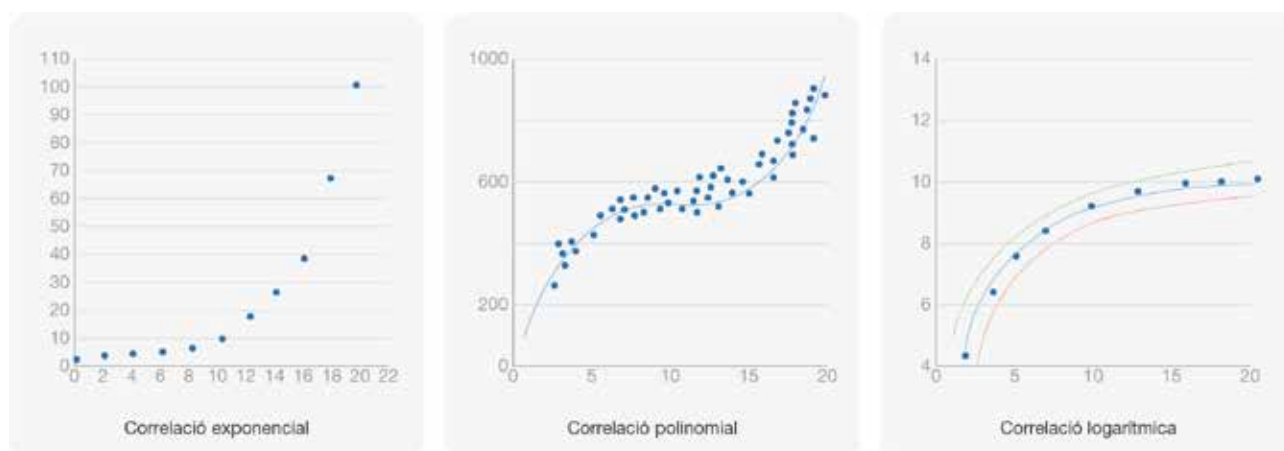


Figura 4.44. De izquierda a derecha, correlación exponencial, polinomial y logarítmica.

- **Las agrupaciones:** si muchos puntos se agrupan en torno a valores similares, quiere decir que tienen características similares (en inglés se denomina *cluster* a una agrupación de elementos).

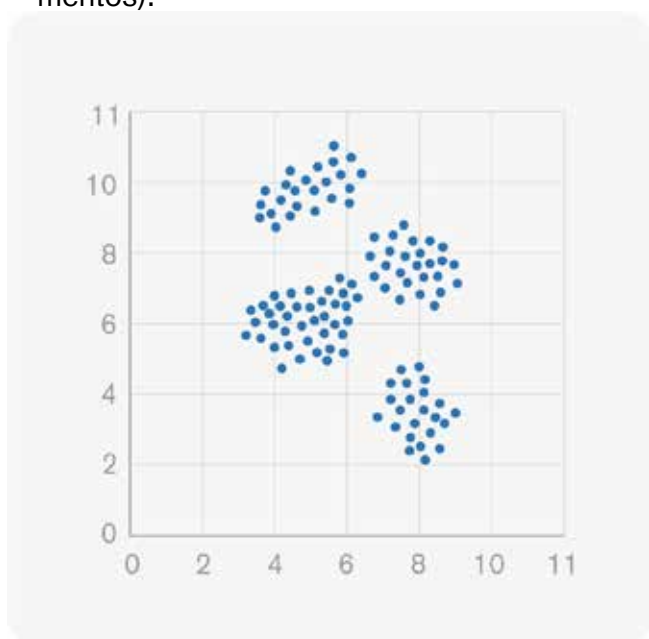


Figura 4.45. Ejemplo de gráfico de dispersión que permite ver agrupaciones.

- **Los valores atípicos** son aquellos valores que quedan fuera de la forma principal que se aprecia en un gráfico de dispersiones y corresponden a excepciones que habrá que saber explicar o bien eliminar del gráfico para no distorsionar su forma. En la siguiente figura corresponden a los puntos azules.

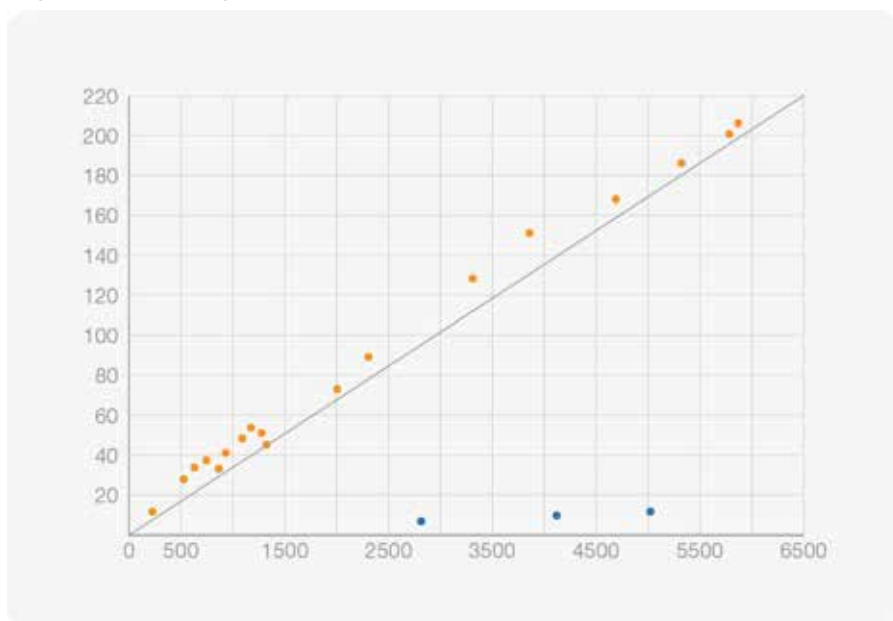


Figura 4.46. Ejemplo de gráfico de dispersión que permite ver, en azul, valores atípicos.

- **Los cuadrantes:** si dividimos el gráfico en cuatro cuadrantes, se pueden ver cuatro agrupaciones de elementos muy diferenciadas, en particular los valores altos por el eje x y el eje y, los que tienen valores bajos por los dos ejes, y los que tienen valores bajos por un eje y altos por el otro, tal y como se puede ver en la siguiente imagen..

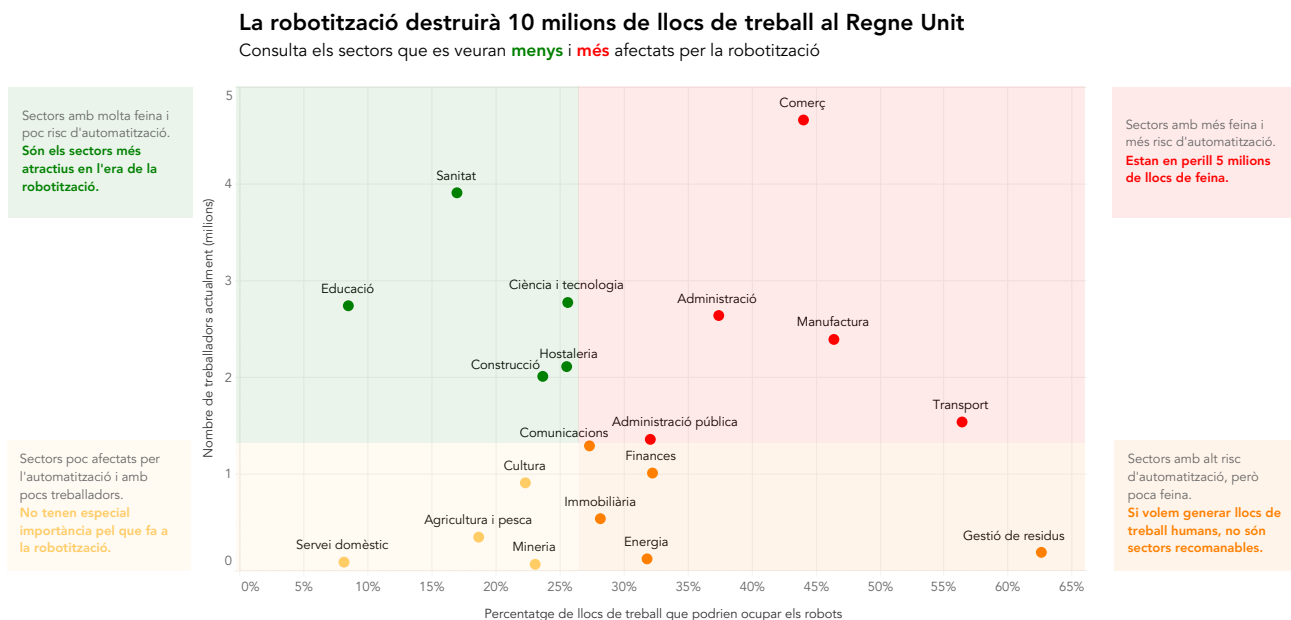


Figura 4.47. La robotización y el mercado laboral. Imagen cedida por OneTandem.
Fuente: <http://onetandem.com/blog/la-robotizacion-mercado-laboral/>.

Finalmente, debe tenerse en cuenta que añadir líneas de referencia en forma de retícula a los gráficos de dispersión es muy útil porque ayudan a estimar los valores de las variables de cada punto.

4.6.2. Gráfico de burbujas

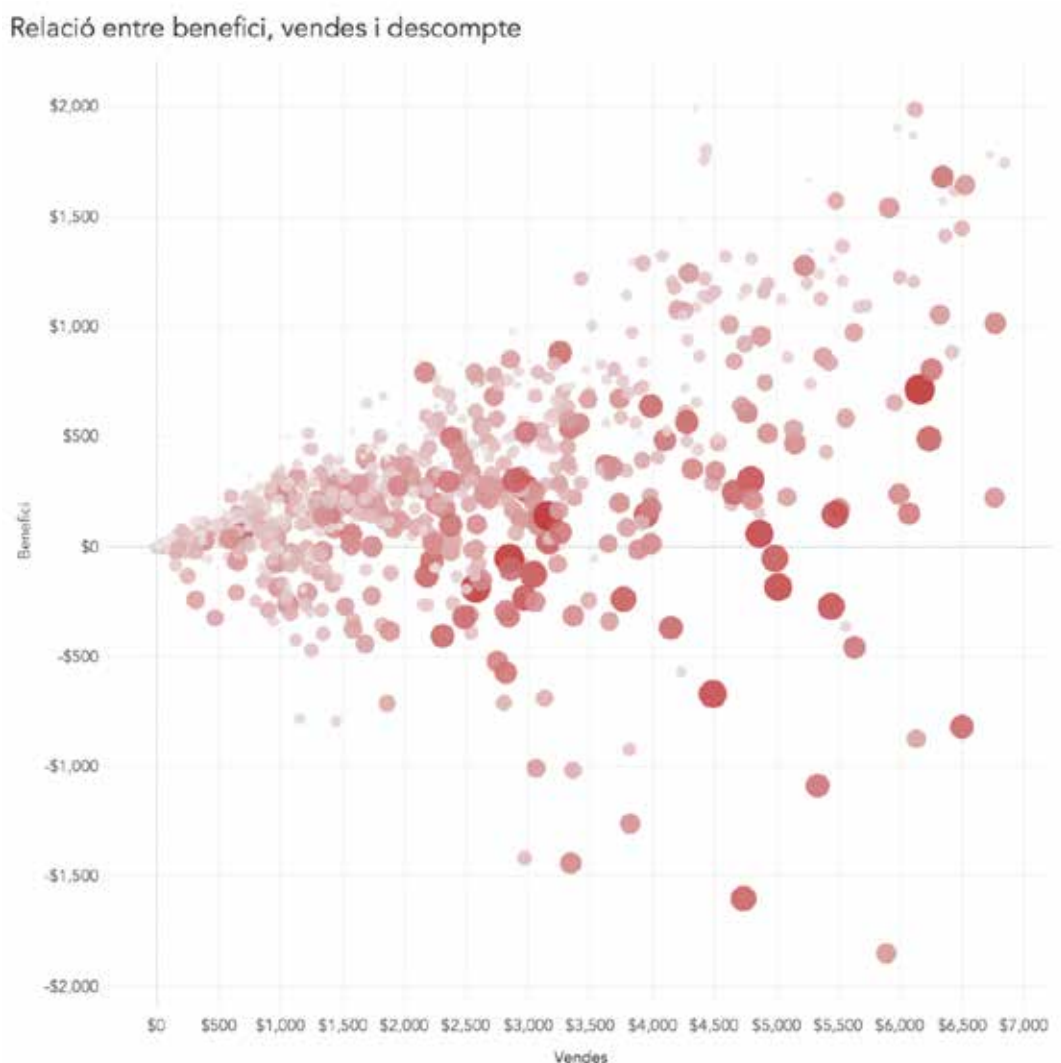


Figura 4.48. Gráfico de burbujas que muestra la relación entre beneficio, ventas y descuento para un conjunto de productos. La medida y color de la burbuja representa la magnitud del descuento. Se puede ver como los productos con beneficio negativo no siempre son los que tienen más descuento.

El gráfico de burbujas es un gráfico de dispersión en que el área y el color de los puntos pueden representar variables adicionales. De esta manera, se pueden representar hasta 5 variables diferentes al mismo tiempo:

- la posición según el eje x
- la posición según el eje y
- el color del punto
- la medida del punto
- la animación para representar la evolución temporal de cada punto

¿Cuándo se debe utilizar?

Cuando se quiere aprovechar lo que ofrece un gráfico de dispersión y, al mismo tiempo, se quiere tener más información sobre cada uno de los elementos representados utilizando la medida, el color y/o la animación.

Recomendaciones

Además de las recomendaciones para el gráfico de dispersión, en el caso del gráfico de burbujas es importante decidir qué variable se utiliza para cada uno de los atributos visuales utilizados (la posición marcada por los ejes x y y , la medida y el color).

Las variables principales del análisis se sitúan a lo largo del eje x y del eje y para ver más fácilmente si están relacionadas entre sí o para ver si hay elementos que se agrupan en torno a ciertos valores.

La medida se acostumbra a utilizar para representar una variable de la cual se quiere ver cuál es el elemento mayor.

Se utiliza el color para distinguir las categorías a las cuales pertenece cada elemento o para comparar valores grandes y pequeños de otra variable (en este caso habrá que utilizar una escala de colores numérica).

Y, finalmente, la animación se utiliza para representar la evolución temporal de los valores de los elementos, de modo que se pueda ver como los puntos se mueven arriba o abajo, a la derecha o a la izquierda, a medida que pasa el tiempo.

4.6.3. Coordenadas paralelas

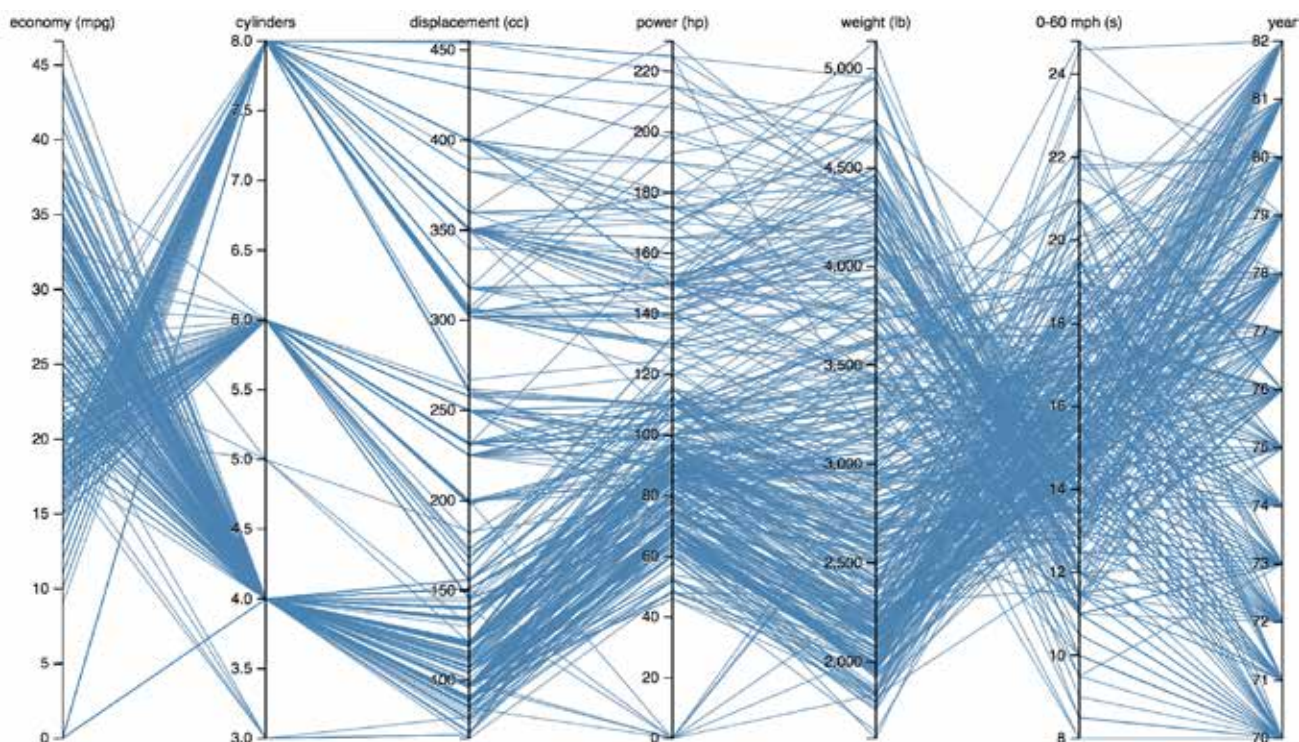


Figura 4.49. Comparación de modelos de coche a partir de diferentes características.

Fuente: <https://bl.ocks.org/mbostock/7586334>.

Las coordenadas paralelas permiten explorar las relaciones entre diferentes variables para un conjunto de elementos. Cada variable se representa con un eje vertical, y los elementos del conjunto de datos se representan uniendo los puntos que indican el valor de cada variable en cada eje.

¿Cuándo se debe utilizar?

Cuando se quieren explorar conjuntos de datos multidimensionales. En particular, con este gráfico se pueden descubrir tendencias o valores atípicos a través del uso de filtros interactivos.

Recomanacions

Las coordenadas paralelas son especialmente útiles en su versión interactiva. La interacción que más a menudo se utiliza es la de filtrado, que hace que el usuario seleccione un rango de valores en un eje y vea así como se comportan en el resto de variables. En la siguiente imagen, por ejemplo, se han filtrado los elementos por un rango determinado de la variable *weight* (lb). Entre otras cosas, se puede ver que estos elementos acostumbra a tener valores bajos de la variable *economy* (mpg) y, por lo tanto, es posible que haya una relación inversa entre estas dos variables.

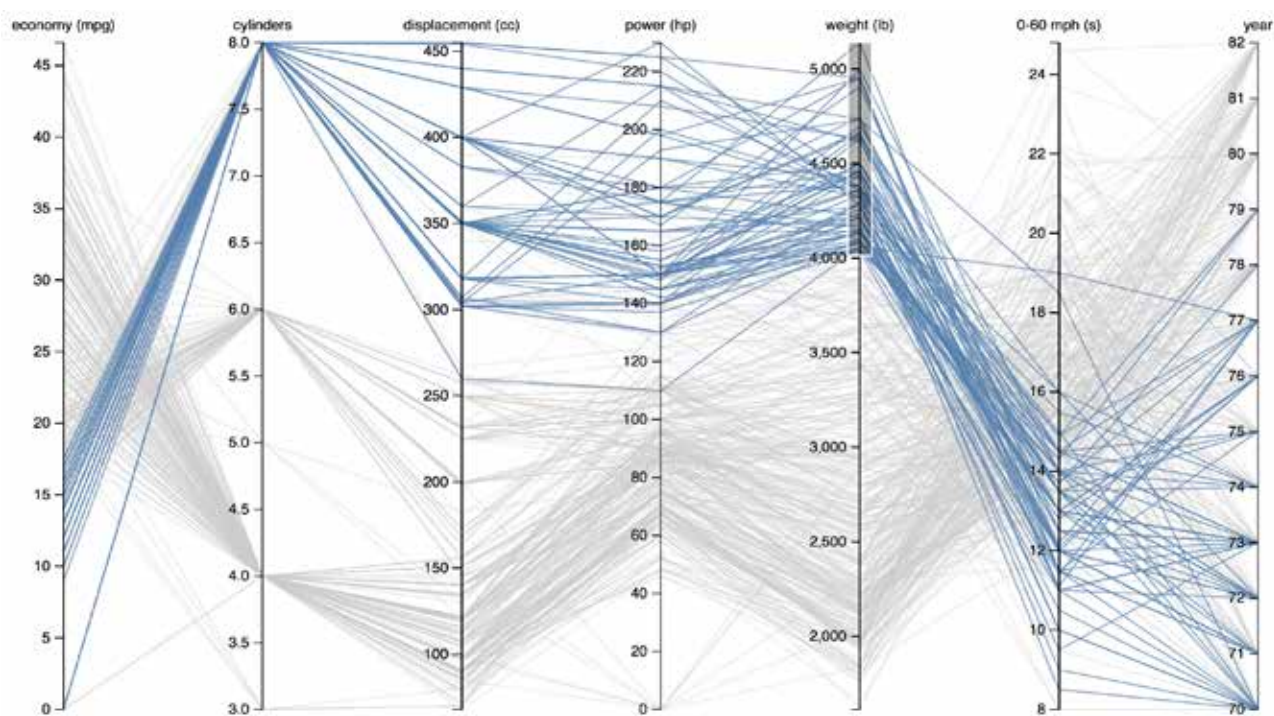


Figura 4.50. Comparación de modelos de coche a partir de diferentes características filtrado por la variable peso (*weight*).
Fuente: <https://bl.ocks.org/mbostock/7586334>.

4.7. Conexiones, relaciones y redes

Los gráficos de conexiones, relaciones y redes ayudan a entender la relación entre los elementos de un conjunto de datos. Por ejemplo, las relaciones entre miembros de una red social.

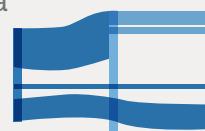
Diagrama nodo-arista

Para detectar patrones que surgen de las relaciones entre los elementos del conjunto de datos



Diagrama de Sankey

Para representar el cambio de una variable a través del flujo entre diferentes estados.



4.7.1. Diagrama de nodos y aristas

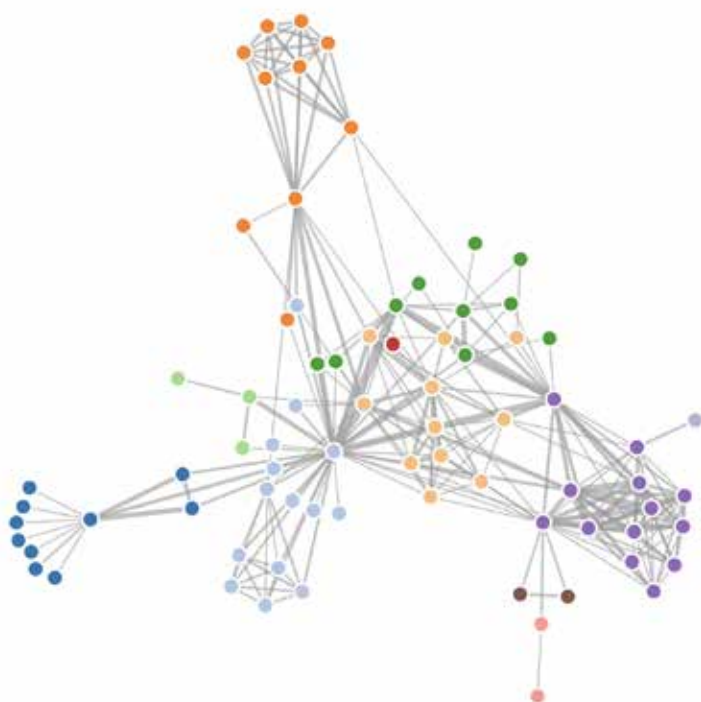


Figura 4.51. Red de personajes en la obra *Los miserables*. Cada unión indica que aparecen en la misma escena.
Fuente <https://bl.ocks.org/mbostock/4062045>.

Los diagramas nodo-arista utilizan círculos para representar elementos, que se denominan nodos, y aristas que se representan como líneas entre nodos y que indican una relación entre ellos.

Las conexiones se pueden establecer partiendo de diferentes factores, como, por ejemplo, relaciones de amistad entre miembros de una red social, o personajes que aparecen en las mismas escenas de una película. Pueden tener una dirección que indica la naturaleza de la relación y que se representa con aristas que acaban en forma de flecha. Por ejemplo, en una red como Facebook la relación entre personas es bidireccional, mientras que en Twitter, no (el usuario A puede seguir al usuario B, pero no hace falta que el usuario B siga al usuario A).

La entidad matemática que surge de establecer la conexión entre nodos a través de aristas se llama “grafo”.

¿Cuándo se debe utilizar?

Cuando se quieren detectar patrones que surgen de las relaciones entre los elementos del conjunto de datos. En general, nos ayuda a encontrar elementos muy conectados o elementos que quedan más aislados dentro de la red. En el ejemplo se pueden ver hasta ocho grupos diferentes de personajes en *Los miserables*. Algunos quedan mezclados en el centro de la red mientras que otros muestran comunidades muy claras y revelan que estos personajes no interactúan con otros personajes de la obra.

Recomendaciones

La complejidad principal recae en la manera de situar los nodos en el espacio. Generalmente conviene seguir los siguientes principios:

- Minimizar los encabalgamientos entre nodos y los cruces entre aristas
- Incluir toda la red en la misma pantalla.
- Favorecer el descubrimiento de grupos de elementos similares (*clusters*).

El cumplimiento de estos principios suele conseguirse aplicando los llamados algoritmos dirigidos por fuerzas que, según la estructura de la red, deciden la mejor posición para cada nodo.

Además, a menudo se utilizan variables del conjunto de datos para modificar la medida o el color de los nodos y/o las aristas.

4.7.2. Diagrama de Sankey

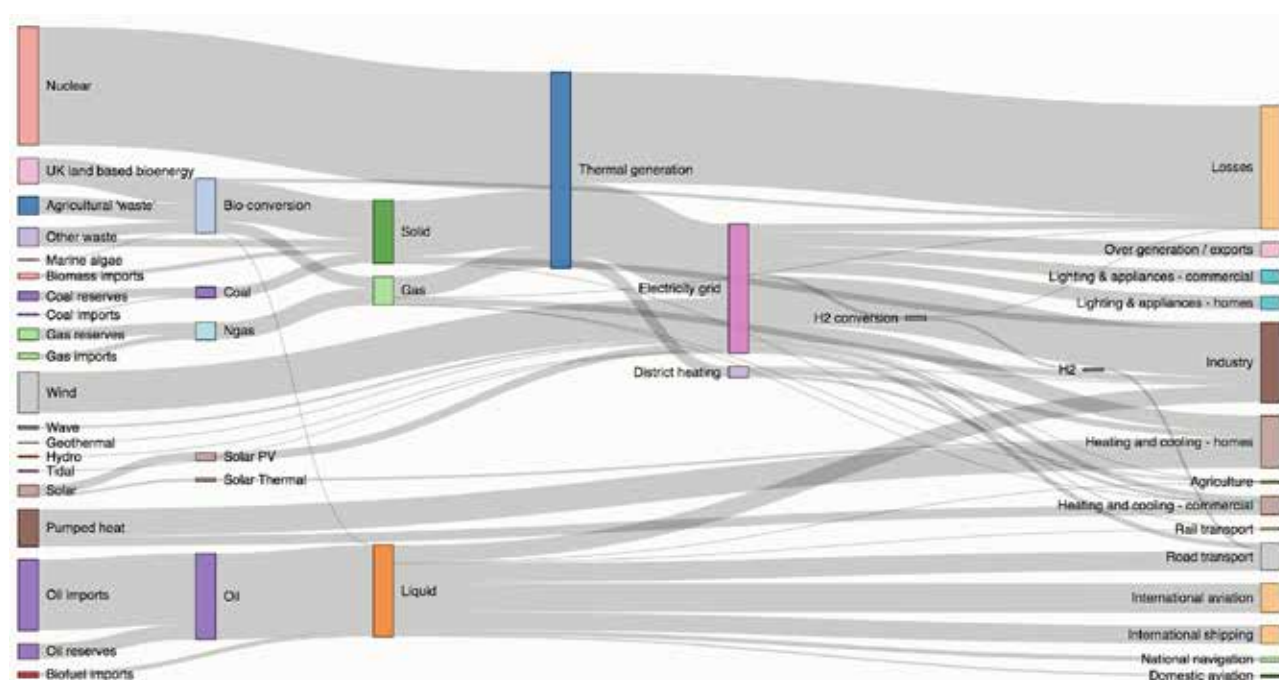


Figura 4.52. Diagrama de Sankey que muestra la cadena de producción y consumo de energía eléctrica.

Fuente: <https://bost.ocks.org/mike/sankey/>.

El diagrama de Sankey muestra diferentes categorías o estados a través de los cuales una variable va cambiando de valor. También muestra que una categoría es la suma de los valores de diferentes categorías, tal y como se aprecia en el ejemplo.

¿Cuándo se debe utilizar?

Cuando se quiere representar el cambio de una variable a través del flujo entre diferentes estados. El diagrama de Sankey muestra cómo una variable se va descomponiendo en los diferentes pasos y/o categorías. Por ejemplo, cuando se representa la cadena de producción y consumo de energía eléctrica, como se puede ver en la imagen superior.

Recomendaciones

Lo más importante a la hora de crear un diagrama de Sankey es decidir en qué orden vertical se sitúan las diferentes categorías que se representan en cada uno de los estados. De la elección de la posición resultarán diferentes grados de encabalgamiento entre aristas que dificultarán más o menos la lectura del gráfico.

5. Principios del diseño

Al desarrollar una visualización de datos, además de hacer un análisis de datos, conviene pensar cuál es la mejor manera de representarlos. Es por eso que conviene tener unos fundamentos de los principios del diseño, que nos permitirán entender por qué tomamos ciertas decisiones de representación visual y, por lo tanto, podremos actuar en consecuencia. En este capítulo trataremos los principales aspectos del diseño que deben tenerse presentes en la visualización de datos.

Previamente, se debe tener en cuenta que las visualizaciones de datos de la Generalitat de Catalunya se elaboran de acuerdo con la normativa de identidad corporativa. Por lo tanto, los departamentos, organismos autónomos adscritos y las empresas públicas que dependen deben aplicar las pautas siguientes, definidas en la web Identidad corporativa:

- Gráficos y tablas <http://identitatcorporativa.gencat.cat/ca/aplicacions/grafics-i-taules/>
- Infografías <http://identitatcorporativa.gencat.cat/ca/aplicacions/infografies/>

5.1. Pensamiento visual

La vista es nuestro sentido más importante. Un 70% de los órganos receptores del cuerpo humano están concentrados en los ojos. Además, los sensores visuales están conectados directamente con el sistema cognitivo. Por lo tanto, si entendemos el funcionamiento de la percepción visual, podremos diseñar visualizaciones que aprovechen al máximo nuestras capacidades cognitivas. Veámoslo con un ejemplo.

La siguiente tabla de valores representa las ventas trimestrales de una empresa a dos segmentos de clientes (pequeña empresa y gran empresa).

	2014 T1	2014 T2	2014 T3	2014 T4	2015 T1	2015 T2	2015 T3	2015 T4	2016 T1	2016 T2
Gran empresa	\$13.992	\$27.245	\$32.993	\$54.206	\$22.280	\$18.980	\$35.252	\$52.245	\$26.578	\$42.418
Petita empresa	\$21.965	\$47.852	\$96.104	\$100.177	\$40.963	\$56.330	\$79.399	\$89.844	\$46.672	\$70.139

Figura 5.1. La mesa muestra las ventas trimestrales de una empresa a clientes de pequeña empresa y de gran empresa.

La tabla nos permite localizar de manera sencilla un valor concreto. Pero si queremos extraer más información, podemos optar por la siguiente visualización:

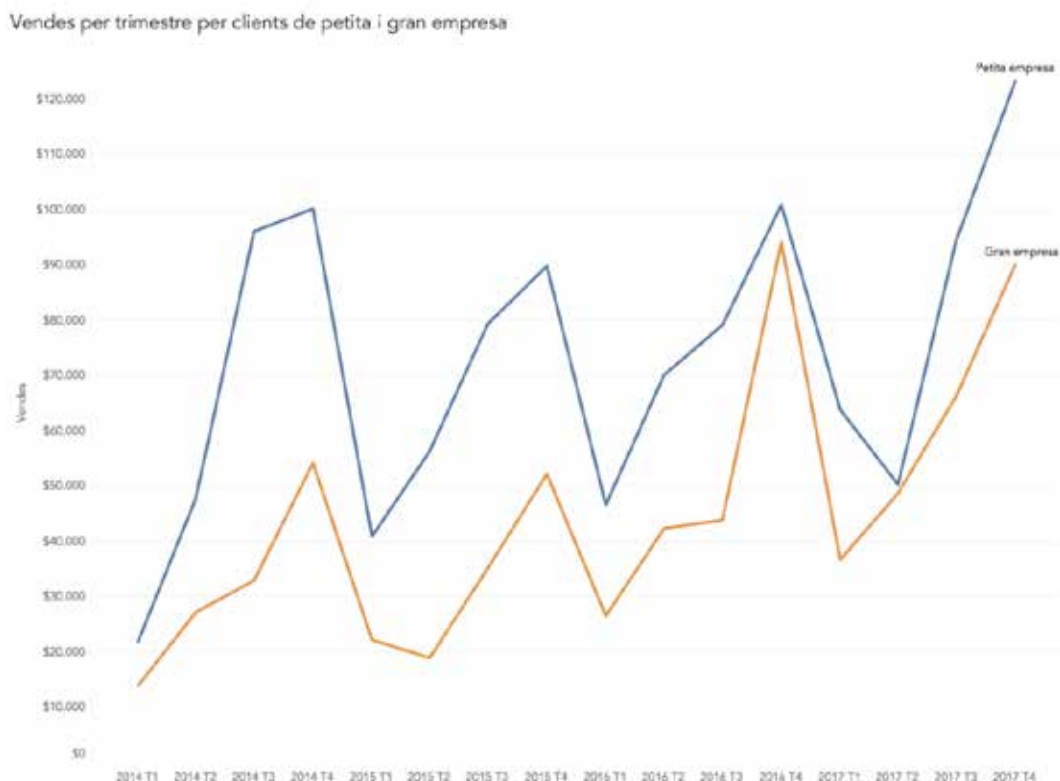


Figura 5.2. El gráfico de líneas muestra las ventas trimestrales de una empresa a clientes de pequeña empresa y de gran empresa.

Gracias a la representación con un gráfico de líneas podemos ver la tendencia cíclica y globalmente ascendente de las ventas. ¿Por qué un gráfico de líneas lo permite y una tabla no? Para entenderlo, deben tenerse presentes los tres tipos de memoria que actúan en el cerebro:

- memoria icónica
- memoria a largo plazo
- memoria de trabajo

La memoria icónica se ocupa de la información que recibe el cerebro procedente de los receptores visuales. Se procesa rápidamente y sin que seamos conscientes: decimos procesamiento “preatentivo”. No pretende el análisis exhaustivo de todo lo que vemos, sino que extrae un subconjunto de elementos relevantes como el color o la forma.

La memoria a largo plazo es aquella de acceso más lento, que nos permite almacenar una información que hemos tratado mucho o que hemos estudiado detenidamente. Por ejemplo, la llave de acceso a nuestro correo electrónico, o el material que hemos estudiado para un examen.

La memoria de trabajo procesa información proporcionada por la memoria icónica y la de largo plazo. Tiene una capacidad limitada: diferentes estudios⁸ dicen que puede llegar a almacenar entre 5 y 9 elementos (en función de la persona y el tipo de tarea que esté realizando). Esta limitación hace que entendamos mejor los datos del ejemplo cuando estas están codificadas en un gráfico de líneas (un elemento) en lugar de estar codificadas en una tabla (tantos elementos diferentes como valores).

Así pues, a la hora de hacer buenas visualizaciones de datos, hará falta:

- escoger los atributos visuales adecuados para aprovechar la capacidad de procesamiento preatentivo de la memoria icónica (en particular, color y forma)
- respetar las limitaciones de la memoria de trabajo
- ser consistente con el uso de la maquetación y formado de los elementos para favorecer los mecanismos de la memoria a largo plazo.

5.2. Color

El color es un atributo visual *preatentivo* (procesado por la memoria icónica), y el más utilizado juntamente con la forma. Para usarlo para comunicar información debe distinguirse entre datos cuantitativos o bien datos categóricos.

En el caso de disponer de datos cuantitativos, tenemos dos opciones:

- utilizar una escala donde se utilice un único color con diferentes saturaciones, tal y como se puede ver en la figura 5.3.
- en caso de que dispongamos de un valor central, como por ejemplo el cero, utilizar una escala dicotómica, como podemos ver en la figura 5.4.

100	100	100	100	100	75.3
41.8	82.8	79.8	71.1	70.9	51.9
68.4	80.7	60.6	73.6	61.5	67.1
66.8	79.4	65.3	67.5	68.1	55.9
79.1	96.6	50.8	56.2	66.5	55.2
52.1	88.5	57.1	46.2	44.2	68.1
48.5	66.7	49.3	60	44.9	100

Figura 5.3. Escala de un único color, que utiliza la saturación para distinguir elementos.

⁸ Miller, George Armitage. *The magical number seven, plus or minus two: Some limits on our capacity for processing information*. Psychological Review, 1956



Figura 5.4. Escala dicotómica de tres colores, con diferentes saturaciones.

Aunque el uso del color es una manera de codificar los datos muy intuitiva, no es una buena elección cuando se necesita saber cuánto mayor es un valor respecto de otro. Por ejemplo, podemos percibir que un color es más oscuro que otro, pero nos resulta prácticamente imposible percibir que un color es dos veces más oscuro que otro.

En el caso de disponer de datos categóricos, el color nos servirá para distinguir las categorías entre sí. Por lo tanto, habrá que buscar colores que sean bien diferentes.



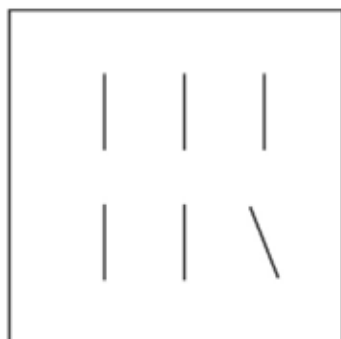
Figura 5.5. Escala de colores categórica proporcionada por la librería D3js. Fuente: <https://d3js.org/>

Aunque exista un gran número de colores, es difícil encontrar más de diez colores que sean marcadamente diferentes entre sí. Por lo tanto, el color no es un buen atributo visual en caso de que tengamos demasiadas categorías por representar. Asimismo, recordamos que la memoria de trabajo tiene dificultades trabajando con más de 5 a 9 elementos diferentes (en particular, más de 5 a 9 colores). En estos casos, hará falta crear una agrupación alternativa de datos, o bien utilizar mecanismos de filtraje que nos permitan ir de la globalidad a la particularidad interactivamente.

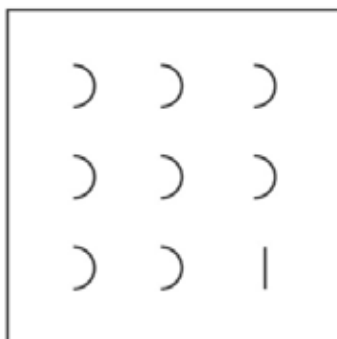
Finalmente, hace falta tener en cuenta que aproximadamente de un 5% a un 10% de la población sufre daltonismo y, por lo tanto, tiene dificultades para distinguir el rojo y el verde. Existen escalas de colores especialmente pensadas para estas personas, que recomendamos utilizar.

5.3. Forma

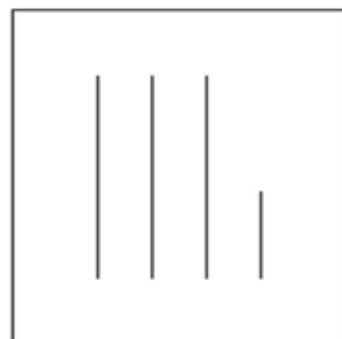
La forma es el otro atributo visual *preatentivo* más utilizado, además del color. La siguiente imagen muestra diferentes maneras de utilizar la forma:



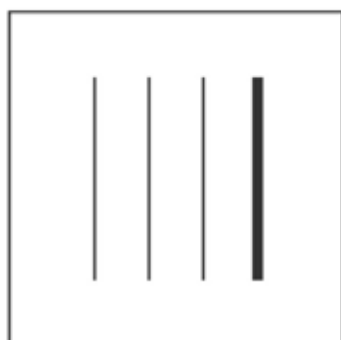
Orientació i coliniaritat



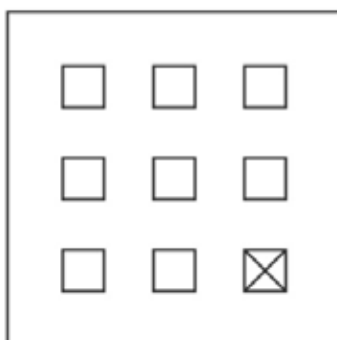
Curvatura



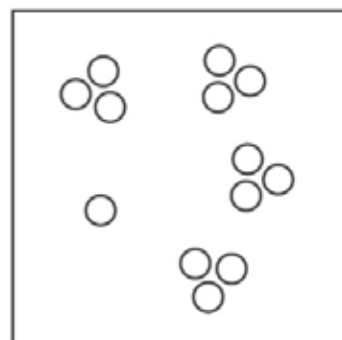
Llargada



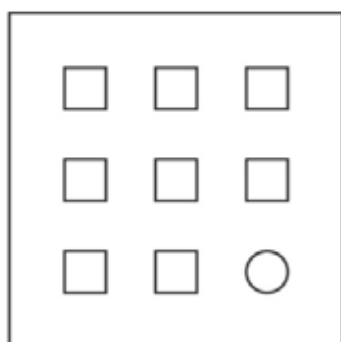
Amplada



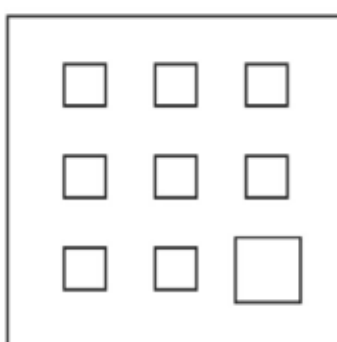
Amb marques



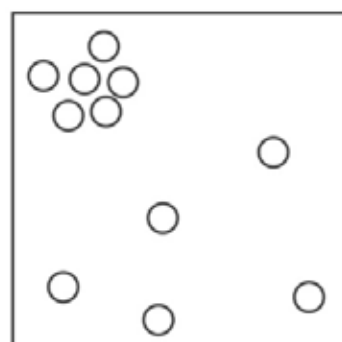
Numeració



Formes diferents



Mida



Agrupació espacial

Figura 5.6. "Diferentes maneras de utilizar el atributo preatentivo de la forma".

La longitud es el atributo que nuestro cerebro procesa más fácilmente. Por eso los gráficos de barras son tan populares. La medida también se utiliza bastante, por ejemplo en un gráfico de burbujas. La agrupación espacial funciona muy bien en un gráfico de dispersión donde podemos detectar un grupo de elementos claramente diferenciados del resto. El resto de atributos se utilizan en menor grado, pero también nos pueden resultar útiles en ciertas situaciones.

5.4. Interacción

Los elementos de interacción nos permiten adaptar la visualización de datos a unas necesidades informativas concretas. Ben Shneiderman definió en el año 1996 el mantra de la búsqueda visual de información (*Information Seeking Mantra*) como: “mostrar una perspectiva general, hacer zoom y filtrar, y después proporcionar detalles bajo demanda”. Esta manera de entender la visualización requiere del uso de la interacción. A continuación, explicamos los tres tipos de interacciones más comunes.

Descripciones emergentes

Las llamadas descripciones emergentes (*tooltips*) son pequeñas cajas con texto y/o visualizaciones que aparecen cuando seleccionamos un elemento de nuestra visualización. Son el tipo de interacción básica para proporcionar detalles sobre un elemento por el cual se interesa al usuario.

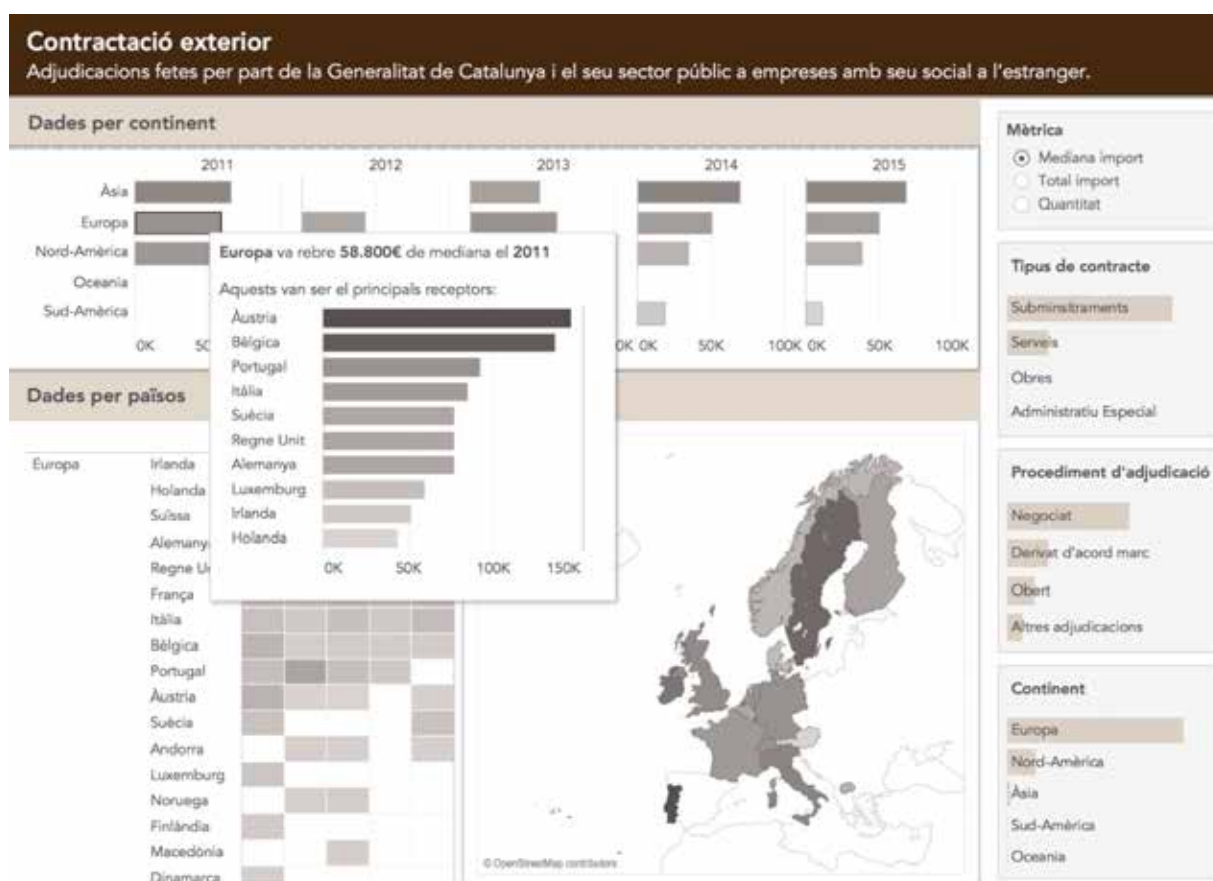


Figura 5.7. Cuando pasamos por encima de la barra “Europa”, una descripción emergente nos muestra un gráfico de barras con el detalle de los países receptores.

Enlace y marcaje

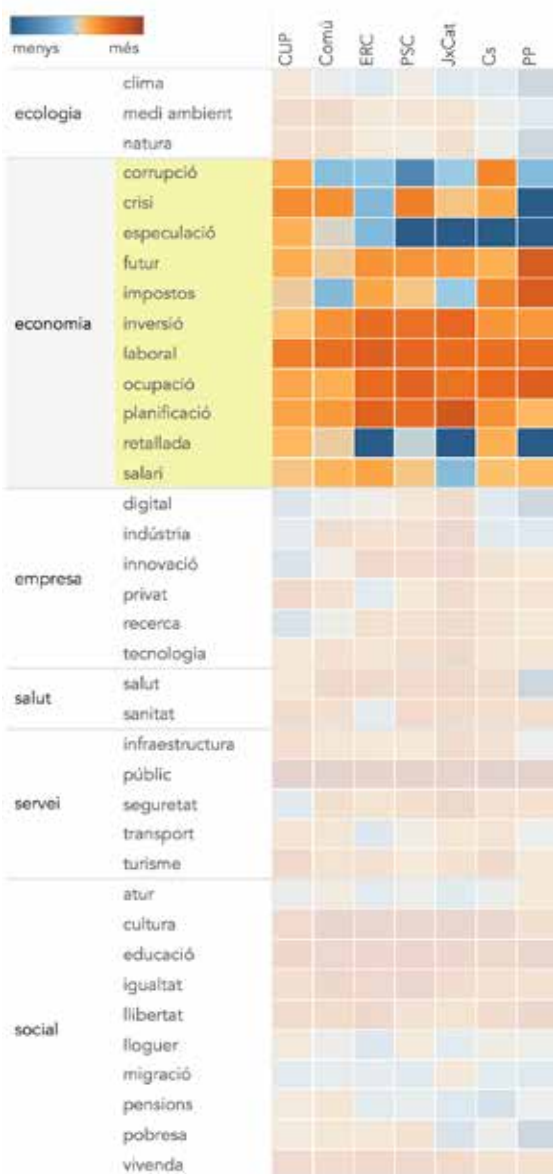
El marcaje (*brushing*) es una técnica de interacción que permite que el usuario seleccione un conjunto de elementos de la visualización y que hace que queden resaltados en unas o más visualizaciones. Cuando hay más de una visualización sobre los mismos datos disponible, el hecho de que todas ellas resalten los mismos elementos se llama “enlace” (*linking*).

En la siguiente imagen podemos ver un ejemplo de visualización que utiliza la técnica del enlace y marcaje. En la visualización titulada *De que hablan los partidos del 21-D* se representa, mediante un gráfico de intensidades por colores, la frecuencia con que cada partido político utiliza una palabra en su programa electoral. La intensidad de los colores nos ayuda a ver si una palabra se ha utilizado más o menos. Paralelamente, unos gráficos de barras muestran la utilización de las diferentes palabras ordenadas de más a menos utilizadas. Como se puede apreciar en la imagen, cuando se resalta un conjunto de palabras, estas se marcan también en los gráficos de barras y nos permiten ver si estas son las más o menos utilizadas en cada programa electoral.

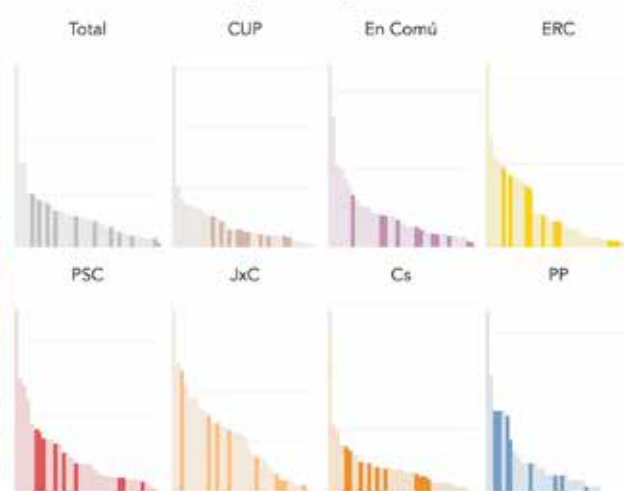
De què parlen els partits del 21-D?

Visualitzem el nombre de mencions de diferents conceptes en els seus programes

Quins termes es mencionen més?



Quina és la distribució per cada partit?



Limitacions de l'anàlisi

Aquesta visualització de dades és útil per obtenir una panoràmica general dels programes dels diferents partits polítics. Ara bé, no sempre ho farà de manera acurada. Que un partit mencioni poc un terme no vol dir necessàriament que no li doni importància. Per exemple, un partit potser menciona poc el terme "corrupció", però les seves polítiques són molt fermes contra la corrupció.

Per entendre bé els programes, cal llegir-los. Per això, aquí us oferim enllaços a cada programa dels partits polítics que es presenten el 21-D: ..

CUP Comú ERC PSC
JxCat Cs PP

Metodologia

Per un recull de 45 paraules, hem comptat quantes vegades es mencionen a cada programa dels partits polítics que es presenten el 21-D. Hem ponderat els resultats per tal de poder comparar independentment de la longitud dels programes. Gràcies a <https://twitter.com/danirdelpz> per la recopilació i tractament de les dades.

Visualització realitzada per OneTandem

ONETANDEM

Figura 5.8. Ejemplo de enlace y marcaje utilizado en la visualización titulada "De qué hablan los partidos del 21-D" hecha por la empresa OneTandem.

Fuente: <https://public.tableau.com/profile/onetandem#!/vizhome/ProgrameseleccionsGeneralitatCatalunya21D/heatmap>

5.5. Recomendaciones generales

A continuación se describen una serie de aspectos a tener en cuenta a la hora de diseñar nuestras visualizaciones.

5.5.1. Menos es más

Edward Tufte es uno de los autores más respetados de la visualización de datos. Su libro *The visual display of quantitative information*⁹ se considera todo un clásico de la disciplina. En este libro, describe la excelencia gráfica de la siguiente manera:

“Por encima de todo, se tienen que representar datos (...). La excelencia gráfica es aquella que da al usuario el mayor número de ideas, en el periodo más corto de tiempo, utilizando la mínima cantidad de tinta, en el espacio más pequeño posible”

El siguiente GIF animado muestra claramente cómo aplicar este principio al diseño de gráficos:

<http://www.darkhorseanalytics.com/blog/data-looks-better-naked>

5.5.2. Equilibrio entre funcionalidad y estética

El equilibrio entre funcionalidad y estética está supeditado a los objetivos de la visualización. Generalmente, si buscamos la complicidad del público general, la estética puede llegar a prevalecer por encima de la funcionalidad, siempre y cuando seamos rigurosos. Por ejemplo, el Better Life Index de la OCDE, mostrado en la figura siguiente, es un ejemplo de visualización donde la estética busca la complicidad del usuario, por encima de la funcionalidad analítica.

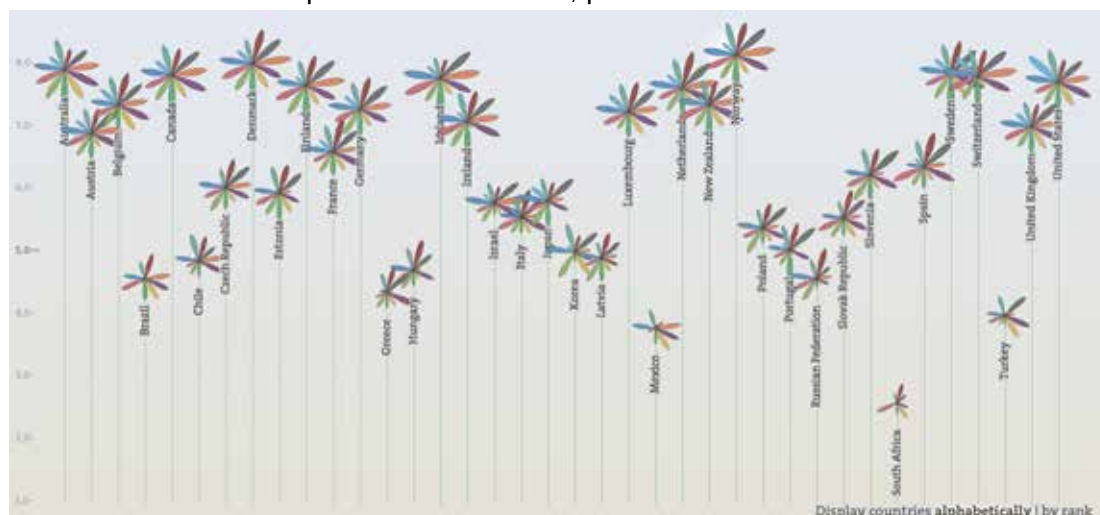


Figura 5.9. Captura parcial de la visualización del "Better Life Index" de la OCDE. Fuente: <http://www.oecdbetterlifeindex.org/>

Sin embargo, es muy importante entender que el objetivo principal de la visualización de datos en un entorno analítico debe ser generar conocimiento.

⁹ Tufte, Edward R. *The visual display of quantitative information*. Graphic Press, 1983

5.5.3. La forma sigue a la necesidad

La forma escogida dependerá siempre del objetivo informacional. No podemos separar la forma de la necesidad. Alberto Cairo pone un buen ejemplo en su libro *El arte funcional* en el que se muestra la visualización siguiente:

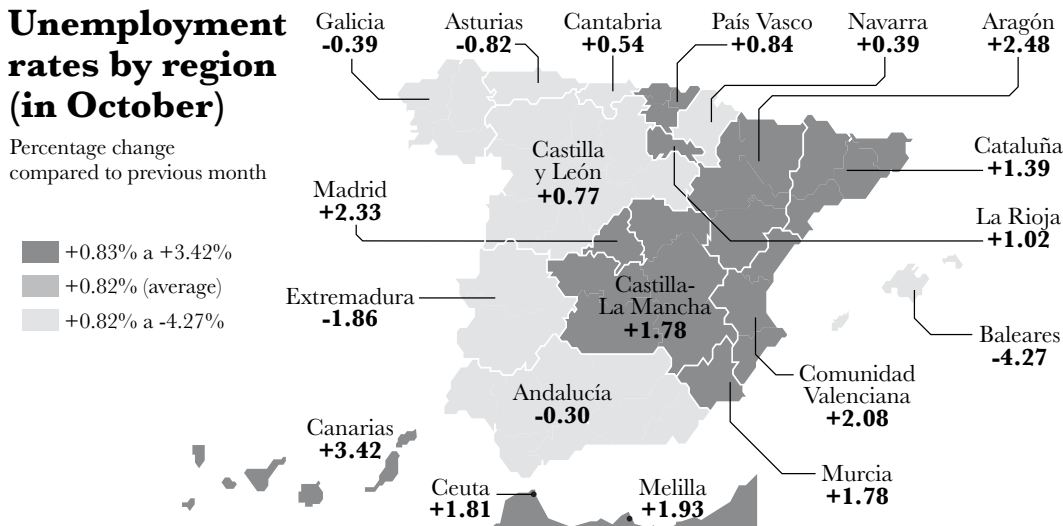


Figura 5.10. Representación de la tasa de paro por comunidad autónoma. Imagen cedida por Alberto Cairo, de su libro *El arte funcional*¹⁰.

El objetivo de la visualización es descubrir qué comunidades autónomas han mejorado más y cuáles están teniendo más problemas en comparación con el mes anterior. Sin embargo, la forma escogida hace que esta tarea sea complicada, ya que solo se han utilizado tres gradaciones de color y, dentro de las comunidades con un mismo color, no queda más remedio que inspeccionar los números, intentar memorizarlos y después tratar de inferir la orden de los valores.

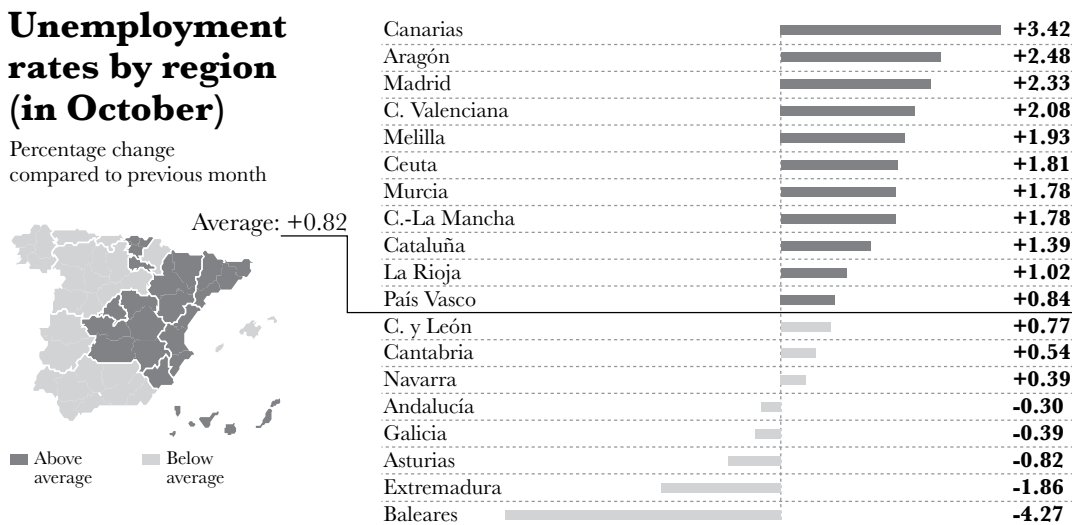


Figura 5.11. Rediseño planteado por Alberto Cairo. Imagen cedida por el mismo autor, de su libro *El arte funcional*.

¹⁰ Cairo, Alberto. *El arte funcional. Infografía y visualización de información*. Alamut 2011.

El rediseño del mapa propone la utilización de un gráfico de barras que permite ordenar fácilmente las comunidades autónomas según el cambio en la tasa de paro. El mapa se mantiene, pero no para mostrar valores numéricos, sino para evidenciar patrones geográficos (que tendría que ser siempre la utilidad de un mapa).

5.5.4. Texto

A veces, en una visualización de datos, no se da al texto la importancia que merece. Por ejemplo, un titular adecuado puede ser la clave para atraer la atención del lector. Asimismo, los textos que acompañan los gráficos en forma de titulares o anotaciones son necesarios para asegurar la correcta comprensión. Incluso el texto puede ser el elemento principal de la visualización. Lo podemos ver claramente con un ejemplo.

El gráfico siguiente muestra el número de pasos que ha realizado una persona a lo largo de un día.

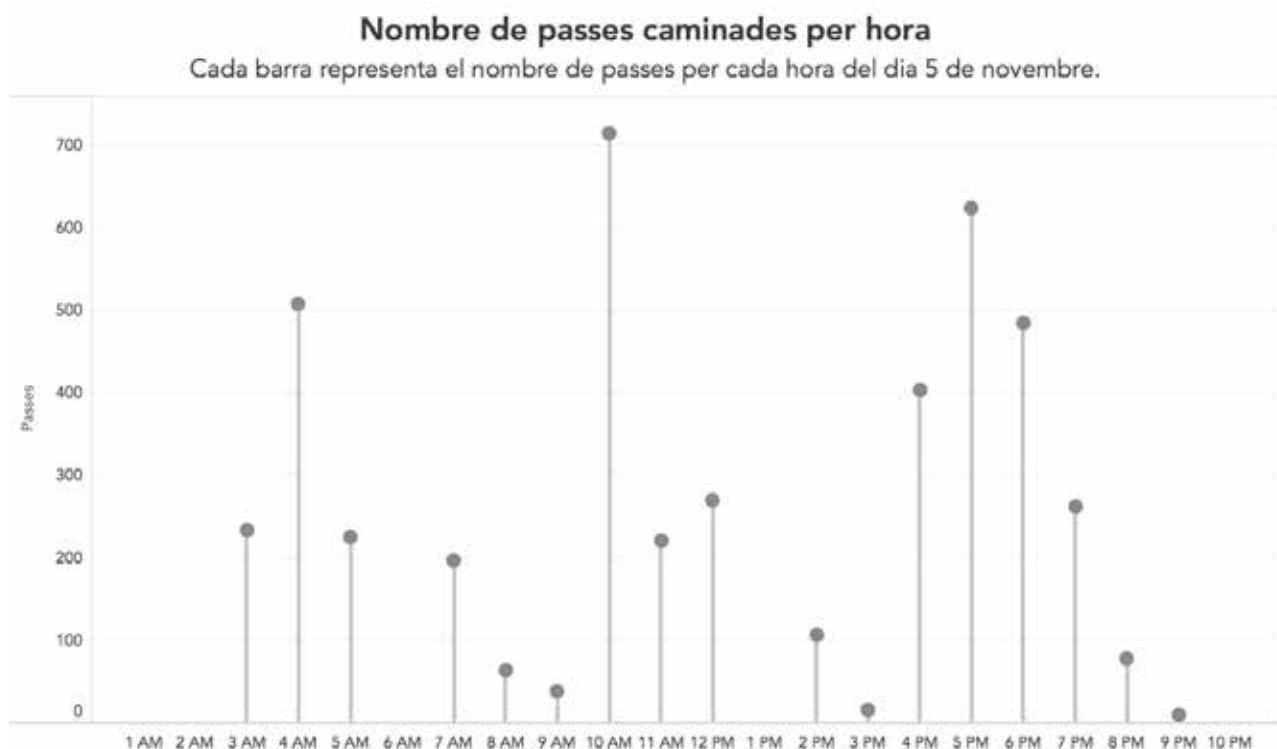


Figura 5.12. Número de pasos en un día.

Se utiliza un texto genérico que describe el contenido del gráfico. El texto no es incorrecto, pero no ayuda a entender el significado de los datos. Fijaos cómo cambia el mismo gráfico, utilizando otros textos:

El camí de la vida

La història del dia en què **va néixer el meu fill Teo**, explicada a través de les passes que vaig fer cada hora.

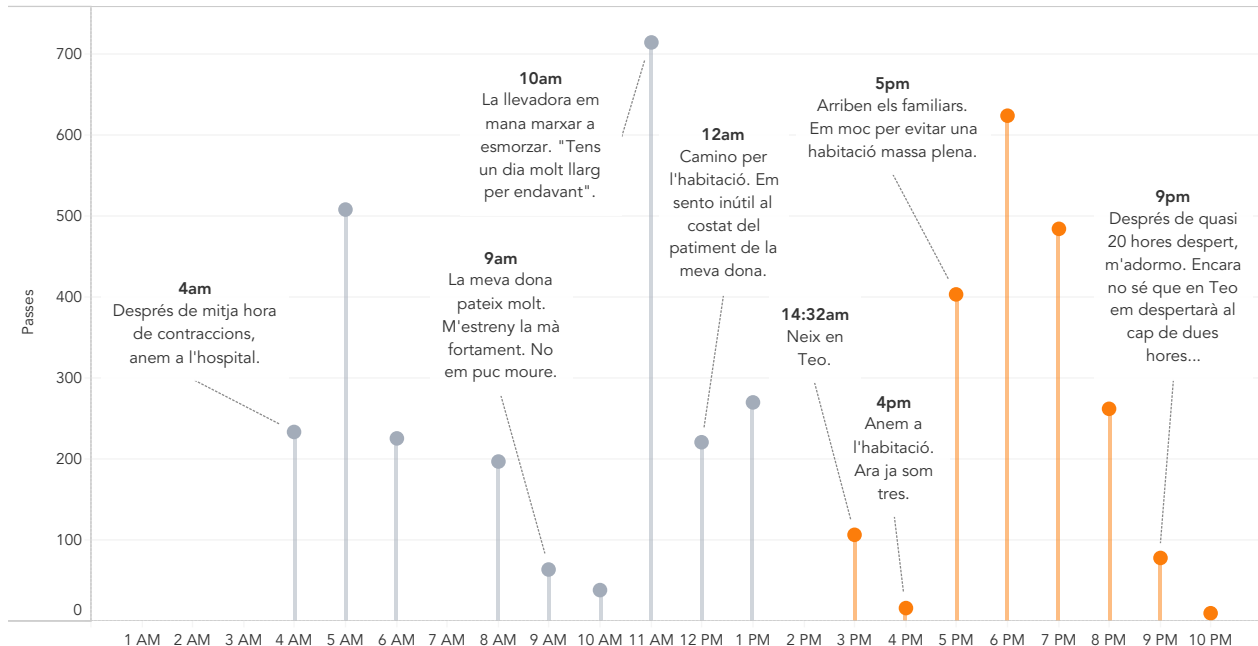


Figura 5.13. "El camino de la vida". Imagen cedida por OneTandem.

Un gráfico simple se convierte en una historia personal, incluso emotiva. Los datos son exactamente los mismos, pero, gracias al texto, los situamos en contexto, entendemos qué hay detrás y qué significado tienen. Como vemos, en una visualización, el texto puede ser incluso más importante que los propios datos numéricos.

5.5.5. Maquetación

Cuando hablamos de maquetación nos referimos a la disposición de los diferentes elementos de la visualización de datos: gráficos, textos, imágenes, filtros, etc. En general, recomendamos reservar espacio para:

- la información sobre la metodología y las fuentes de datos, que generalmente irá al pie de la visualización
- los filtros y selectores, que generalmente irán agrupados en una columna a la derecha o la izquierda, o en una fila en la parte superior
- los mecanismos de ayuda, que generalmente irán al lado de cada gráfico (en este caso también se puede situar un icono de ayuda general en la esquina superior derecha).

También conviene recordar que generalmente el público lee de arriba a abajo y de izquierda a derecha y, por lo tanto, la información más importante se debe situar en la esquina superior izquierda.

6. Bibliografía

Qlik. *Data Equality Campaign*. <<http://dataequality.org/>> [recurso electrónico]. [Consultado: noviembre 2017].

Ware, Colin. *Information Visualization, Third Edition: Perception for Design*. Morgan Kaufmann, 2012

Wikipedia. *Interactive Storytelling*.

<https://ca.wikipedia.org/wiki/Projecci%C3%B3_cartogr%C3%A0fica> [Consulta: noviembre 2017]

Nussbaumer, Cole. *A Data Visualization Guide for Business Professionals*. Wiley, 2015

Wikipedia, *Five Ws*. <https://en.wikipedia.org/wiki/Five_Ws> [Consulta: noviembre 2017]

Wikipedia, *Projecció cartogràfica*.

<https://ca.wikipedia.org/wiki/Projecci%C3%B3_cartogr%C3%A0fica> [Consulta: noviembre 2017]

Miller, George Armitage. *The magical number seven, plus or minus two: Some limits on our capacity for processing information*. Psychological Review, 1956

Tufte, Edward R. *The visual display of quantitative information*. Graphic Press, 1983

Cairo, Alberto. *El arte funcional. Infografía y visualización de información*. Alamut 2011.

A continuación citamos algunos libros que pueden ser de utilidad para profundizar en el conocimiento de la materia de la visualización de datos.

Few, Stephen. *Now You See It*. Analytics Press, 2009.

Meirelles, Isabel. *Design for Information*. Rockport Publishers, 2013.

Cotgreave, Andy; Shaffer, Jeffrey; Wexler, Steve. *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business*. Wiley, 2017.

Cairo, Alberto. *The Truthful Art*. [s. n.] 2016.

7. Anexos

7.1. Glosario

Análisis de *clústers*: El análisis de *clusters* es una técnica de análisis de datos que permite, automáticamente, segmentar un conjunto de datos en subgrupos (denominados *clusters*) que tengan valores similares a sus variables.

Dimensión: Aquellas variables de tipos textuales o categóricos que nos permiten segmentar nuestros datos.

Indicador: También denominado métrica. Un indicador es una variable de los datos de tipo numérico que puede ser medida y agregada.

Cuartil: Los cuartiles de un conjunto ordenado de datos son los tres puntos de corte que dividen el conjunto de datos en cuatro grupos de la misma medida.

(definición de Wikipedia: <https://ca.wikipedia.org/wiki/Quartil>).

Variable: Todo atributo de nuestros datos, ya sea numérico o de texto. Por ejemplo, en una hoja de cálculo, las variables de nuestro conjunto de datos son las columnas.

Encontraréis más términos relacionados con esta materia en la *Terminología de la visualización de datos* del Termcat: <http://gen.cat/termvisdades>

7.2. Herramientas

La disciplina de la visualización de datos está en pleno auge. Por eso constantemente aparecen herramientas que facilitan el proceso de creación de visualizaciones. A continuación se presentan las herramientas más interesantes actualmente.

Tableau (<https://www.tableau.com/>): Tableau se ha convertido en la herramienta referente en el mundo de la visualización de datos, especialmente cuando se utiliza en los campos de la inteligencia de negocio o la analítica visual. Tableau posibilita cambiar el tipo de visualización con un solo clic, y facilita la creación de gráficos interactivos. Además, la herramienta facilita la conexión a múltiples fuentes de datos que van desde las hojas de cálculo más sencillas, a bases de datos, o a plataformas de datos masivos.

Tableau se puede utilizar en las fases de prototipaje y de finalización de las visualizaciones, pero también puede ser una buena herramienta para explorar visualmente los datos en la fase de estrategia.

Aunque Tableau es fácil de utilizar para crear gráficos sencillos, la curva de aprendizaje es bastante grande cuando se quieren hacer visualizaciones adelantadas. Por último, es interesante saber que, aunque es una herramienta de pago, también tiene una versión gratuita (Tableau Public <https://public.tableau.com/s/>).

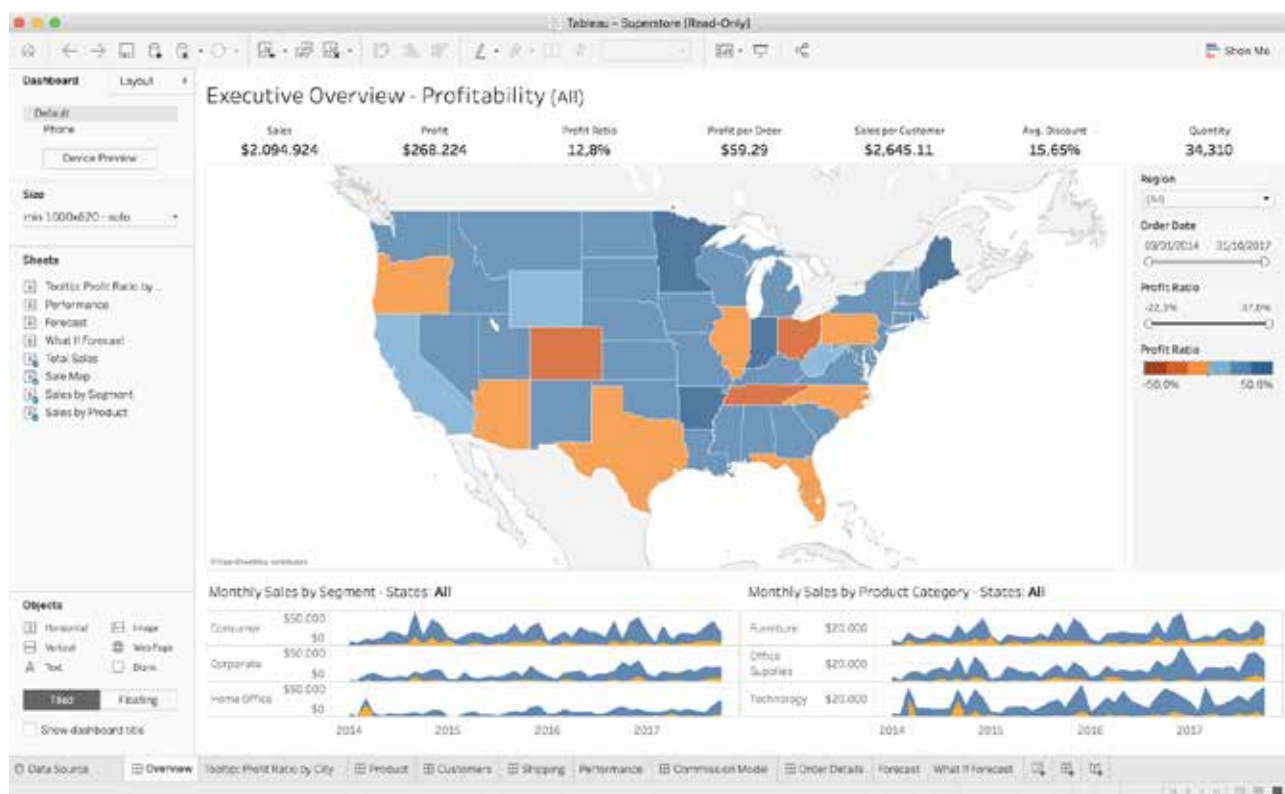


Figura 7.5. Captura de pantalla de uno de los *dashboards* que proporciona Tableau en su versión *Desktop*.

DataWrapper (<https://www.datawrapper.de/>): esta herramienta ha ganado últimamente mucha popularidad. A pesar de ser mucho más sencilla que Tableau, permite subir los datos a una hoja de cálculo y escoger entre una gran variedad de gráficos. Cada uno dispone de una serie de opciones para personalizarlos y adecuarlos a nuestras necesidades.

Al tratarse de una herramienta web, DataWrapper permite incrustar de un modo muy sencillo las visualizaciones que se crean en una página web o blog. Sin embargo, debe tenerse en cuenta que DataWrapper es una herramienta de pago.

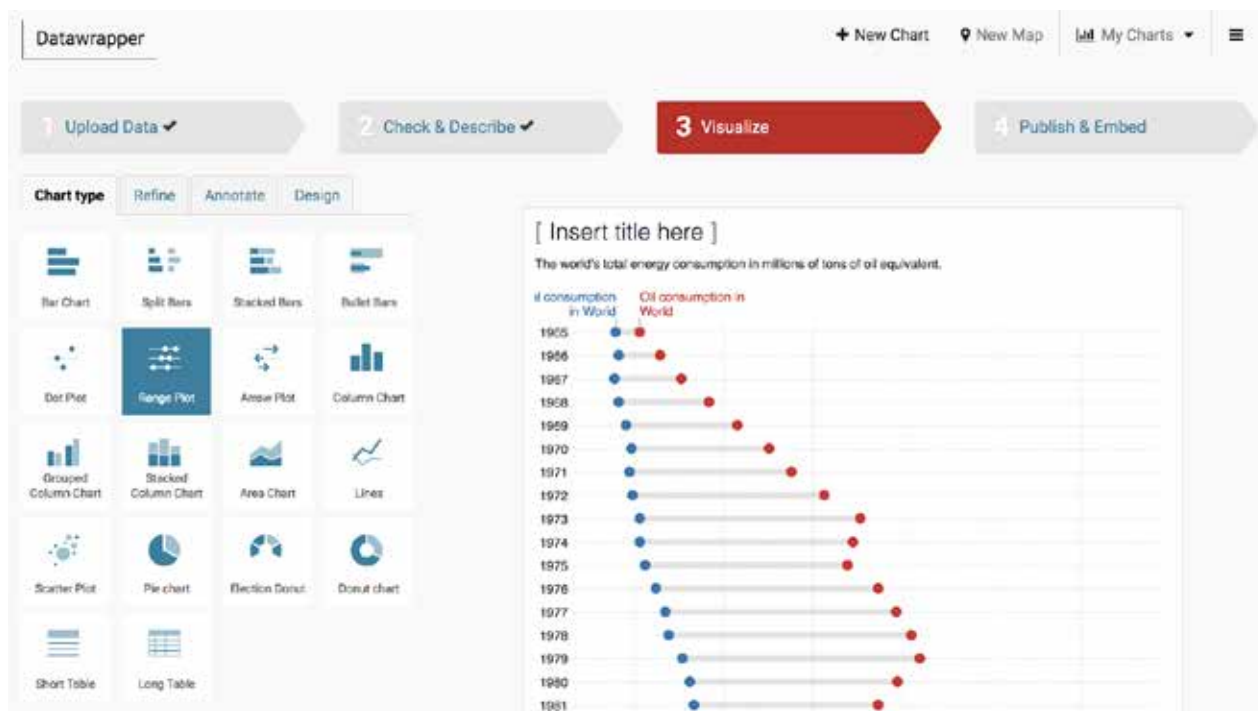


Figura 7.6. Captura de pantalla de DataWrapper.

Adobe Illustrator (<https://www.adobe.com/es/products/illustrator.html>): Esta popular herramienta de ilustración se utiliza mucho para esbozar, prototipar y finalizar visualizaciones. A menudo se utiliza para añadir los últimos detalles a las visualizaciones hechas con otras herramientas. No obstante, la curva de aprendizaje es muy grande. Adobe Illustrator es una herramienta de pago.

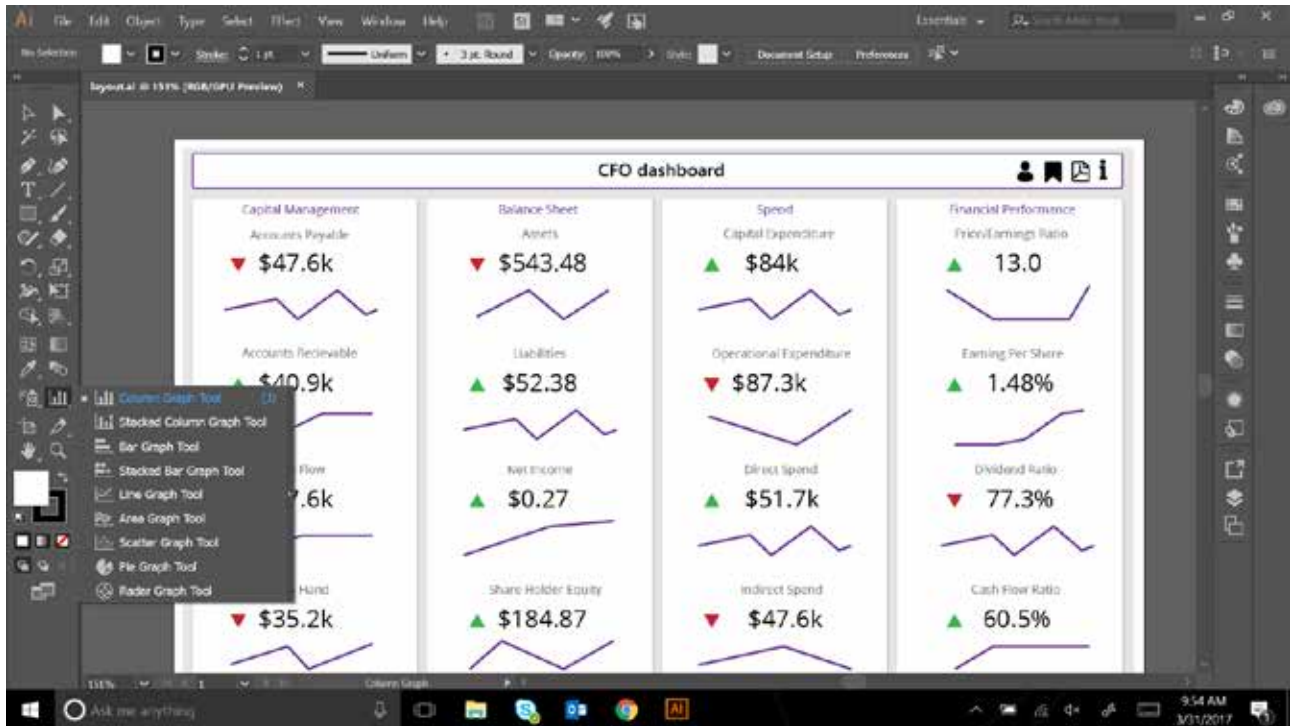


Figura 7.7. Captura de pantalla de un cuadro de mando hecho con Adobe Illustrator.
Fuente: https://cdn-images-1.medium.com/max/1600/1*h7ol80RsF6ajio5yboEAJg.png.

SketchApp (<https://www.sketchapp.com/>): SketchApp se ha convertido en una alternativa al Adobe Illustrator que utilizan los diseñadores de interfaces de usuario y los visualizadores de datos. Es mucho más sencilla de utilizar que el Adobe Illustrator y resulta especialmente adecuada para la fase de prototipaje de nuestra visualización. SketchApp es una herramienta de pago.



Figura 7.8. Captura de pantalla de un cuadro de mando prototipado con SketchApp.

Fuente: <https://www.sketchappsources.com/resources/source-image/collection-of-charts-bykova.png>.

Microsoft Excel (<https://products.office.com/en/excel>): Esta popular hoja de cálculo todavía es una herramienta referente en el mundo del análisis de datos. Históricamente ha facilitado la creación de todo tipo de gráficos, aunque poco a poco se han ido descubriendo carencias en el campo de la visualización. En las últimas versiones se ha ido adaptando y cada vez proporciona más variedad de visualizaciones modernas.

Microsoft Excel es una herramienta adecuada en la fase de prototipaje y en la de finalización. Es de pago.

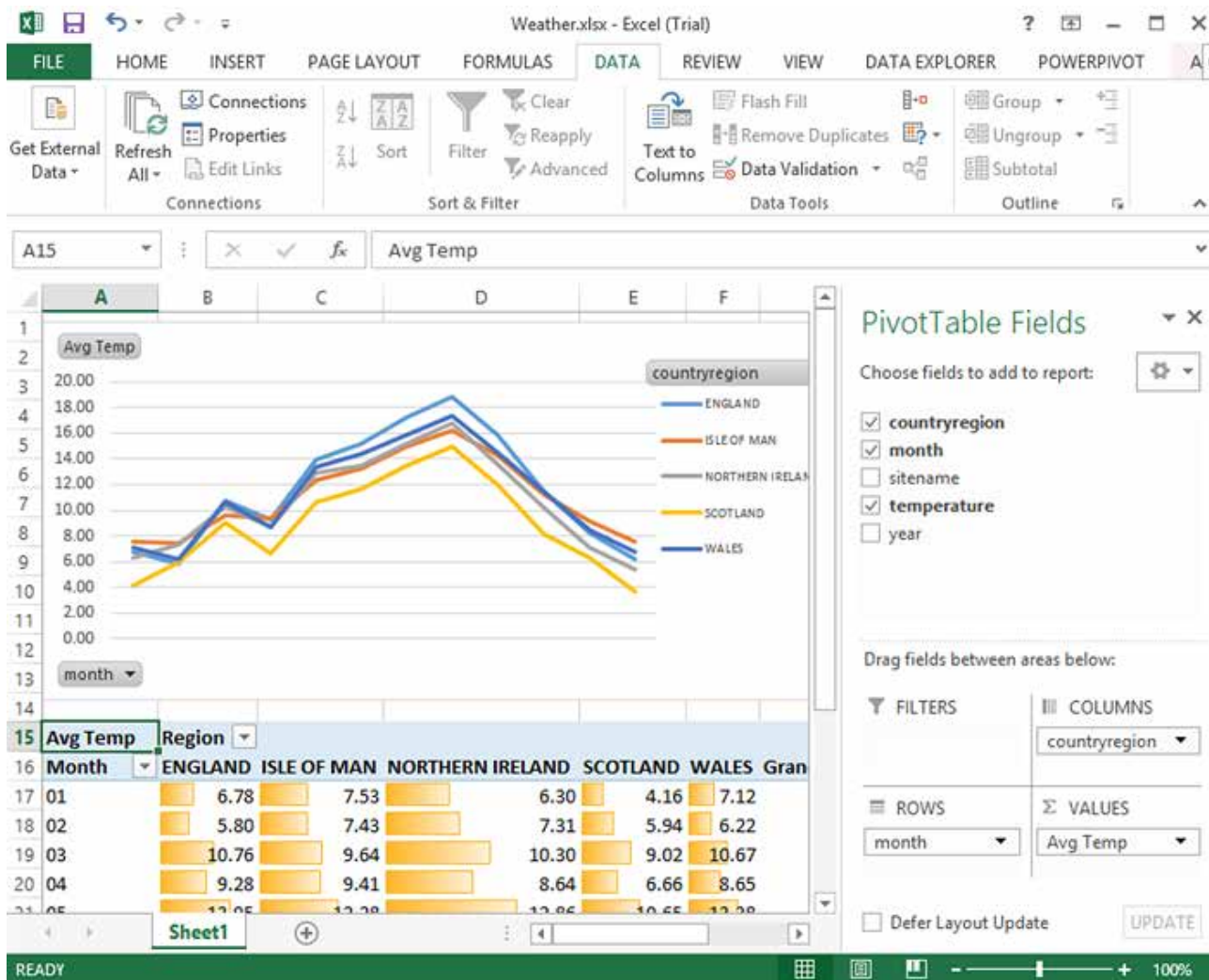
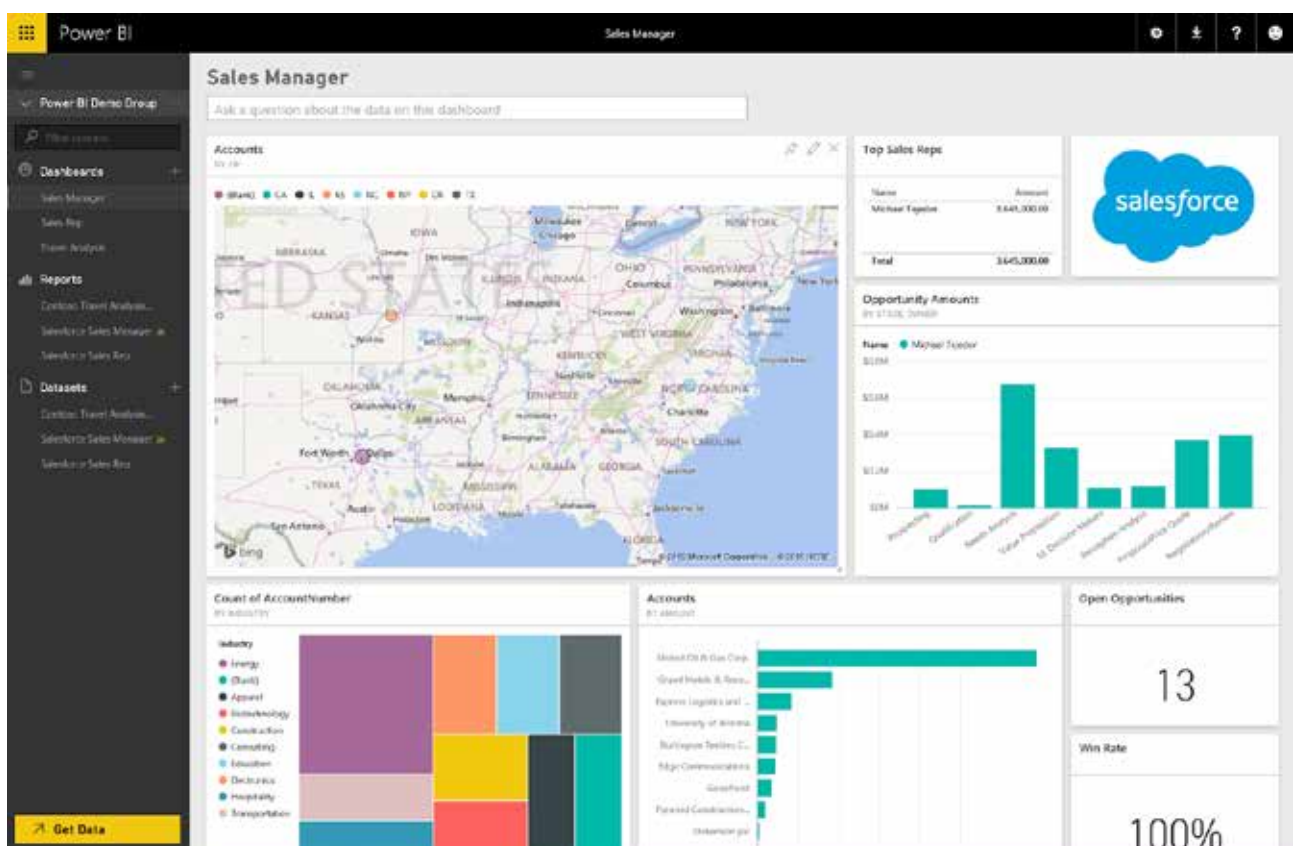


Figura 7.9. Captura de pantalla de un cuadro de mando desarrollado con Microsoft Excel.

Fuente: <https://msdn.microsoft.com/en-us/library/dn749860.aspx>.

Microsoft PowerBI se puede utilizar en las etapas de prototipaje y de finalización. La herramienta dispone de planes de utilización, y uno de ellos es gratuito.



Fuente: <https://blogs.microsoft.com/blog/2015/09/18/microsoft-expands-computer-science-education-partnership-with-sales-force-and-your-mind-with-hacking-mars-design-challenge-weekend-reading-sept-18-edition/>.

Instamaps (<https://www.instamaps.cat>): Es una herramienta gratuita para hacer mapas dirigida a usuarios no expertos; los mapas se pueden publicar en internet de forma rápida y sencilla, en abierto o privadamente, dentro de blogs o páginas web.

Puede posicionar elementos puntuales como una plaza, una fuente o un comercio, itinerarios como rutas o carreras o bien señalar polígonos como una zona industrial o una área de estudio.

Instamaps puede emplear información contenida en archivos geográficos (shp, gpx, kml, etc.) u otros archivos (xlsx, csv, tablas de Google, etc.), así como archivos y servicios procedentes de portales de datos abiertos, geoservicios wms u otra información en línea.

Instamaps es una plataforma abierta a todo el mundo creada por el Instituto Cartográfico y Geológico de Cataluña (<https://www.icgc.cat>).

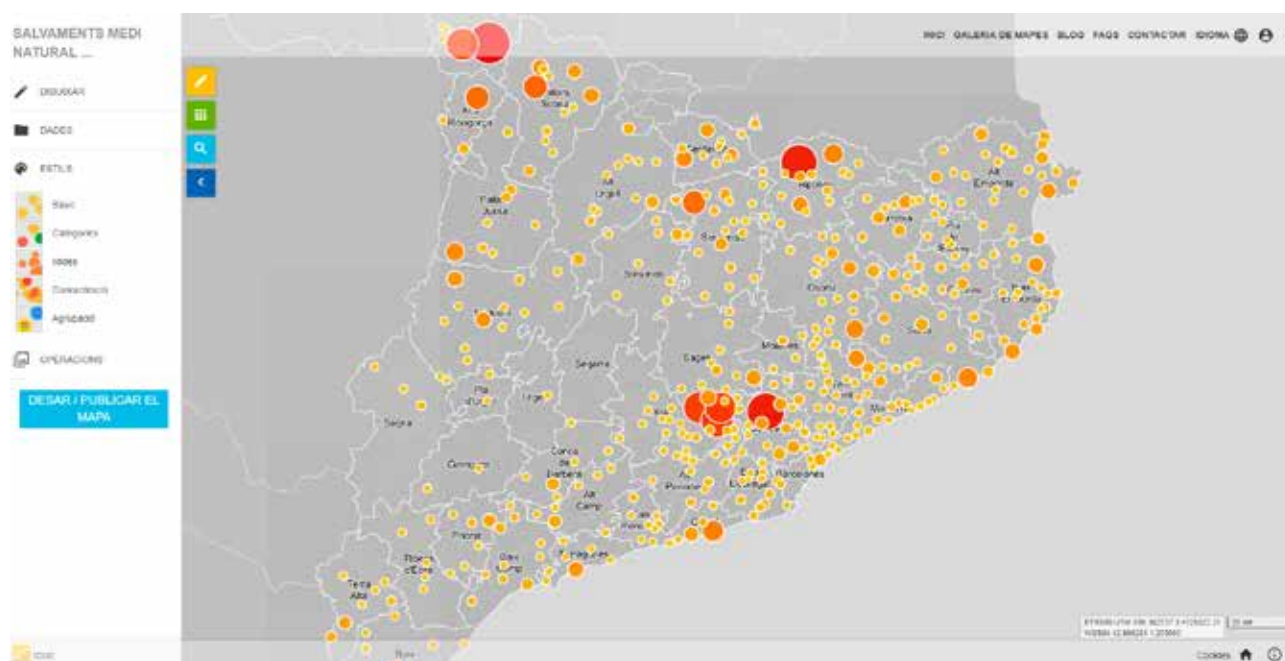


Figura 7.11. Captura de pantalla del mapa de Salvamentos del Medio Natural

Fuente: https://www.instamaps.cat/instavisor/498de3e1152b0d1d5ab20cc6ba5a8956/SALVAMENTS_MEDI_NATURAL_2017.html#8/41.882/1.208

D3 (<https://d3js.org/>): a pesar de no ser una herramienta, conviene mencionar D3, una librería hecha con el lenguaje de programación Javascript. Esta librería es una de las más utilizadas por los profesionales del sector. Como no es una herramienta, sino una librería que facilita la programación de visualizaciones para la web, no ofrece límites a la hora de crear visualizaciones e interacciones que vayan más allá de las soluciones más comunes y estándares.



Figura 7.11. Captura de un collage hecho con D3 que muestra visualizaciones desarrolladas con esta librería.
Fuente: <https://d3js.org/>.

R (<https://www.r-project.org/>) : A caballo entre las librerías de programación y el software más avanzado, R permite hacer análisis de datos muy complejos y generar visualizaciones gracias a paquetes como ggplot (<https://www.statmethods.net/advgraphs/ggplot2.html>) o Shiny (<https://shiny.rstudio.com/>). Concretamente, con Shiny se pueden crear visualizaciones interactivas con una sintaxis parecida al HTML (lenguaje de programación para crear páginas web). Resulta bastante compleja y es utilizada básicamente por perfiles técnicos.

