Universitat Oberta de Catalunya

uoc.edu

Aspectes a considerar per la Pràctica de visualització de dades

Autor: Xavier Giménez
Coordinador: Julià Minguillón

Índex

Selecció del conjunt de dades

Objectiu

Seleccionar un conjunt de dades adequat per a crear una visualització

- **Rellevància:** el conjunt de dades hauria de ser d'actualitat, significatiu (evitar dades sintètiques), d'interès general, i permetre plantejar preguntes interessants
- **Dimensions:**
  - De l'ordre de 1000-10000 files
  - De l'ordre de 10-100 columnes
- **Característiques:** combina dades numèriques i categòriques
- **Conté alguna jerarquia:** categoria / subcategoria
- **Atenció a temes sensibles:** dades personals, des-anonimització, etc.
- **Quina és la llicència del conjunt de dades? D'on s'ha obtingut?**

Preparació

El conjunt de dades ha de ser inspeccionat per detectar possibles necessitats més endavant

- **Valors perduts**
  - Eliminar files / columnes amb un excés de dades perdudes
- **Homogeneïtzar dades categòriques (p. ex. “Barcelona” / “BCN”)**
- **Detectar valors extrems en dades numèriques**
  - Box-plots
  - Histograma
- **Extracció de característiques**
- **Combinar-ho amb altres conjunts de dades**

Fonts de dades

Infinites opcions

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

Exemples

Temes que poden ser adequats

- **Covid:** casos, morts, vacunes, etc., per àrea geogràfica, per període de temps, per grup d'edat, per gènere, ...
- **Escalfament global:** mesures històriques de temperatura, emissions de gasos d'efecte hivernacle, mesures del desgel en els pols, ...
- **Govern obert:** dades de despesa pública, gestió governamental de pressupostos, adjudicació d'obra pública, …

Exemples

Temes que poden ser adequats

Qualsevol tema pot ser adequat, partint sempre d'uns requisits:

- **Veracitat de la font de dades**: comprovar quina entitat, organisme o institució ha creat les dades.
- **Rigor en el procés d'obtenció de dades**: identificar possibles *data mining pitfalls*: errors en l'obtenció de les dades, poca cobertura de la totalitat de la mostra, ambigüitat, etc.
- **Context:** Considerar afegir més fonts d'informació si les dades que disposem no són autoexplicatives per sí mateixes (p.e. visualitzar dades de creixement econòmic sense tindre dades històriques de la inflació).

Propostes

Dades que poden generar visualitzacions interessants

- **Índex de carrers:** a partir dels carrers d'un municipi, classificar-los (manualment!) en funció de diversos criteris, p.ex. gènere (dona / home / neutral), origen (religiós / històric / geogràfic / cultural / …), longitud, … Es poden creuar amb dades de tràfic, població, etc.

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

- **Wikipedia**: anàlisi de les últimes modificacions realitzades, per usuari, tipologia, mida de la modificació, etc.

  https://ca.wikipedia.org/wiki/Especial:Canvis_recents
  http://hint.fm/projects/historyflow/

Propostes

Dades que poden generar visualitzacions interessants

- **Escalfament global:** donar valor a les dades històriques produïdes per les diferents agències nacionals de meteorologia per visualitzar canvis en la temperatura global del planeta.

  https://showyourstripes.info/s/globe

- **Migració**: Visualitzar històries humanes (un mateix ha d'obtenir i crear les dades!) sobre immigrants que cerquen asil fora del seu país d'origen.

  http://www.storiesbehindaline.com/

Exemple pràctic

Font: Eurostat, el portal d'estadística de la Comissió Europea

- Publica indicadors estadístics en l'àmbit europeu, com a resultat de la col·laboració entre els diferents instituts estadístics dels països membres de la UE.

- Ofereix en format *Open Data* conjunts de dades estadístics i indicadors sobre diverses àrees tals com Economia i finançament, Indústria, Serveis, Agricultura, Ciència, etc., així com indicadors sobre benestar social, desenvolupament sostenible, economia circular, ...

Exemple pràctic

Font: Eurostat, el portal d'estadística de la Comissió Europea

- Podem assumir que la font de dades és fiable i que està lliure d'errors metodològics relatius a la seva obtenció i processament.

- Inclou una secció amb exemples interactius de visualitzacions de dades per a explorar les dades:

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

Exemple pràctic

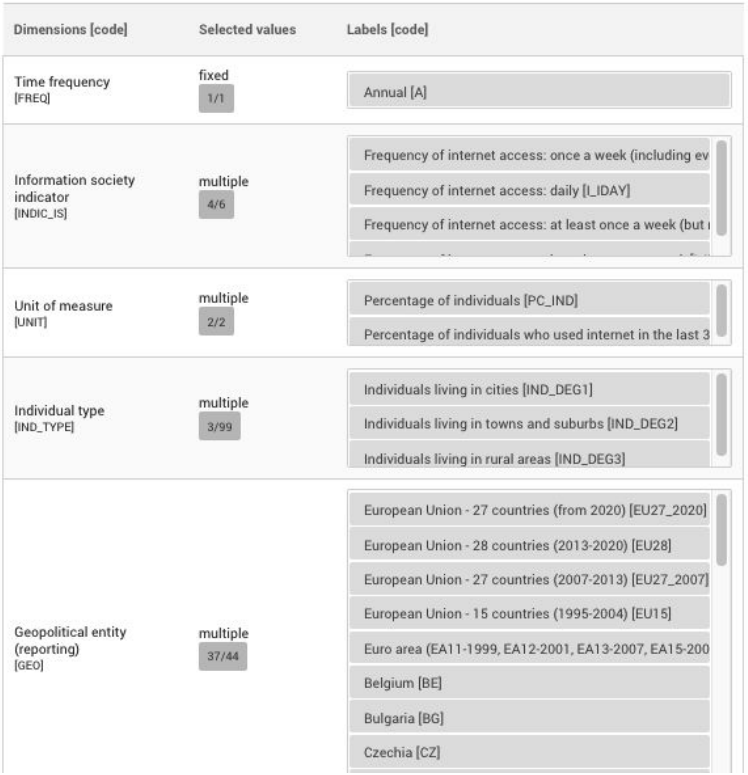
***Dataset*: Freqüència d'ús d'internet per part de la ciutadania**

- Conjunt de dades format per 22 columnes i 6479 files, en format llarg per les variables categòriques i en format ample per les columnes amb informació temporal.
- Reflecteix situacions reals (i habituals): el conjunt de dades no és 100% perfecte ni complet, com succeeix amb la majoria de conjunts de dades.
- Múltiples valors categòrics:
  - **Freqüència d'accés a internet** (6 valors): diari, 1 cop / setmana, 1 cop / mes, etc.
  - **Tipus de població** (99 valors): individus segregats per edat, educació, tipus de feina, situació geogràfica, etc.
  - **País o composició històrica de la UE** (44 valors)
  - **Anys** (18 valors, 2003-2020)
  - ...

Exemple pràctic: Freqüència d'ús d'internet per part de la ciutadania

El conjunt de dades inclou certes **característiques rellevants** (i apropiades per a l'ús de la visualització de dades), com per exemple:

- Multitud de possibles valors categòrics, amb un major nombre de possibles combinacions de dades.



Exemple pràctic: Freqüència d'ús d'internet per part de la ciutadania

El conjunt de dades inclou certes **característiques rellevants** (i apropiades per a l'ús de la visualització de dades), com per exemple:

- Multitud de possibles valors categòrics, amb un major nombre de possibles combinacions de dades.
- Sèries de dades incompletes, amb diversa casuística: dades no disponibles, valors poc fidedignes, sèries temporals incompletes, 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
(bn) break in time series, not significant
(n) not significant
(u) low reliability

Exemple pràctic: Freqüència d'ús d'internet per part de la ciutadania

El conjunt de dades inclou certes **característiques rellevants** (i apropiades per a l'ús de la visualització de dades), com per exemple:

- Multitud de possibles valors categòrics, amb un major nombre de possibles combinacions de dades.
- Sèries de dades incompletes, amb diversa casuística: dades no disponibles, valors poc fidedignes, sèries temporals incompletes, etc.
- Estructura: el *dataset* no està en un format adequat (p.ej. *tidy-data*) pel seu ús en visualitzacions.

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

Exemple pràctic: Freqüència d'ús d'internet per part de la ciutadania

Aquest conjunt de dades **és adequat per la visualització de dades,** ja que:

- Prové d'una organització fiable i rigorosa.
- Conté una alta densitat de dades, tan numèriques com categòriques.
- Permet realitzar un gran ventall de preguntes (p.e., ús de la tecnologia segons diversos àmbits: regional, per grups de població, etc.)
- És fàcilment combinable amb altres conjunts de dades d'interès (p.e., indicadors socioeconòmics europeus).

El procés de visualitzar informació

Plantejament de preguntes

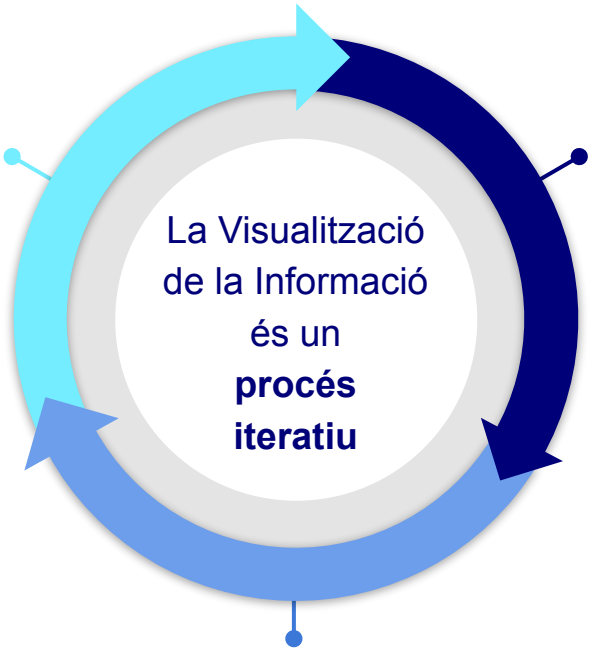
- **Usuaris**: Determinar a qui va dirigida la visualització.
    - Els meus usuaris són el públic en general o un segment concret, p.ex. experts?
    - Cal proporcionar un context previ a la visualització de les dades?

- **Preguntes**: Què vull respondre amb la visualització?
    - El dataset em permet contestar-les amb precisió?
    - Quines dades concretes del meu conjunt de dades vull explorar?
    - Necessito més fonts de dades per donar context al meu conjunt de dades?

És important revisar les decisions preses en la selecció del conjunt de dades i veure si n'hi ha prou per aconseguir els objectius desitjats.

Plantejament de preguntes

3. Avaluació:
El meu conjunt de dades és suficient per a respondre les preguntes? Em cal afegir noves dades? ¿Sorgeixen més preguntes degut a la complexitat de les dades?



1. Elecció de les dades:
el conjunt de dades ha de tindre els requisits per a poder plantejar preguntes rellevants

2. Preguntes: exposar preguntes amb un objectiu clar i un àmbit definit, que es puguin quantificar i que no depenguin de suposicions

Exploració de les dades

En funció de la naturalesa del conjunt de dades (dades en forma tabular, numèriques i categòriques), serà interessant realitzar una fase d'inspecció o un E.D.A. (*Explorative Data Analysis*) per tal de:

- Mostrar quines distribucions segueixen els diferents valors presents en les dades.
- Detectar patrons, tendències i valors atípics.
- Detectar associacions entre variables.

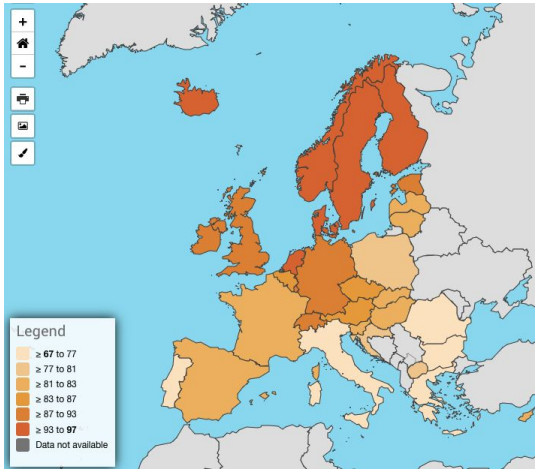
En altres casos (dades no estructurades, grafs, texte, mapes, etc.) l'equivalent a aquests anàlisis pasa per plantejar-se altres preguntes, com per exemple:

- Detectar relacions entre entitats en dades no tabulars.
- Mostrar patrons geogràfics (fluxes de movilitat, fenòmens poblacionals).
- Detectar temàtiques en dades no estructurades (caracteritzar documents de texte).

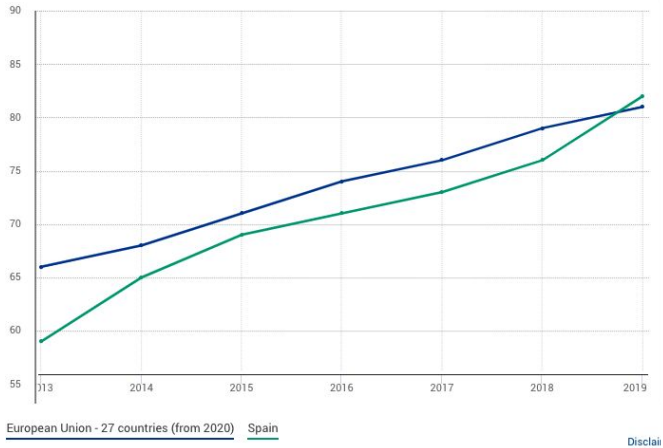
Exploració de les dades

Seguint amb el nostre exemple pràctic de conjunt de dades (Freqüència d'ús d'internet per part de la ciutadania), l'exploració consisteix en visualitzacions que ens permetin entendre la informació present en les dades, per exemple:

Quin % de població a Europa fa diariament ús d'internet?



Quina ha sigut l'evolució de l'ús d'internet a Espanya en comparació amb la Unió Europea?

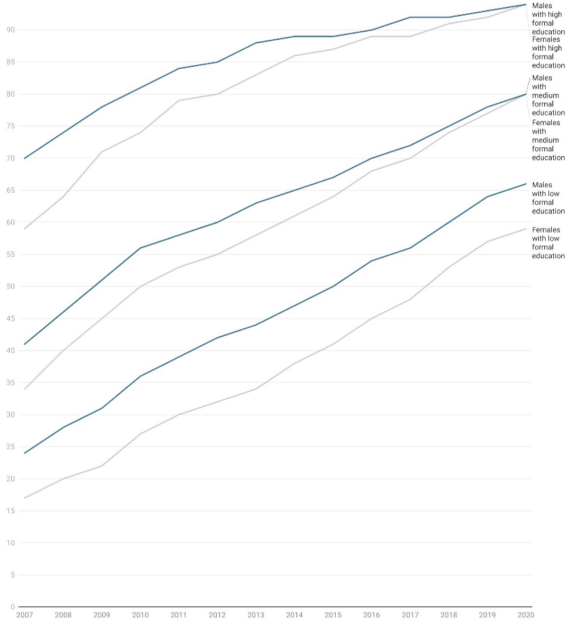


Exploració de les dades

Seguint amb el nostre exemple pràctic de conjunt de dades (Freqüència d'ús d'internet per part de la ciutadania), l'exploració consisteix en visualitzacions que ens permetin entendre la informació present en les dades, per exemple:

Quines diferències hi ha en l'ús d'internet entre els diferents grups poblacionals?

Com difereixen entre sí els grups segons gènere i nivell d'estudis?



Exploració de les dades

La fase d'exploració de dades té que proporcionar:

- Una idea clara de la informació continguda en el conjunt de dades.
- Detectar inconsistències i/o errors en les dades.
- Validar la factibilitat de les preguntes que pretenem resoldre.
- Facilitar noves preguntes i/o exposar la necessitat d'afegir noves fonts de dades.

Per exemple, el mapa d'ús d'internet ens mostra clarament unas diferències Nord-Sud que no s'observen fàcilment en les dades quan estan en un format tabular, i que ens poden fer pensar en associacions amb altres variables que també mostrin aquestes diferències.

Això ens pot portar a integrar més dades o plantejar noves preguntes: l'objectiu final és **concretar les preguntes específiques** que volem respondre a través de la visualització, amb **les dades necessàries**, i **escollir la representació visual que millor s'ajusti** a les casuístiques que volem visualitzar.

Disseny: Com representar la informació

Considerar si es vol afrontar el projecte considerant les dues grans distincions que existeixen (infografia vs. visualització de dades), així com els seus diferents tipus i categories. En aquest sentit convé tindre en compte els continguts docents exposats en "**Introducció a la visualització de la informació**".

L'elecció de quina representació / tipus de gràfic emprar es basa en optar por l'opció que representi de forma més eficaç els aspectes més rellevants del nostre conjunt de dades, però també respondre a les preguntes plantejades sobre el mateix.

Visual encodings: Determinar quins atributs visuals (posició, forma, mida, color,...) visualitzen d'una forma més adient la dada que representen. Dades ordinals, quantitatives o categòriques s'ajustaran millor a certs atributs visuals que altres, per exemple.

Disseny: Com representar la informació

De la mateixa manera, recordar els conceptes mostrats a la "**Guia per crear una visualització**" del material docent:

- Ús adequat del color.
- Proporcionar context mitjançant l'ús del text.
- Principis de disseny considerant els atributs preatentius.
- Bones pràctiques (ràtio dades/tinta, bonic vs funcional).
- Patrons d'interacció per amplificar la visualització (filtratge, *brushing, progressive disclosure*).

Trobareu també a la "*Data Visualization Checklist*" de Stephanie Evergreen una guia útil per tindre una idea més objectiva de la qualitat de les vostres visualitzacions.

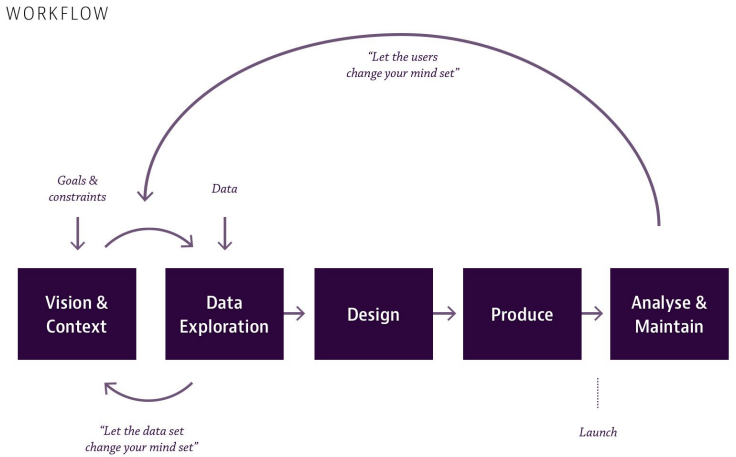
El procés de visualització

La visualització de dades és un procés multidisciplinar. Com a tal implica diverses etapes que pertanyen a diferents àmbits i que impliquen adquirir habilitats diferents (computació, disseny, interacció, etc.)

És important seguir una metodologia, com algunes de les propostes per diversos especialistes en el camp de la visualització de la informació, com ara Fry o Stefaner.

L'objectiu de la pràctica és que treballeu aquest aspecte reflexiu del procés: **què estic fent, per què, què n'aprenc?**

Process of Computational Information Design, Ben Fry.



Information Visualization workflow, Moritz Stefaner

El procés de visualització

El procés de visualització és un procés d'**exploració contínua**, visualitzant les dades es responen a unes preguntes preestablertes, però en sorgeixen de noves, per la qual cosa es creen noves visualitzacions que fan tornar a començar el procés.

La visualització d'informació ha de convertir **dades en coneixement**, per la qual cosa ha de ser un **procés enriquidor**. Quan es tinguin clar el que es vol visualitzar, ens hem de preguntar què hem après de:

- El conjunt de dades escollit
- Les visualitzacions generades
- El procés mateix de creació

És important, doncs, saber quin coneixement hem extret: Quins han estat els fets més interessants? Quines respostes han resultat inesperades? Quines idees preconcebudes han resultat ser falses? I la més difícil: què no hem tingut en compte?

Resum

Així, doncs, la visualització d'informació va més enllà de la creació de gràfics de forma automàtica mitjançant programari, és un procés que ha d'assegurar la generació de coneixement a partir de les dades. A continuació, un resum de la metodologia exposada anteriorment:

- Seleccionar un **conjunt de dades adient** per crear una visualització
- Realitzar una inspecció de la qualitat les dades
- Determinar els meus usuaris i **quines preguntes vull respondre**
- Realitzar una **exploració visual** de les dades
- Considereu si he d'iterar en el procés (incorporar més dades, noves preguntes, etc.)
- Realitzar una **fase de disseny i refinar aspectes visuals** que asseguren
- que les dades / informació es mostren amb rigor
- que les preguntes plantejades poden ser contestades
- **Avaluació del procés**: Què s'ha après durant el procés? s'han de millorar aspectes de la meva visualització?

Universitat Oberta
de Catalunya

UOC.universitat
@UOCuniversidad
UOCuniversitat