Universitat Oberta de Catalunya

uoc.edu

Aspectos a tener en cuenta para la Práctica de visualización de datos

Autor: Xavier Giménez
Coordinador: Julià Minguillón

Índice

Selección del conjunto de datos

Objetivo

Seleccionar un conjunto de datos adecuado para crear una visualización

- **Relevancia:** el conjunto de datos debería ser de actualidad, significativo (huir de datos sintéticos), de interés general, y permite plantear preguntas interesantes
- **Dimensiones:**
  - Del orden de 1000-10000 filas
  - Del orden de 10-100 columnas
- **Características:** combina datos numéricos y categóricos
- **Contiene alguna jerarquía:** categoría / subcategoría
- **Atención a temas sensibles:** datos personales, des-anonimización, etc.
- **¿Qué licencia tiene el conjunto de datos? ¿De dónde se ha obtenido?**

Preparación

El conjunto de datos debe ser inspeccionado para detectar posibles necesidades más adelante

- **Valores perdidos**
  - Eliminar filas / columnas con un exceso de datos perdidos
- **Homogeneizar datos categóricos (p.e. “Barcelona” / “BCN”)**
- **Detectar valores extremos en datos numéricos**
  - Box-plots
  - Histograma
- **Extracción de características**
- **Combinarlo con otros conjuntos de datos**

Fuentes

Infinitas opciones

- **UCI Machine Learning repository**
  - https://archive.ics.uci.edu/ml/datasets.php
- **Kaggle**
  - https://www.kaggle.com/datasets
- **Datos abiertos**
  - https://governobert.gencat.cat/ca/dades_obertes/
  - https://datos.gob.es/
  - https://data.europa.eu/es
- **Statista (disponible en el aula)**
- **...**

Ejemplos

Temas que pueden ser adecuados

- **Covid:** casos, fallecidos, vacunas, etc., por área geográfica, por periodo de tiempo, por grupo de edad, por género, ...
- **Calentamiento global:** mediciones históricas de temperatura, emisiones de gases de efecto invernadero, mediciones del deshielo en los polos, ...
- **Gobierno abierto:** datos de gasto público, gestión gubernamental de presupuestos, adjudicación de obra pública, …

Ejemplos

Temas que pueden ser adecuados

Cualquier tema puede ser adecuado, partiendo siempre de unos requisitos:

- **Veracidad de la fuente de datos**: comprobar qué entidad, organismo o institución ha creado los datos.
- **Rigor en el proceso de obtención de datos**: identificar posibles *data mining pitfalls*: errores en la captura de datos, poca cobertura de la muestra total, ambigüedad, etc.
- **Contexto:** Considerar añadir más fuentes de información si los datos de que disponemos no son autoexplicativos por sí solos (p.e. visualizar datos de crecimiento económico sin tener datos históricos de la inflación).

Propuestas

Datos que pueden generar visualizaciones interesantes

- **Callejero:** a partir de las calles de un municipio, clasificarlas (manualmente!) en función de diversos criterios, p.e. género (mujer / hombre / neutral), origen (religioso / histórico / geográfico / cultural / …), longitud, … Se pueden cruzar con datos de tráfico, población, etc.

  https://geochicasosm.github.io/lascallesdelasmujeres/

- **Wikipedia**: análisis de las últimas ediciones realizadas, por usuario, tipología, tamaño de la edición, etc.

  https://es.wikipedia.org/wiki/Especial:CambiosRecientes
  http://hint.fm/projects/historyflow/

Propuestas

Datos que pueden generar visualizaciones interesantes

- **Calentamiento global:** dar valor a los datos históricos producidos por las diferentes agencias nacionales de meteorología para visualizar cambios en la temperatura global del planeta.

  https://showyourstripes.info/s/globe

- **Migración**: Visualizar historias personales (uno mismo tiene que obtener y crear los datos!) sobre inmigrantes en busca de asilo fuera de su país de origen.

  http://www.storiesbehindaline.com/

Ejemplo práctico

Fuente: Eurostat, el portal de estadística de la Comisión Europea

- Publica indicadores estadísticos a nivel europeo, como resultado de la colaboración entre los distintos institutos estadísticos de los países miembros de la UE.

- Ofrece en formato *Open Data* conjuntos de datos estadísticos e indicadores sobre diversas áreas tales como Economía y financiación, Industria, Servicios, Agricultura, Ciencia, etc., e indicadores sobre bienestar social, desarrollo sostenible, economía circular, ...

Ejemplo práctico

Fuente: Eurostat, el portal de estadístico de la Comisión Europea

- Podemos asumir que la fuente de datos es fiable y que está libre de errores metodológicos durante su captura y procesamiento.

- Incluye una sección con ejemplos interactivos de visualizaciones de datos para explorar los datos:

  https://ec.europa.eu/eurostat/web/main/data/visualisation-tools

Ejemplo práctico

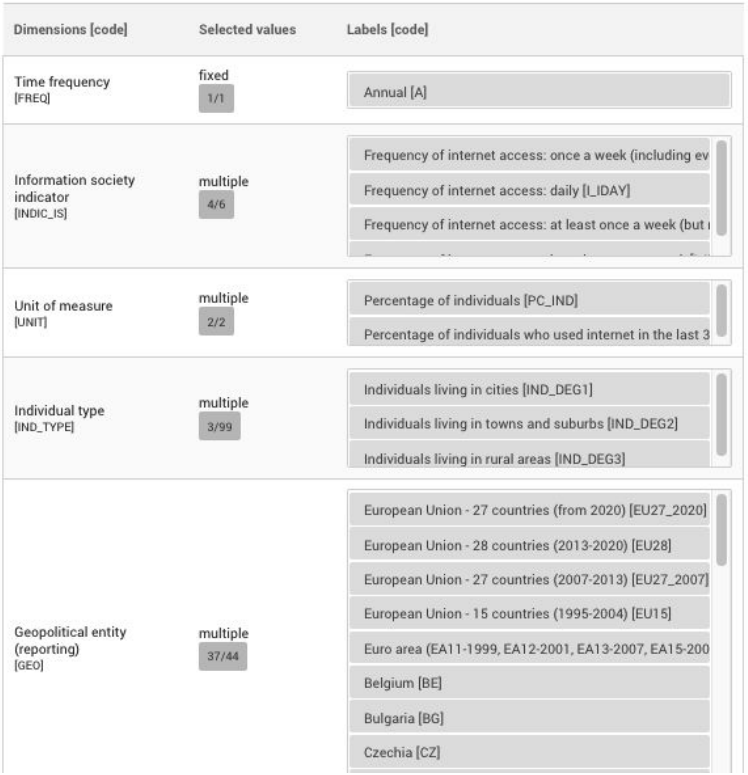
***Dataset*: Frecuencia de uso de internet en la población**

- Conjunto de datos compuesto por 22 columnas y 6479 filas, en formato largo para las variables categóricas y en formato ancho para las columnas con información temporal.
- Refleja situaciones reales (y habituales): el conjunto de datos no es 100% perfecto ni completo, como sucede con la mayoría de *datasets*.
- Múltiples valores categóricos:
  - **Frecuencia de acceso a internet** (6 valores): diario, 1 vez / semana, 1 vez / mes, etc.
  - **Tipo de población** (99 valores): individuos segregados por edad, educación, tipo de empleo, situación geográfica, etc.
  - **País o composición histórica de la UE** (44 valores)
  - **Años** (18 valores, 2003-2020)
  - ...

Ejemplo práctico: Frecuencia de uso de internet en la población

El dataset incluye ciertas **características relevantes** (y apropiadas para el uso de la visualización de datos), como por ejemplo:

- Multitud de posibles valores categóricos, con un aún mayor número de posibles combinaciones de datos.



Dimensions [code]	Selected values	Labels [code]
Time frequency [FREQ]	fixed 1/1	Annual [A]
Information society indicator [INDIC_IS]	multiple 4/6	Frequency of internet access: once a week (including ev Frequency of internet access: daily [I_IDAY] Frequency of internet access: at least once a week (but
Unit of measure [UNIT]	multiple 2/2	Percentage of individuals [PC_IND] Percentage of individuals who used internet in the last 3
Individual type [IND_TYPE]	multiple 3/99	Individuals living in cities [IND_DEG1] Individuals living in towns and suburbs [IND_DEG2] Individuals living in rural areas [IND_DEG3]
Geopolitical entity (reporting) [GEO]	multiple 37/44	European Union - 27 countries (from 2020) [EU27_2020] European Union - 28 countries (2013-2020) [EU28] European Union - 27 countries (2007-2013) [EU27_2007] European Union - 15 countries (1995-2004) [EU15] Euro area (EA11-1999, EA12-2001, EA13-2007, EA15-200 Belgium [BE] Bulgaria [BG] Czechia [CZ]

Ejemplo práctico: Frecuencia de uso de internet en la población

El dataset incluye ciertas **características relevantes** (y apropiadas para el uso de la visualización de datos), como por ejemplo:

- Multitud de posibles valores categóricos, con un aún mayor número de posibles combinaciones de datos.
- Series de datos incompletas, con diversa casuística: datos no disponibles, valores poco fidedignos, series temporales incompletas, etc.

TIME / GEO	2013	2014	2015	2016	2017	2018	2019
European Union - 27 countries (from 2020)	76	78	80	82	84	85	88
European Union - 28 countries (2013-2020)	77	80	81	83	85	86	89
European Union - 27 countries (2007-2013)	77	80	81	84	85	86	:
European Union - 25 countries (2004-2006)	:	:	:	:	:	:	:
European Union - 15 countries (1995-2004)	79	81	83	85	86	87	:
Euro area (EA11-1999, EA12-2001, EA13-2007, EA15-2...	77	78	81	82	84	86	88
Belgium	80	81	82	83	86	86	87
Bulgaria	65	68	69	70	73	74	75
Czechia	73	84	83	84	86	88 (b)	89
Denmark	94	95	94	96	97	96	97
Germany (until 1990 former territory of the FRG)	82	85	87	88	89	93	93
Estonia	83	85 (b)	88	88	90	89	91
Ireland	81	83	83	84	85	86	91
Greece	:	65	72	74	75	78	80
Spain	71	76	79	80	84	85	90
France	81	82	82	84	85	86	89
Croatia	74	74	79	78	74	82	86
Italy	61	64	67	70	73	75	77
Cyprus	67	70	74	76	84	87	88
Latvia	76	77	81	82 (b)	82	84	87
Lithuania	75	78	76	81	84	85	87
Luxembourg	94	94	96	97	96	94 (b)	94
Hungary	80	83	82	87	83	82	86
Malta	64	68	74	78	81	80	86
Netherlands	92	92	93	92 (b)	95	94	96

Special value:
(:) not available

Available flags:
(b) break in time series
(n) not significant
(bn) break in time series, not significant
(u) low reliability

Ejemplo práctico: Frecuencia de uso de internet en la población

El dataset incluye ciertas **características relevantes** (y apropiadas para el uso de la visualización de datos), como por ejemplo:

- Multitud de posibles valores categóricos, con un aún mayor número de posibles combinaciones de datos
- Series de datos incompletas, con diversa casuística: datos no disponibles, valores poco fidedignos, series temporales incompletas, etc.
- Estructura: el dataset no está en un formato adecuado (p.ej. *tidy-data*) para su uso en visualizaciones.

TIME / GEO	2013	2014	2015	2016	2017	2018	2019
European Union - 27 countries (from 2020)	76	78	80	82	84	85	88
European Union - 28 countries (2013-2020)	77	80	81	83	85	86	89
European Union - 27 countries (2007-2013)	77	80	81	84	85	86	:
European Union - 25 countries (2004-2006)	:	:	:	:	:	:	:
European Union - 15 countries (1995-2004)	79	81	83	85	86	87	:
Euro area (EA11-1999, EA12-2001, EA13-2007, EA15-2...	77	78	81	82	84	86	88
Belgium	80	81	82	83	86	86	87
Bulgaria	65	68	69	70	73	74	75
Czechia	73	84	83	84	86	88 (b)	89
Denmark	94	95	94	96	97	96	97
Germany (until 1990 former territory of the FRG)	82	85	87	88	89	93	93
Estonia	83	85 (b)	88	88	90	89	91
Ireland	81	83	83	84	85	86	91
Greece	:	65	72	74	75	78	80
Spain	71	76	79	80	84	85	90
France	81	82	82	84	85	86	89
Croatia	74	74	79	78	74	82	86
Italy	61	64	67	70	73	75	77
Cyprus	67	70	74	76	84	87	88
Latvia	76	77	81	82 (b)	82	84	87
Lithuania	75	78	76	81	84	85	87
Luxembourg	94	94	96	97	96	94 (b)	94
Hungary	80	83	82	87	83	82	86
Malta	64	68	74	78	81	80	86
Netherlands	92	92	93	92 (b)	95	94	96

Special value:
(:) not available

Available flags:
(b) break in time series
(n) not significant
(bn) break in time series, not significant
(u) low reliability

Ejemplo práctico: Frecuencia de uso de internet en la población

Este conjunto de datos **es adecuado para la visualización de datos** ya que:

- Proviene de una organización fiable y rigurosa.
- Tiene un alta densidad de datos, tanto numéricos como categóricos.
- Permite realizar multitud de preguntas (p.e., uso de la tecnología según diversos ámbitos: regional, por grupos de población, etc.)
- Es fácilmente combinable con otros conjuntos de datos de interés (p.e., indicadores socio-económicos europeos).

El proceso de visualizar información

Planteamiento de preguntas

- **Usuarios:** Determinar a quién va dirigida la visualización.
  - ¿Mis usuarios son el público en general o un segmento concreto, p.e. expertos?
  - ¿Hay que proporcionar un contexto previo a la visualización de los datos?

- **Preguntas:** ¿Qué quiero responder con la visualización?
  - ¿El *dataset* me permite contestarlas con precisión?
  - ¿Qué datos específicos de mi *dataset* quiero explorar?
  - ¿Necesito más fuentes de datos para dar contexto a mi conjunto de datos?

Es importante revisar las decisiones tomadas en la selección del conjunto de datos y ver si es suficiente para obtener los objetivos deseados.

Planteamiento de preguntas

3. Evaluación:
¿Mi conjunto de datos es suficiente para responder a las preguntas? ¿Necesito añadir nuevos datos? ¿Surgen más preguntas debido a la complejidad de los datos?

1. Elección de los datos:
el conjunto de datos debe tener los requisitos para poder plantear preguntas interesantes

2. Preguntas: exponer preguntas con un objetivo claro y un ámbito definido, que se puedan cuantificar y que no dependan de suposiciones

Exploración de los datos

En función de la naturaleza del conjunto de datos (datos en forma tabular, numéricos y categóricos), será interesante realizar una fase de inspección o incluso un E.D.A. (*Explorative Data Analysis*) para:

- Mostrar qué distribuciones siguen los distintos valores presentes en los datos.
- Detectar patrones, tendencias y valores atípicos.
- Detectar asociaciones entre variables.

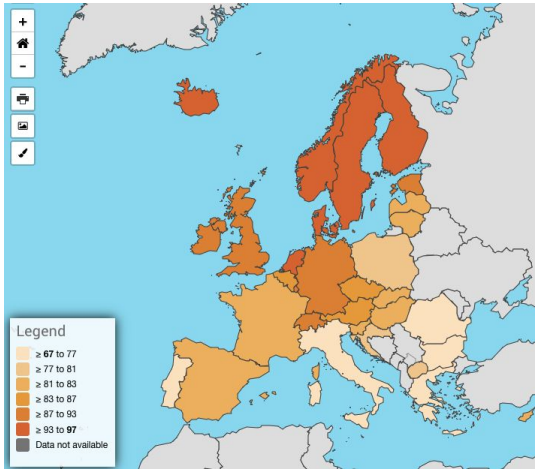
En otros casos (datos no estructurados, grafos, texto, mapas, etc.) el equivalente a este análisis pasa por plantearse otras preguntas, como por ejemplo:

- Detectar relaciones entre entidades en datos no tabulares.
- Mostrar patrones geográficos (flujos de movilidad, fenómenos poblacionales).
- Detectar temáticas en datos no estructurados (caracterizar documentos de texto).

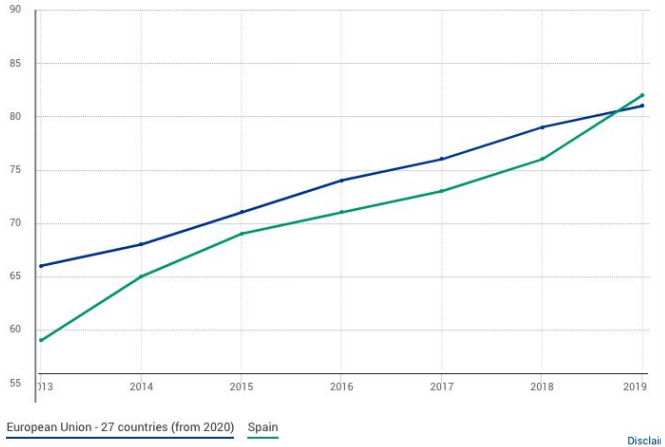
Exploración de los datos

Siguiendo con nuestro ejemplo práctico de conjunto de datos (Frecuencia de uso de internet en la población), la exploración consiste en visualizaciones que nos permitan entender la información contenida en los datos, por ejemplo (https://ec.europa.eu/eurostat/web/main/data/visualisation-tools):

¿Qué % de población en Europa hace uso diario de internet?



¿Cuál ha sido la evolución del uso de internet en España en comparación con la Unión Europea?

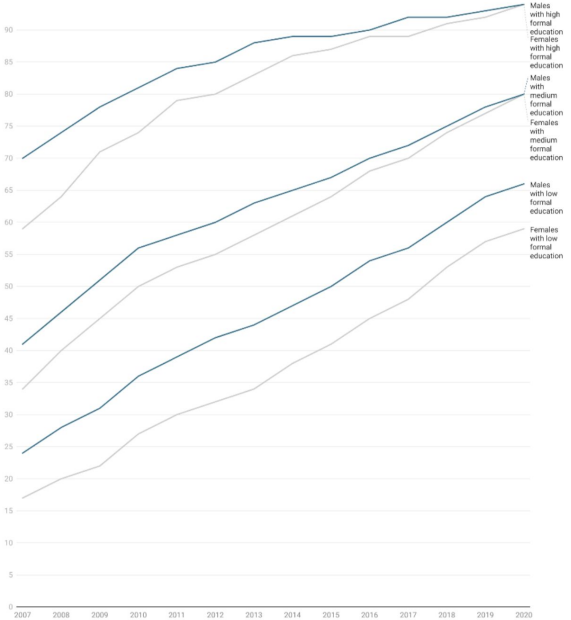


Exploración de los datos

Siguiendo con nuestro ejemplo práctico de conjunto de datos (Frecuencia de uso de internet en la población), la exploración consiste en visualizaciones que nos permitan entender la información contenida en los datos, por ejemplo:

¿Qué diferencias hay en el uso diario de internet entre distintos grupo poblacionales?

¿Cómo difieren entre sí grupos según el género y nivel de estudios?



Exploración de los datos

La fase de exploración de datos tiene que proporcionar:

- Una idea clara de la información que contiene el conjunto de datos.
- Detectar inconsistencias y/o errores en los datos.
- Validar la factibilidad de las cuestiones que queremos resolver.
- Facilitar nuevas preguntas y/o exponer la necesidad de añadir nuevas fuentes de datos.

Por ejemplo, el mapa del uso de Internet nos muestra claramente unas diferencias Norte-Sur que no se observan fácilmente en los datos cuando están en forma tabular, y que pueden hacernos pensar en asociaciones con otras variables que también muestren diferencias en el mismo eje Norte-Sur.

Esto puede llevar a integrar más datos o a plantearse nuevas preguntas: El objetivo final es **concretar las preguntas específicas** que queremos responder mediante la visualización de datos, **con los datos necesario para ello**, y **escoger la representación visual que mejor se ajuste** a la casuísticas que se quiere visualizar.

Diseño: Cómo representar la información

Considerar si se quiere abordar el proyecto considerando las dos grandes distinciones que existen (infografía vs visualización de datos), así como sus diferentes tipos y categorías. En este sentido conviene tener en cuenta los contenidos docentes expuestos en "**Introducción a la visualización de la información**".

La elección de qué representación / tipo de gráfico usar se basa en optar por la opción que represente de forma más eficaz los aspectos más relevantes de nuestro conjunto de datos, pero también responder a las preguntas planteadas sobre el mismo.

Visual encodings: Determinar qué atributos visuales (posición, forma, tamaño, color,...) visualizan de una forma más adecuada el dato que representan. Datos ordinales, cuantitativos o categóricos se adecuarán mejor a ciertos atributos visuales que a otros, por ejemplo.

Diseño: Cómo representar los datos

Del mismo modo, recordar los conceptos mostrados en la "**Guía para crear una visualización**" del material docente:

- Uso adecuado del color.
- Proporcionar contexto mediante el uso del texto.
- Principios de diseño considerando los atributos preatentivos.
- Buenas prácticas (ratio datos/tinta, bonito vs funcional).
- Patrones de interacción para amplificar la visualización (filtrado, *brushing*, *progressive disclosure*).

Encontraréis también en la "*Data Visualization Checklist*" de Stephanie Evergreen una guía útil para tener una idea más objetiva de la calidad de vuestras visualizaciones.

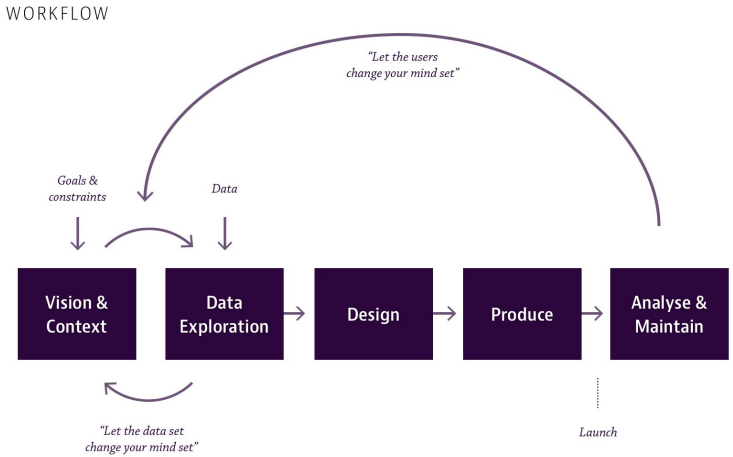
El proceso de visualización

La visualización de datos es un proceso multidisciplinar. Como tal implica diversas etapas que pertenecen a distintos ámbitos y que implican adquirir habilidades distintas (computación, diseño, interacción, etc.)

Es importante seguir una metodología, como algunas de las propuestas por diversos especialistas en el campo de la visualización de la información, como Fry o Stefaner.

El objetivo de la práctica es que trabajéis este aspecto reflexivo del proceso: **¿qué estoy haciendo, por qué, qué aprendo?**

Process of Computational Information Design, Ben Fry.



Information Visualization workflow, Moritz Stefaner

El proceso de visualización

El proceso de visualización es un proceso de **exploración continua**, visualizando los datos se responden a unas preguntas preestablecidas, pero surgen de nuevas, por lo que se crean nuevas visualizaciones que hacen empezar el proceso de nuevo.

La visualización de información debe convertir **datos en conocimiento**, por lo que tiene que ser un **proceso enriquecedor**. Cuando se tengan claro lo que se quiere visualizar, debemos preguntarnos qué hemos aprendido de:

- El conjunto de datos escogido
- Las visualizaciones generadas
- El proceso mismo de creación

Es importante, pues, saber qué conocimiento hemos extraído: ¿Cuáles han sido los hechos más interesantes? ¿Qué respuestas han resultado inesperadas? ¿Qué ideas preconcebidas han resultado ser falsas? Y la más difícil: ¿Qué no hemos tenido en cuenta?

Resumen

Así pues, la visualización de información va más allà de la creación de gráficos de forma automática mediante *software*, es un proceso que tiene que asegurar la generación de conocimiento a partir de los datos. A continuación un resumen de la metodología expuesta anteriormente:

- Seleccionar **un conjunto de datos adecuado** para crear una visualización
- Realizar una inspección de la calidad los datos
- Determinar mis usuarios y **qué preguntas quiero responder**
- Realizar una **exploración visual** de los datos
- Considerar si debo iterar en el proceso (incorporar más datos, nuevas preguntas, etc.)
- Realizar una **fase de diseño y refinar aspectos visuales** que aseguran
    - que la datos / información se muestran con rigor
    - que las preguntas planteadas pueden ser contestadas
- **Evaluación del proceso**: ¿Qué se ha aprendido durante el proceso? ¿Debo mejorar aspectos de mi visualización?

Universitat Oberta
de Catalunya

UOC.universitat
@UOCuniversidad
UOCuniversitat