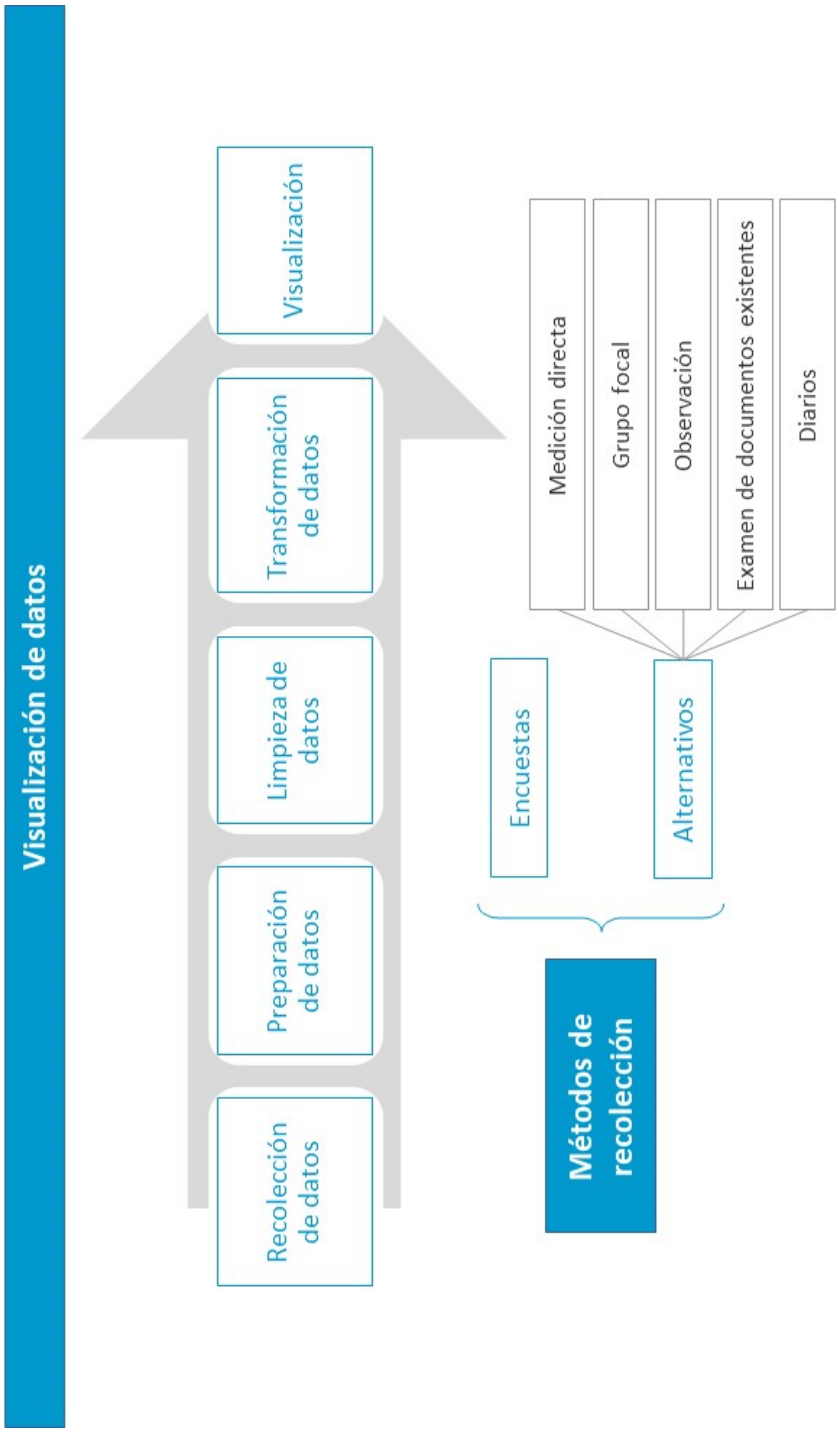
Visualización Interactiva de la Información

Tema 3. Trabajar con datos

Índice

Esquema



3.1. ¿Cómo estudiar este tema?

Para estudiar este tema deberás leer las **Ideas clave** desarrolladas en este documento, que se complementan con lecturas y otros recursos para que puedas ampliar los conocimientos sobre el mismo.

En este tema se estudian conceptos relacionados con el trabajo con datos y se explican las distintas etapas del proceso a seguir:

- Se presentan **los diferentes tipos de variables,** como fundamento del trabajo con datos.
- Se abordan cuestiones relativas a la **recolección de datos** y sus métodos.
- Se trata la **preparación y limpieza de los datos**, para homogeneizar y eliminar errores de conjuntos de datos.
- Se abordan las **transformaciones de datos**.
- Finalmente, se trata la **visualización de datos**, como etapa final.

3.2. Fundamentos de los datos

En la actualidad es habitual escuchar la expresión «la información es poder». Pero, en realidad, la idea que subyace no es válida sino que deberíamos aceptar mejor la expresión «**el conocimiento es poder, la información no**», ofrecida por el filósofo R. Lewis.

En efecto, la cantidad de información a nuestro alcance, desde la popularización de Internet, es realmente abrumadora. La capacidad de almacenarla, incluso a nivel personal, es cada vez mayor con la ayuda de sistemas de almacenamiento en la nube como Dropbox, Google Drive, One Drive, etc.

Resulta sumamente complicado absorberla y extraer de ella el conocimiento que necesitamos. Así, queda cada vez más patente que lo complicado ya no es tanto capturar y almacenar datos, como su análisis para poder representar posteriormente de forma adecuada y comprensible dicha información tanto de forma estática como dinámica.

Para comunicar adecuadamente de forma gráfica nuestra visión sobre un conjunto de datos, estos se han de mostrar atendiendo a **criterios estratégicos**. Es importante visualizar esos datos de manera que **se puedan comprender adecuadamente** y, gracias a ello, tomar decisiones informadas y correctas.

Un buen ejemplo de lo que NO se debe hacer lo vemos en la siguiente figura, extraída a partir de un buen volumen de datos de ventas de productos de una hoja Excel. ¿Realmente es fácil determinar que el producto D se vende más que el producto C? ¿Se ve claramente que el producto A está en tercera posición?

Figura 1: Gráfico de tarta en 3D.

Cuando se trata de convertir datos en información, los datos de los que se parte tienen una gran importancia. Los datos pueden ser simples **factoides** (de los que alguien más ya ha hecho todo el análisis) o **transacciones en bruto**, donde la exploración se deja enteramente al usuario.

De cara a su correcta visualización, un paso imprescindible es ser capaz de analizar e identificar los tipos de variables que conforman un conjunto de datos, ya que de la tipología de estas variables dependerá en gran medida qué expresión gráfica de los datos será más eficaz.

Entre los **tipos de variables** podemos diferenciar las siguientes:

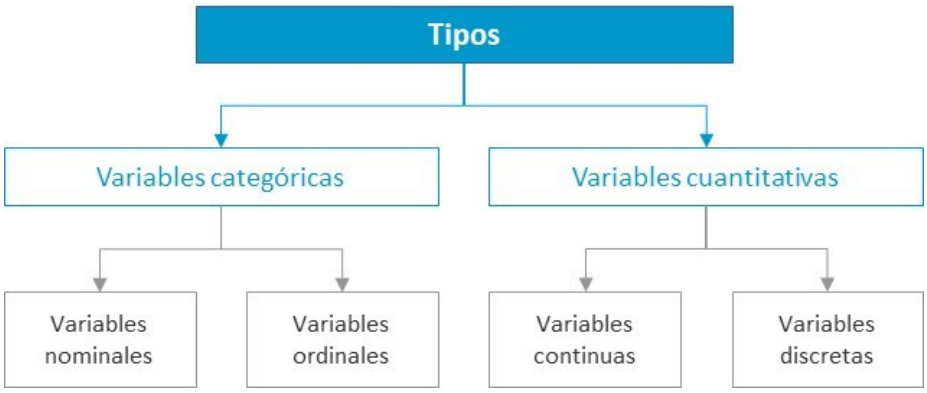


Figura 2: Tipos de variables.

Ideas clave

L a s **variables categóricas** son aquellas cuyos posibles valores representan categorías, mientras que en las **variables cuantitativas** los posibles valores representan numéricamente cantidades o mediciones

En las variables categóricas diferenciamos entre **variables nominales**, cuando las diferentes posibles categorías no presentan un orden predeterminado entre sí, como por ejemplo la variable «género literario» (ficción, terror, romance…); y **variables ordinales**, cuando sí existe ese orden predeterminado, como por ejemplo en la variable «titulación académica» (grado, máster, doctorado…).

En las variables cuantitativas diferenciamos entre **continuas o discretas** cuando los valores pueden presentar decimales (ejemplo de variable continua: peso), o cuando los posibles valores son valores absolutos (ejemplo de variable discreta: número de hijos).

3.3. Recolección de datos

La **recolección de datos** es esencial para la posterior visualización adecuada del conocimiento que se desprende de los mismos, por los sesgos que se pueden producir.

Para esta recolección se pueden usar técnicas no triviales, como el correcto **diseño de encuestas** que han de cumplimentar personas, hasta las utilizadas en grandes sistemas que coleccionan datos de cientos de miles de dispositivos **IoT** (Internet de las Cosas) en repositorios de información en bruto, para su posterior análisis con herramientas de ***Big Data*.**

Centrándonos en la creación de cuestionarios, no se trata simplemente de poner unas cuantas preguntas y ver qué contestan los usuarios. Siempre el punto de partida ha de ser identificar el propósito del cuestionario antes de crearlo.

Una buena encuesta recolecta datos fiables y verificables que nos permitirán realizar afirmaciones concretas.

En general, una encuesta es una herramienta útil para hacer alguna (o varias) de las siguientes cosas:

- Describir algo.
- Describir cómo están relacionadas las cosas.
- Explicar una relación.
- Influenciar sobre algo.

Para poder llegar a un propósito claro y bien definido de la encuesta, suele ser útil hacerse las siguientes preguntas:

Ideas clave

- ¿Qué deseamos lograr con la encuesta?
- ¿Qué es, exactamente, lo que queremos saber?
- ¿Por qué es importante saberlo?
- ¿Hay alguna otra información que podría ser útil? ¿Por qué?
- ¿Es la creación de un cuestionario el método adecuado para el tipo de datos que se están recolectando?

Lamentablemente, **en muchas ocasiones se crean encuestas por motivos erróneos**. Se desaprovechará la ocasión de obtener conocimiento útil con los mismos si:

- Realmente no nos importan los resultados, sino que solo queremos mostrar que tenemos datos y números.
- Utilizamos mal los datos recolectados.
- Si nos preocupa más cómo se recibirán los hallazgos que tener resultados fiables.
- Si ya se ha determinado de antemano cómo «deberían» ser los resultados.

En cualquier caso, no todos los tipos de información se recolectan fácilmente usando encuestas. En particular, las encuestas capturan información que tiene lugar en un momento exacto en el tiempo y las cosas que pueden suceder variando en el tiempo pueden ser difíciles de capturar.

Por ello, hay otras herramientas de recolección de datos que se pueden usar como alternativa mejor:

Método	Adecuado si
Medición directa	Los valores deben ser exactos.
Grupo focal	No se sabe aún qué preguntar. Si hay interés en dinámica de grupos o en procesos de decisión.
Observación	Lo que se está midiendo es fácil y públicamente observable y se desea tomar notas del comportamiento de los participantes.
Examen de documentos existentes	Los datos que nos interesan ya están registrados en algún sitio (recibos, facturas, logs de tráfico web, etc.).
Diarios	Se necesita hacer un seguimiento de variables en el tiempo.

Tabla 1: Métodos complementarios de recolección de datos. Adaptado de «Data+Design». Fuente: https://infoactive.co/data-design/

Por último, la realización de un estudio para recolectar datos no es la única forma de iniciar un análisis de datos, pues se pueden utilizar conjuntos de datos de terceros. Se trata de las **fuentes de datos externas**.

Estas fuentes de datos externas pueden ser:

- **Públicas**: hoy día existen muchísimas fuentes de datos externas accesibles gracias al movimiento de contenidos abiertos, donde se comparten (y compartimos) datos para su análisis libre por terceros. Muchas instituciones, gobiernos y organizaciones han establecido políticas de liberación de datos al público en pro de la transparencia.
- **Privadas**: puede suceder que en algún momento tengamos acceso a un conjunto de datos especial debido a nuestro estatus dentro de una red en particular, o bien que se trate de datos a la venta por terceros, para los que se ha de firmar una licencia de acceso y uso especial.

3.4. Preparación y limpieza de datos

Es importante llevar a cabo tareas preliminares de preparación de los datos, especialmente cuando se trata de datos recolectados por terceros. Aunque buena parte de estas tareas se pueden agilizar con herramientas informáticas, casi siempre hay una parte de labor manual que suele hacer de esta fase una etapa ingrata para muchos.

A modo de ejemplo, ¿cuántas veces hemos tenido que dar formato a valores provenientes de otros sistemas, sustituir puntos decimales por comas o eliminar comillas dobles de ciertos datos para trabajar con ellos en nuestro sistema? Esto es algo que casi siempre hay que hacer. Da igual que manejemos millones de datos o una simple y poco voluminosa colección de datos de ingresos y gastos mensuales.

La **preparación de datos** puede tomar muchas formas (complementarias):

- Separar los campos que nos serán útiles (por ejemplo, en las direcciones postales separar calle de ciudad y provincia).
- Identificar datos incompletos y vacíos.
- Uniformizar unidades y realizar conversión de unidades de medida (por ejemplo, si hay cosas expresadas en centímetros o en pulgadas, en metros y kilómetros a la vez).

Tras la preparación de datos, se debe realizar la **limpieza**:

- Identificar datos erróneos o carentes de sentido (por ejemplo, un «1234» en un campo utilizado para nombres de personas).
- Eliminar duplicados.
- Verificar rangos (por ejemplo, un dato de temperatura ambiente difícilmente puede

ser de 224 ºC en la Tierra).

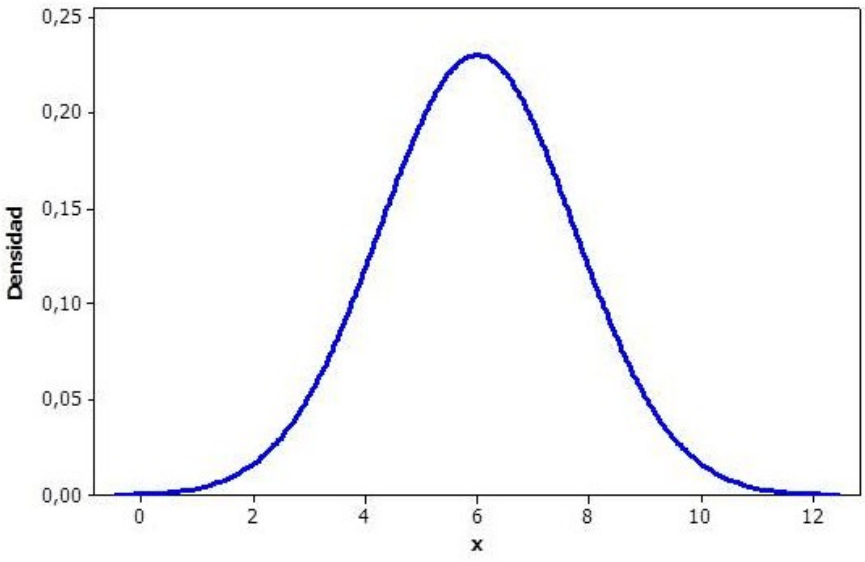
- Verificar ortografía y errores tipográficos.
- Examinar la sintaxis utilizando patrones y expresiones regulares (por ejemplo, para verificar el formato de direcciones de correo electrónico).

3.5. Transformación de datos

La **transformación de datos** es uno de los procedimientos de manipulación más comunes que pueden revelar características ocultas de los datos, que no son observables en su forma original.

Podemos transformar la distribución de los datos para hacerlos **más fáciles de observar** y para que **se cumplan los requisitos de muchas pruebas estadísticas**. Esto normalmente se hace sustituyendo una variable por una función matemática que opera sobre esa variable.

Una de las suposiciones más frecuentes de las pruebas estadísticas es que los datos tengan una **distribución normal** (en forma de campana con datos dispersos en torno a un valor central). Los **valores sesgados**, al contrario de la distribución normal, tienden a tener más observaciones a la izquierda o a la derecha.



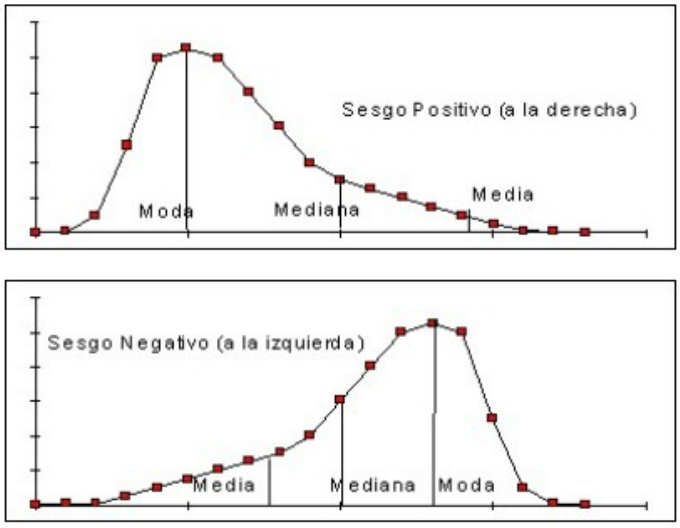


Figura 3: Distribución normal y distribuciones con sesgo positivo y negativo.

Existen muchas transformaciones que se pueden aplicar cuando hay datos sesgados, siendo las más importantes:

- La transformación logarítmica.
- La transformación mediante raíz cuadrada.

Otra forma común de transforman los datos, en el caso de distribuciones cuantitativas, es mediante su rexpresión a través de medidas de tendencia central, como la media o la mediana, lo que nos permite resumir una serie de datos a través de un único valor. La ventaja de usar estas medidas de tendencia central es que simplifican la realidad expresada por el conjunto de los valores, pero su desventaja es que este único valor representativo del conjunto de los valores poco nos dice acerca de la distribución de los valores resumidos.

Un ejemplo lo podemos encontrar en el famoso **cuarteto Anscombe** (recibe el nombre del estadístico que lo ideó), formado por cuatro conjuntos de datos que presentan las mismas propiedades estadísticas:

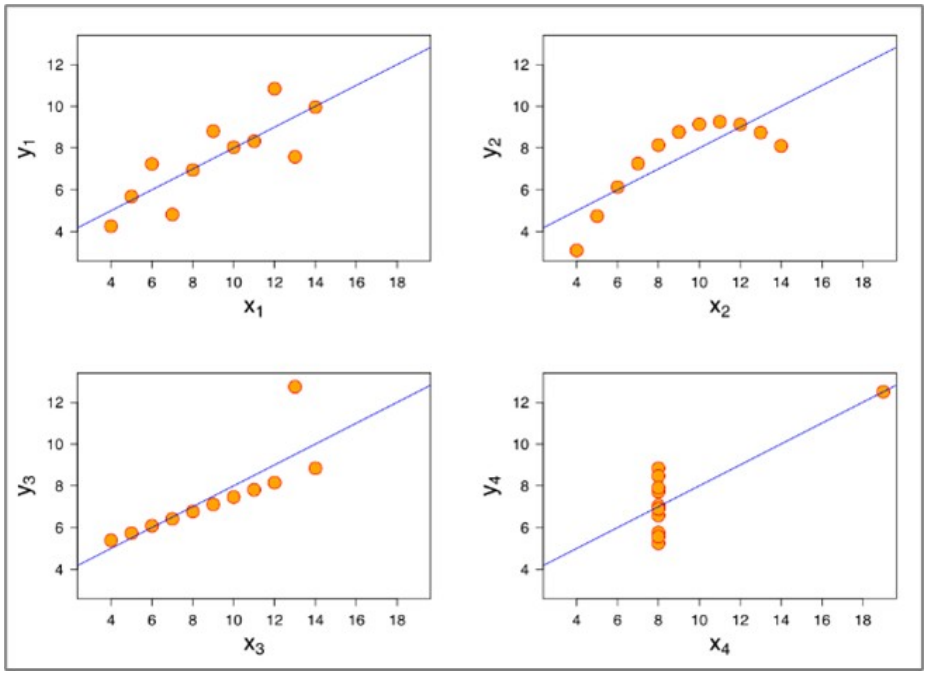


Figura 4: Cuarteto de Anscombe. Fuente: Wikipedia.

En los cuatro casos, la media de la variable Y, la media de la variable X, o el valor de la correlación entre ambas variables, son idénticas. Se trata de un ejemplo que evidencia la importancia de graficar un conjunto de datos para comprenderlos realmente, y que visibiliza la información que perdemos al resumir distribuciones de valores a través de un único valor representativo.

3.6. Visualización de datos

Para realizar una correcta visualización de los datos, en primer lugar, se ha de determinar claramente el **mensaje a comunicar**. Suele ser útil preguntarse para ello, ¿qué es lo que sé?, ¿qué significa?, ¿por qué es importante? Si no se puede resumir el mensaje clave en una frase concisa, probablemente necesitemos darle una vuelta más y repensar la cuestión.

Una vez determinado el mensaje, se debe considerar la importancia de **entender a la audiencia destino**, sobre todo, para determinar cuáles y cuántos de los datos se deben ilustrar. A la hora de decidir cuántos datos se presentan, es esencial entender que más no siempre es sinónimo de mejor, salvo que tenga sentido a la hora de apoyar el mensaje.

En numerosas ocasiones se utiliza la narrativa, entendida como una historia en la que se presentan una serie de **hechos que conducen gradualmente a la audiencia hacia el mensaje clave**. Para probar si la narrativa es efectiva, se puede valorar si al suprimir o reordenar alguna de las ilustraciones el mensaje sigue siendo claro o se desvirtúa de alguna forma.

Finalmente, hay que indicar que existen numerosas técnicas y consejos a la hora de representar visualmente la información, desde el tipo de gráficas (de barras, apiladas, de tarta, etc.) hasta el uso que se ha de hacer de la tipografía y colores, pasando por el etiquetado de los elementos representados. Todo esto es una mezcla de **ciencia y arte** por sí solo y existen numerosos libros y artículos publicados al respecto.

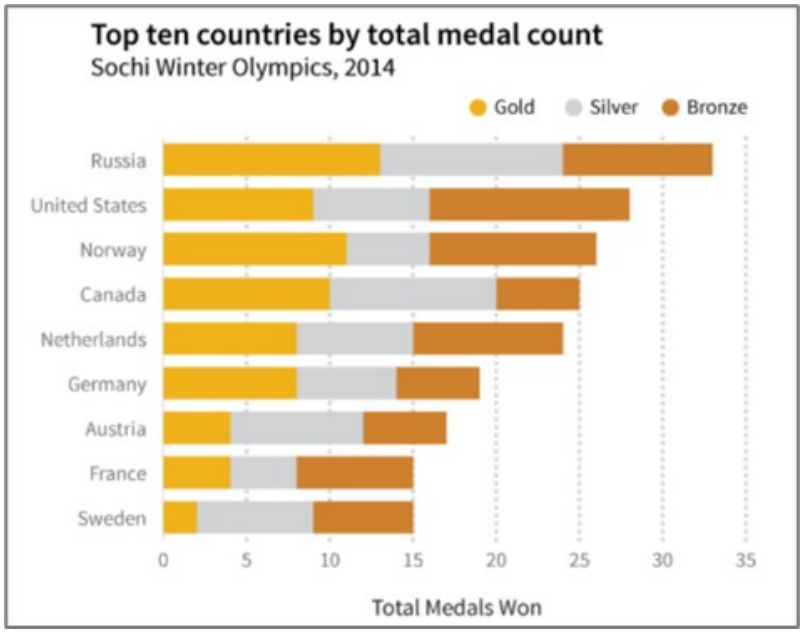


Figura 5: Ejemplo de visualización muy correcta del medallero resultante de los Juegos Olímpicos de Invierno de Sochi en 2014. Fuente: https://infoactive.co/data-design

Visualización de datos: el diseño de la comprensión

Interesante artículo de Ignasi Alcalde en el que se reflexiona sobre el diseño de la comprensión como parte fundamental de la visualización de datos.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web: http://www.ignasialcalde.es/visualizacion-de-datos-el-diseno-de-la-comprension/

Diseño de la información y visualización de datos

Bien sintetizado ensayo de Mauro Villena sobre las necesidades metodológicas relativas al diseño de la información y la visualización de datos para el diseño gráfico actual, donde se expresa claramente la diferencia entre los conceptos de diseño de la información y la visualización de datos.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web: https://wiki.ead.pucv.cl/Dise%C3%B1o_de_la_informaci%C3%B3n_y_Visualizaci%C3%B3n_de_datos,_necesidades_metodol%C3%B3gicas_para_el_Dise%C3%B1o_Gr%C3%A1fico_actual.

¿Qué papel tiene la visualización de datos en el diseño de la interacción?

Interesante vídeo del profesor Andy Cargile sobre el papel que juega la visualización de datos en el diseño de la interacción.

Accede al vídeo desde el aula virtual o a través de la siguiente dirección web: https://vimeo.com/111141729

Open Refine

Open Refine es una herramienta *open source* para la limpieza y preparación de datos. En su sección de documentación encontramos la explicación detallada de sus funcionalidades.

Accede al recurso desde el aula virtual o a través de la siguiente dirección web: https://openrefine.org/docs

Data plus Design: A simple introduction to preparing and visualizing information

Completo y valioso documento elaborado colaborativamente por Trina Chiasson, Dyanna Gregory y más de 50 personas, sobre el proceso de diseño de la visualización, en el que se hace hincapié en la transformación de datos (explicada para no expertos en estadística) y en la propia representación de los mismos.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web:
https://infoactive.co/data-design

Bibliografía

Kirk, A. (2012). *Data Visualization: A sucessfull Design Process*. Birmingham (UK): Packt Publishing.

Murray, S. (2013). *Interactive Data visualization for the Web*. Massachusetts (EUA): O'Reilly.

Telea, A. C. (2014). *Data Visualization: Principles and practice, second Edition*. Ohio (EUA): CRC Press.

Test

1. La información:

A. Es el verdadero poder en sí misma y en términos absolutos.

B. Permite extraer de ella, mediante distintas técnicas, conocimiento, que es realmente lo valioso.

C. Solo tiene sentido si se puede representar en 3D.

D. Todas las anteriores son correctas.

2. Para comunicar adecuadamente de forma gráfica nuestra visión sobre un conjunto de datos, estos se han de mostrar siguiendo:

A. Criterios estratégicos.

B. Criterios económicos.

C. Criterios psicosociales.

D. Todas las anteriores son correctas.

3. Entre las variables categóricas distinguimos entre

A. Variables nominales y discretas.

B. Variables ordinales y nominales.

C. Variables cuantitativas y nominales.

D. Variables discretas y continuas.

4. Las medidas de tendencia central:

A. Poco nos dicen sobre la distribución de los valores resumidos.

B. Deben ser utilizadas siempre que podamos.

C. Deben evitar ser utilizadas siempre.

D. Ninguna de las anteriores es correcta.

5. Son motivos erróneos para la creación de una encuesta:

A. Que realmente no nos importen los resultados, sino solo mostrar que tenemos datos y números.

B. Utilizar mal los datos recolectados.

C. Que preocupe más cómo se recibirán los hallazgos que tener resultados fiables.

D. Todas las anteriores son correctas.

6. Son herramientas de recolección de datos

A. Medición directa, grupo focal y observación.

B. Medición directa, heurísticas isobáricas y Examen de análisis previos.

C. Grupo focal, examen de análisis previos y fabricación de factoides.

D. Todas las anteriores son correctas.

7. Indica cuál de las siguientes afirmaciones es correcta:

A. La limpieza de datos se ha de realizar antes de la preparación de los mismos.

B. La verificación de rangos de datos es parte de la preparación de datos.

C. La uniformización de unidades de medida es parte de la preparación de datos.

D. Todas las anteriores son correctas.

8. El cuarteto de Anscombe:

A. Son cuatro profesionales de la visualización muy conocidos.

B. Visibiliza la información que ganamos al resumir distribuciones.

C. Evidencia la importancia de graficar un conjunto de datos para comprenderlos.

D. Todas las anteriores son incorrectas.

9. Una distribución normal de los datos:

A. Es aquella que tiene forma de campana, con datos en torno a un valor central.

B. Es aquella que tiene sesgo, con un elevado número observaciones a la derecha o a la izquierda.

C. Es la que normalmente se tiene en los datos, con independencia de su naturaleza.

D. Todas las anteriores son correctas.

10. Para la visualización de datos:

A. Se ha de determinar claramente el mensaje a comunicar.

B. Se debe entender a la audiencia destino.

C. En ocasiones se utiliza la narrativa.

D. Todas las anteriores son correctas.